Modeling and Inference of Connectivity in the Brain from EEG and Exogenous Stimulation

by

Jui-Yang Chang

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Electrical Engineering)

at the

UNIVERSITY OF WISCONSIN–MADISON

2018

Date of final oral examination: 08/31/2017

The dissertation is approved by the following members of the Final Oral Committee:
Barry Van Veen, Professor, Electrical and Computer Engineering
Robert Nowak, Professor, Electrical and Computer Engineering
Akabar Sayeed, Professor, Electrical and Computer Engineering
Rebecca Willett, Associate Professor, Electrical and Computer Engineering
Brad Postle, Professor, Psychology

To my parents and my wife.

ACKNOWLEDGMENTS

First of all, I would like to thank my advisor, Professor Barry Van Veen, for his generosity in sharing his research ideas. Without his ideas, I would not have finished my dissertation on this topic. Over the past seven years, Professor Van Veen have also provided me support in almost every aspect – finance, research collaboration opportunities, career and life choice mentoring – I cannot thank Professor Van Veen more for his guidance and teaching. I learned a lot from his thinking and him as a person.

Secondly, I could not have done my research without Professor Marcello Massimini, Professor Giulio Tononi and Professor Brad Postle and their groups. They and their associates and students have been really generous in sharing their data and collaborating with me. Among their associates I would like to thank especially Dr. Andrea Pigorini and Dr. Matteo Fecchio for their ideas, effort, and constant support.

Thirdly, I would like to thank my doctoral committee members for their guidance and feedback. In particular, I would like to thank Professor Rob Nowak for leading me into the field of machine learning through his teaching and sharing. My research and career choices have been drawing significant influence from this field.

I would also like to thank my lab mates and collaborators : Dr. TJ Colgan (for his humor and caring), Dr. Patrick Cheung (for sharing his research ideas), Dr. Ricardo Pizarro, Prathyusha Sarma (for her sharing of research ideas and career path), Shiwei Zhou and Dr. Sheida Malekpour. I could not have completed my coursework and research without their support. I also would like to thank Professor Laurent Lessard for having me as a teaching assistant in his course. I learned a lot from his passion for teaching and his insights into optimization and machine learning. I also want to thank my former roommate, Dr. Tzu-Chi Lin for his sharing of his ideas and suggestions on job hunting.

I would also like to thank the faculty and staff of the Department of Electrical and Computer Engineering and the Department of Computer Science, for creating a great learning environment. I also want to thank the National Institutes of Health for funding my PhD research. Thank you to Ms. Julie Rae and Ms. Kelly Hayek for their guidance on job searching.

I have left many mentors and friends unmentioned, from whom I received guidance and support at the University of Wisconsin–Madison and National Tsing

Hua University. Their kindness and generosity will be memorized.

Finally, I would like to thank my parents, my wife, Ya-Fang and my brother for their endless and unconditional support. The whole journey to the degree would not have begun without them.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Brain stimulations have seen various applications in medical treatment, pre-surgical evaluation, and understanding of brain functions. Despite the popularity, the functional interaction among different regions during the stimulation remains unclear. In this work, we propose to use the multivariate autoregressive model with exogenous stimulation (MVARX) to model evoked response excited by brain stimulation and use the model to advance our understanding of brain networks during stimulation sessions.

We start by showing that the MVARX model well characterizes cortical signals by modeling intracerebral electroencephalography (EEG) excited by current stimulation. The application of the model in learning structural properties is demonstrated with models learned under two different structural hypotheses from subjects in wakefulness and sleep. We also estimate MVARX model from intracerebral EEG with the self-connected group lasso (SCGL) regularization and present the capability of the SCGL in identifying sparsity pattern of coefficient matrices in wakefulness and sleep.

Building on the results in cortical level, we propose to model the sensor level evoked response excited by brain stimulation with a linear state space model in which the cortical activities are characterized with the MVARX model. We use the transcranial magnetic stimulation (TMS)-EEG data to demonstrate the applicability of the model. The regions of interest in the model are selected with a data driven approach based on cortical signal power and the model parameters are estimated with an expectation-maximization algorithm. We demonstrate that the model is capable of modeling the TMS-EEG evoked response and that the feedback connections in model is necessary in characterizing the data in wakefulness.

Finally, the linear state space model is further considered from the Bayesian paradigm in which we consider imposing two classes of priors with differing structural preferences. We propose a variational inference procedure for learning the posterior distribution of the parameters. We demonstrate that the structural preferences encoded in the priors encourage identification of underlying network patterns with several simulated studies.

1 INTRODUCTION

Brain stimulation is widely used in treatment of brain illness [1, 2], pre-surgical evaluation [3], and understanding of brain functionality [4]. A variety of stimulation modalites are used, including transcranial magnetic stimulation (TMS), intracerebral electrical stimulation, and deep brain stimulation. Despite the broad application of brain stimulation, the functional interactions among different brain areas in response to stimulation remain largely unknown. Therefore it is of great interest to study the interaction mechanism from a systematic point of view.

Dynamic causal modeling (DCM) has been a popular method for analyzing magneto-/electroencephalography (M/EEG) evoked responses [5, 6]. The model is a Bayesian nonlinear state space model where the source activities are characterized with neural mass models, which are formulated as differential equations, and the sensor-space measurements are modeled as linear combinations of the source activities. In DCM, the stimulation directly excites neural populations in the source model and generates the evoked response according to the differential equations. Despite being physiologically plausible, the very large number of parameters in DCM makes it infeasible to consider networks with more than a few, e.g., 5-7, nodes. Moreover, DCM recommends learning the model parameters from the average evoked response, which is obtained by averaging over multiple trials [7]. The limited number of data samples in the average evoked response makes it challenging to estimate the parameters reliably. Furthermore, the focus on evoked response also means that DCM doesn't characterize the trial-by-trail fluctuations in the data, which also contain rich information about the brain [8].

Another popular model for modeling brain data is the multivariate autoregressive (MVAR) model. The evoked response due to stimulation manifests a time-varying mean, which is contrary to the stationary assumption of the MVAR model. In the literature, adaptive MVAR models have been proposed to study evoked responses. In [8], the authors proposed viewing each trial with multiple overlapping short windows and subtracting ensemble means from each time point in the window to make each window of constant mean. MVAR models are then estimated from each window. Subtracting the mean limits this procedure from characterizing information in the average evoked responses and degrades the signal-to-noise ratio (SNR) of the data, which in turn enlarges the mean squared error of the estimated parameters. Another approach [9] models the evoked response with time varying MVAR coefficients and considers modeling to be a tracking problem. Recursive least squares (RLS) algorithms are used to estimate the coefficients. Nevertheless, neither of the adaptive MVAR models directly consider the exogenous input in the model as in DCM. Furthermore, the adaptive MVAR models proposed in these works require a preprocessing step to estimate cortical signals from the measurements, which limits the modeling procedures to high SNR data.

The MVAR model is linear and simpler to work with than nonlinear models like DCM. Moreover, as was shown in [10], consistent estimates of the MVAR parameters can be estimated from low SNR data by modeling EEG with a linear state space model and jointly estimating the cortical source orientation, cortical signal and MVAR model parameters with the expectation-maximization (EM) algorithm. The estimation performance obtained with the state space approach and the simplicity of MVAR models suggest that MVAR models have promising potential for modeling responses evoked by brain stimulation.

The objective of this research is to extend the MVAR model to directly accommodate the exogenous stimulation. We propose to model the evoked response to be a multivariate autoregressive with exogenous stimulation (MVARX) process. Our MVARX model represents the stimulation as an input that is passed through a bank of FIR filters, which model the effect of fibers of passage from the stimulation location to the cortical regions of interest (ROIs). We consider issues around model selection and validation. The EM algorithm proposed in [10] is extended to estimate MVARX model parameters from scalp recordings such as TMS-EEG. Lasso regularization and Bayesian priors are utilized to incorporate structural hypotheses on the MVARX model in a data driven way. We apply the MVARX model to human brain activity to gain new insights into brain connectivity.

Specifically, in Chapter 2 we consider modeling intracerebral EEG excited by current stimulation with the MVARX model. We perform model selection with cross-validation and model checking with a residual whiteness test. We show that the estimated models work well in characterizing the evoked response and one-step predicting fluctuations in single trial data. We compare the structures of the brain networks during wakefulness and sleep by estimating models with two different structural hypotheses and assess the performance of the models in characterizing the evoked responses. Finally we show an application of the MVARX model in measuring level of consciousness with the integrated information theory.

Building from the results in modeling intracerebral EEG, in Chapter 3 we further consider identifying the structure of network with a data-driven way utilizing the self-connected group lasso penalty. We show that the sparse models outperform the unconstrained models in characterizing the data. We conclude that the identified sparse network structure agrees with physiological understandings of the brain.

In Chapter 4 we consider scalp EEG responses triggered by exogenous stimulation with a linear state space model. The scalp measurements are linear combinations of cortical signals, which in turn are modeled as an MVARX process. We propose jointly estimating the model parameters and the cortical signals with an EM algorithm. We use TMS-EEG measurements to demonstrate validity of our methods. The model is applied to estimate models under two different structural hypotheses and to demonstrate the difference between data in wakefulness and sleep.

Finally Chapter 5 considers the linear state space model from the Bayesian point of view. We consider prior distributions over the parameters and estimate approximate posterior distribution of the parameters with a variational inference procedure. We show the estimation procedure for priors under two classes of structural assumptions and compare the priors by studying several simulations.

2 MULTIVARIATE AUTOREGRESSIVE MODELS WITH EXOGENOUS INPUTS FOR INTRACEREBRAL RESPONSES TO DIRECT ELECTRICAL STIMULATION OF THE HUMAN BRAIN

2.1 Introduction

The remarkable cognitive abilities of the healthy human brain depend on an exquisite balance between functional specialization of local cortical circuits and their functional integration through long-range connections [1] [2]. Hence, there is considerable interest in characterizing long-range cause and effect or directional interactions in the human brain. Multivariate autoregressive (MVAR) models, sometimes referred to as vector autoregressive (VAR) models, have been widely applied to study directional cortical network properties from both intracranial data (e.g., [11, 12, 13, 14, 15]) and scalp EEG or MEG (e.g., [16, 17]). An MVAR model describes each signal as a weighted combination of its own past values and the past values of other signals in the model — an autoregression — plus an error term. The weights relating the present of one signal to the past of another capture the causal or directed influence between signals. A variety of different metrics for summarizing the directed interactions in MVAR models have been proposed, including directed transfer functions [18], directed coherence [19], conditional Granger causality [20], and integrated information [21].

MVAR models assume the data is stationary and of constant mean. While stationarity and constant mean may be reasonable assumptions for a relatively short duration of spontaneous data, evoked or event-related data appear to violate these assumptions. For example, the mean or average response to a stimulus varies with time. An MVAR model fit to data with a time-varying mean results in spurious interactions because the assumption of stationarity is violated. Adaptive or time-varying methods have been developed to relax stationarity assumptions [8, 22, 23]. For example, a time-varying mean response is removed by subtracting the

[1] Reprinted from [J.-Y. Chang, A. Pigorini, M. Massimini, G. Tononi, L. Nobili, and B. D. Van Veen, "Multivariate autoregressive models with exogenous inputs for intracerebral responses to direct electrical stimulation of the human brain," Frontiers in human neuroscience, vol. 6, 2012.].

[2] This research was supported in part by the National Institute of Biomedical Imaging and Bioengineering under grant R21EB009749.

ensemble average [8] and the MVAR model parameters are allowed to vary with time. Adaptive models require specification of an adaptation rate parameter that effectively determines how much of the past data is used to estimate the present model parameters, or equivalently, how fast the model is changing. Models that use fast adaptation are able to track faster changes in the underlying data, but employ less data to estimate model parameters and consequently possess more variability in the estimated model parameters (see [23] for assessment of these issues).

During the presurgical evaluation of drug-resistant epileptic patients, direct electrical stimulation of the brain is systematically performed for diagnostic purposes to identify the epileptogenic zone [24]. Electrical stimulation generates a time-varying response at the recording sites. In this paper we propose describing the response of the brain using stationary MVAR models with an exogenous input (MVARX) derived from the stimulus characteristics. MVARX models are commonly used in econometric time series analysis [25]. The advantage of the MVARX model is that it does not require subtraction of the mean and consequent reduction in signal-to-noise ratio (SNR) or the complication of time-varying models to capture the response evoked by direct electrical stimulation. The model captures both the mean evoked response and the background activity present during the recordings. We demonstrate the effectiveness of the MVARX model using intracerebral recordings from epilepsy patients.

Direct electrical stimulation of the brain presents several modeling challenges. Although the timing and location of the stimulus is known precisely, the response of the brain in the near vicinity of the stimulus cannot be measured due to electrical artifacts and the propagation of the stimulus to more distant sites depends on the topology of axons in the vicinity of the stimulation site [26]. Electrical stimulation creates action potentials in neurons whose axons pass near the stimulus site. These neurons synapse both near and distant to the stimulation site, so the stimulus actually activates multiple, a priori unknown areas. The MVARX model explicitly accounts for this effect with a bank of finite impulse response (FIR) filters that capture the impact of the exogenous input, i.e., stimulus, on all recording sites. The exogenous input filter coefficients and the MVAR model parameters are simultaneously estimated from the recordings and knowledge of the stimulation times using a least squares procedure. The exogenous input filter coefficients describe the conduction paths from the stimulus site to each recording site, while the MVAR model parameters capture the causal interactions between recording sites.

The MVARX model is applied to 10 datasets collected from three subjects in wakefulness and NREM sleep. Two stimulation levels are studied in one subject, and two stimulation sites in another. The data consists of the intracranial response to 30 current impulses separated by one second. A cross-validation procedure is introduced for choosing the memory in the MVARX model. We demonstrate that a stationary MVARX model accurately describes the activity evoked by direct electrical stimulation. Comparison to a series of univariate autoregressive models with exogenous inputs (ARX) reveals that causal interactions must be modeled to accurately describe the measured activity. The series of ARX models result in much larger modeling error than the MVARX model. One-step prediction performance is used to demonstrate that the MVARX model also captures spontaneous fluctuations in the recorded data. The MVARX model errors pass a whiteness test while the univariate ARX models do not, further supporting the applicability of the MVARX model.

The MVARX models are employed to contrast integrated information in wakefulness and sleep. Integrated information is a measure of the extent to which the information generated by the causal interactions in the model cannot be partitioned into independent subparts of the system. Hence, integrated information measures the balance between functional specialization and integration represented by the model. Theoretical considerations [27, 28, 29, 30] indicate that integrated information should be less in sleep than in wakefulness. This prediction is confirmed in all 10 datasets using our MVARX model.

This paper is organized as follows. Section 2.2 describes the data and preprocessing procedures. Section 2.3 defines the MVARX model, introduces the method for estimating the model parameters, including our cross-validation approach for selecting model memory, and presents the residual whiteness test. Section 2.4 demonstrates the effectiveness of the proposed model using the 10 datasets described above and Section 2.5 applies the MVARX models to contrast integrated information in wakefulness and sleep. This paper concludes with a discussion in Section 2.6. For notation, boldface lower and upper case symbols represent vectors and matrices, respectively, while superscript T denotes matrix transpose and superscript -1 denotes matrix inverse. The trace of a matrix $\mathbf{A}$ is $\mathrm{tr}[\mathbf{A}]$ and the determinant is $\det(\mathbf{A})$. $\mathsf{E}\{a\}$ denotes the expectation of a random variable a. The Euclidean norm of a vector $\mathbf{x}$ is $||\mathbf{x}||_2 = \sqrt{\mathbf{x}^\mathsf{T}\mathbf{x}}$. The number of elements in a set S is $|S|$. $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ means that the vector $\mathbf{x}$ is normally distributed with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

2.2 Data

Subjects and experimental protocol

Three subjects with long-standing drug-resistant focal epilepsy participated in this study. All patients were candidates for surgical removal of the epileptic focus. During pre-surgical evaluation the patients underwent individual investigation with stereotactically implanted intracerebral multilead electrodes for precise localization of the epileptogenic areas [31]. All patients gave written informed consent before intracerebral electrode implantation as approved by the local Ethical Committee. Confirmation of the hypothesized seizure focus and localization of epileptogenic tissue in relation to essential cortex was achieved by simultaneous scalp and intracerebral electrode recording, as well as intracerebral stimulation during wakefulness and sleep to further investigate connectivity of epileptogenic and healthy tissue [32, 33]. The decision on implantation site, duration of implantation and stimulation site(s) was made entirely on clinical needs. Stereoelectroencephalography (SEEG) activity was recorded from platinum-iridium semiflexible multilead intracerebral electrodes, with a diameter of 0.8 mm, a contact length of 2 mm, an intercontact distance of 1.5 mm and a maximal contact number of 18 (Dixi Medical, Besançon, France) [31]. The individual placement of electrodes was ascertained by post-implantation tomographic imaging (CT) scans. Scalp EEG activity was recorded from two platinum needle electrodes placed during surgery at "10–20" positions Fz and Cz on the scalp. Electroocular activity was registered at the outer canthi of both eyes, and submental electromyographic activity was acquired with electrodes attached to the chin. EEG and SEEG signals were recorded using a 192-channel recording system (Nihon-Kohden Neurofax-110) with a sampling rate of 1000 Hz. Data was recorded and exported in EEG Nihon-Kohden format [34, 35]. The data for each channel is obtained using bipolar referencing to a neighboring contact located entirely in the white matter. Intracerebral stimulations were started on the third day after electrode implantation. In eight out of ten cases we discuss, stimulation of strength 5 mA

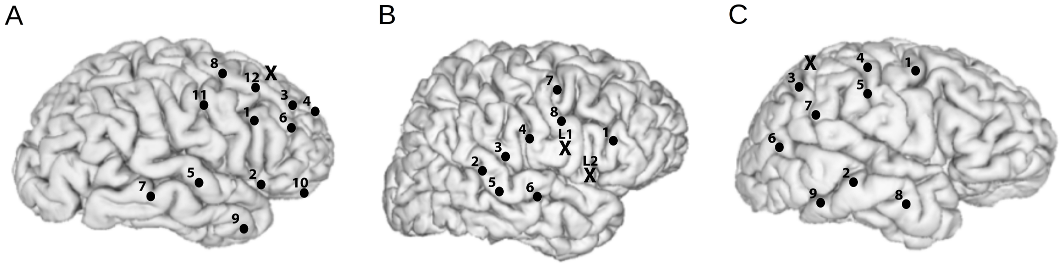


Figure 2.1: Recording and stimulation electrode placements for the subjects. Black dots represents recording channels while black ‘X’ represents stimulating channel(s). (a) Subject A, left hemisphere is shown. 1 - inferior frontal opercular, 2 - anterior horizontal lateral fissure, 3 - middle frontal gyrus, 4 - middle frontal sulcus, 5 - superior temporal sulcus, 6 - inferior frontal sulcus, 7 - middle temporal gyrus, 8 - middle frontal gyrus, 9 - middle temporal gyrus, 10 - orbital gyrus, 11 - precentral gyrus, 12 - superior frontal sulcus, X - middle frontal gyrus. (b) Subject B, right hemisphere is shown. 1 - inferior frontal gyrus, 2 - superior temporal sulcus, 3 - posterior lateral fissure, 4 - postcentral solcus, 5 - superior temporal gyrus, 6 - transversal temporal sulcus, 7 - superior frontal gyrus, 8 - subcentral gyrus. X (L1) - precentral gyrus, X (L2) - subcentral sulcus. (c) Subject C, right hemisphere is shown. 1 - precentral gyrus, 2 - posterior middle temporal gyrus, 3 - inferior parietal lobule, 4 - postcentral gyrus, 5 - postcentral sulcus, 6 - angular gyrus, 7 - supramarginal gyrus, 8 - anterior middle temporal gyrus, inferior temporal gyrus, X - superior parietal lobule.

were performed, while for the other two cases stimulation of 1 mA were applied. At each stimulation session, the stimulation is applied at a single channel and SEEG recordings were obtained from all other channels. A single stimulation session consisted of a 30 impulse stimulation train at intervals of 1 s. Each impulse is of 0.2-ms duration. The channels that were stimulated were chosen based on clinical requirements. All patients included in this study were stimulated during wakefulness and stage 4 of NREM sleep. Sleep staging was performed using standard criteria [36]. Stimulations which elicited muscle twitches, sensations or cognitive symptoms, were excluded from this study, in order to prevent possible awareness of stimulation or alteration of sleep depth.

In our analysis, we consider a subset of 8-12 recording channels of all channels for each subject, as illustrated in Fig. 2.1. The 8-12 channels were selected based on approximately maximizing the distance between the subset of channels that are both artifact free and near the surface of the cortex.

Preprocessing

During each stimulation session, a raw trigger signal that indicates the occurrence of current stimulation with 1 and the absence of stimulation with 0 is collected at a sampling rate of 1000 Hz in addition to the SEEG recordings. We use a Tukey-windowed median filter to remove volume conduction artifacts within 39 ms of each stimulus. First, a median filter of order 19 is applied to the raw data channel by channel. Next, the raw data within a 39-ms window centered at each stimulus is replaced with a weighted average of the raw data and the median filtered data to eliminate the artifact. The weights for the median filtered data take the form of a Tukey window ([37], pg. 69) and are zero for +/- 20 ms away from the stimulus, a cosine rising from 0 to 1 beginning at 19 ms prior to the stimulus and ending at 10 ms prior to the stimulus, unity until 10 ms poststimulus, and then a cosine decreasing from 1 to 0 ending at 19 ms poststimulus. The weighting applied to the raw data are one minus those applied to the median filtered data. Fig. 2.2 illustrates the results of this process. The cleaned data is then lowpass filtered by an FIR filter with passband-edge of 48 Hz and stopband-edge of 49.9 Hz to eliminate 50 Hz powerline contamination, and the lowpass filtered data is downsampled by a factor of 10 to a sampling frequency of 100 Hz. The filtering and downsampling limits the affect of filtering to very edges of the Nyquist band. The stimulated portion of the downsampled data are further segmented into 30 epochs of data $\mathbf{y}_n^{(j)}$, each of which contains 100 samples. The start of each epoch is from 12 samples (0.12 s) before the occurrence of a stimulus and the end is 87 samples (0.87 s) post-stimulus. Similarly, the raw trigger signal is lowpass filtered, downsampled by 10, and partitioned into 100-sample epochs $x_n^{(j)}$.

In principle, filtering the signal may have an impact on model estimation and causality inference [38]. We minimize the potential impact of filtering by specifying the stop-band edge of the lowpass filter close to the Nyquist frequency.

Identification of outlying epochs

An automated procedure is employed to exclude epochs that markedly deviate from the majority of epochs due to nonstationary brain activity or other factors. Let $\mathbf{y}_n^{(j)} = [y_{1,n}^{(j)}, y_{2,n}^{(j)}, \cdots, y_{d,n}^{(j)}]^T$ represent the d channels of recordings at time $n = 1, 2, \cdots, N_j$ from epochs $j = 1, 2, \cdots, J$. For epoch m, we compute the time-varying mean

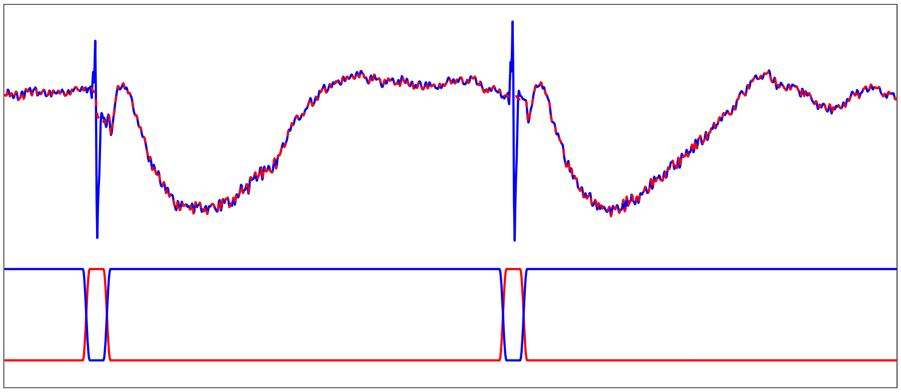


Figure 2.2: Tukey-windowed median filtering for eliminating volume conduction artifacts. The upper trace depicts an example of raw data (blue solid line) and the weighted cosine-tapered median filter output (red dashed line). The lower trace depicts the weighting applied to the raw data (blue solid line) and the median filtered data (red solid line) to eliminate the volume conduction artifact.

$\mu_{\mathbf{y}}^{-m}(n)$ and time-varying covariance matrix $\Sigma_{\mathbf{y}}^{-m}(n)$ from all other epochs. That is,

$$\mu_{\mathbf{y}}^{-m}(n) = \frac{1}{J-1} \sum_{j=1, j \neq m}^{J} \mathbf{y}_n^{(j)} \tag{2.1}$$

$$\Sigma_{\mathbf{y}}^{-m}(n) = \frac{1}{J-2} \sum_{j=1, j \neq m}^{J} \left(\mathbf{y}_n^{(j)} - \mu_{\mathbf{y}}^{-m}(n)\right) \left(\mathbf{y}_n^{(j)} - \mu_{\mathbf{y}}^{-m}(n)\right)^T, \tag{2.2}$$

for $n = 1, \cdots, 100$. Here $m = 1$ to J and J is 30 for all data sets considered. Then the squared Mahalanobis distance [39] between the epoch m and the other epochs is computed as

$$D^2(m) = \sum_{n=1}^{100} \left(\mathbf{y}_n^{(m)} - \mu_{\mathbf{y}}^{-m}(n)\right)^T \left(\Sigma_{\mathbf{y}}^{-m}(n)\right)^{-1} \left(\mathbf{y}_n^{(m)} - \mu_{\mathbf{y}}^{-m}(n)\right). \tag{2.3}$$

Epochs with $D^2(m)$ exceeding

$$100 \cdot d + 60\sqrt{2 \cdot 100 \cdot d} \tag{2.4}$$

Table 2.1: Number of non-outlying epochs used in analysis

Dataset	Wakefulness epochs	Sleep epochs
Subject A, 1 mA	29	25
Subject A, 5 mA	28	22
Subject B, L1	30	24
Subject B, L2	30	29
Subject C	30	29

are declared as outliers and removed from subsequent analysis. Intuitively, if the data is Gaussian, then $D^2(m)$ is Chi-squared distributed with $100 \cdot d$ degrees of freedom. This implies that the threshold rules out an epoch m if $D^2(m)$ exceeds its mean plus 60 standard deviations. Thus this threshold only excludes epochs that have a large deviation from the temporal average of the other epochs. The number of epochs retained for analysis are given in Table 2.1.

2.3 Methods

MVARX model

The MVARX model of order (p, ℓ) describes the data as follows [25]:

$$\mathbf{y}_n^{(j)} = \sum_{i=1}^{p} \mathbf{A}_i \mathbf{y}_{n-i}^{(j)} + \sum_{i=0}^{\ell} \mathbf{b}_i x_{n-i}^{(j)} + \mathbf{w}_n^{(j)}, \tag{2.5}$$

where $x_n^{(j)}$ denotes the input at time n and epoch j. The $d \times d$ matrices $\mathbf{A}_i = \{a_{m,n}(i)\}$ contain autoregressive coefficients describing the influence of channel n on channel m at lag i, and the $d \times 1$ vectors $\mathbf{b}_i = \{b_m(i)\}$ contain filter coefficients from the stimulus to channel m at lag i. The vectors $\mathbf{w}_n^{(j)}$ are $d \times 1$ zero-mean noise vectors with covariance matrix $\mathbf{Q}$ and are assumed to satisfy $E\{\mathbf{w}_n^{(i)}(\mathbf{w}_s^{(j)})^T\} = 0$, for either $i \neq j$ or $n \neq s$. The model assumes that the data is stationary over time and epochs. We further assume that the epochs are of varying lengths N_j and are possibly disconnected in time to accommodate rejection of outlying epochs. Fig. 2.3 depicts a schematic diagram of an example MVARX model. The diagram assumes there are three recording electrodes corresponding to the recordings $y_{1,n}, y_{2,n}$, and $y_{3,n}$ (the epoch index j is omitted in the figure for simplicity). The intracranial EEG signals recorded at the electrodes contain contributions due to the current stimulus response and background brain activity. The exogenous input x_n represents the current stimulation. If $\mathbf{B} = [\mathbf{b}_0, \cdots, \mathbf{b}_\ell]$ is a $d \times (\ell + 1)$ matrix of exogenous input coefficients, then the i-th row of $\mathbf{B}$, $[\mathbf{B}]_{i,:}$, is the impulse response of the filter representing the unknown transmission characteristics between the current stimulus and the i-th recording channel. The autoregressive coefficients $\mathbf{A} = [\mathbf{A}_1, \cdots, \mathbf{A}_p]$ indicate how past values of the recorded signals affect present values. The autoregressive order p determines the time extent of the past that affect the present values and may be regarded as the memory of the system. The signals $w_{1,n}$, $w_{2,n}$, and $w_{3,n}$ can be interpreted as modeling errors or alternatively as a process that generates spontaneous activity.

Electrodes can be used either as stimulating or recording electrodes but cannot be used simultaneously for recording and stimulation. Moreover, the electrodes closest to the stimulation site are affected by huge electrical artifacts and they cannot be used because of consequent low SNR. Hence the recorded data $\mathbf{y}_n^{(j)}$ contains recordings of the effect of the stimulation at distant sites, not the stimulation itself. Stimulation depolarizes the membranes of neurons passing through the neighborhood of the stimulating electrode, possibly creating action potentials in neurons that synapse near the stimulation site and at distant locations [26], a phenomenon termed fibers of passage. Thus, stimulation generates an "input" that is conveyed to potentially all recording sites in a manner that depends on the axonal topology in the vicinity of the stimulation site. This topology and consequent stimulation effects are usually unknown and described in our MVARX model by the exogenous input filters $\mathbf{B}$. In our model we assume the exogenous input is given by the trigger signal associated with delivery of a current pulse, so $\mathbf{B}$ captures both the shape of the delivered stimulus and the unknown direct propagation of the input to each recording site.

Denote $\mathbf{y}_{n,s}^{(j)}$ and $\mathbf{y}_{n,e}^{(j)}$ as the spontaneous activity and stimulus response to the exogenous input, respectively, at time n from epoch j. Equation (2.5) can be alternatively expressed as

$$\mathbf{y}_n^{(j)} = \mathbf{y}_{n,s}^{(j)} + \mathbf{y}_{n,e}^{(j)} \tag{2.6}$$

$$\mathbf{y}_{n,s}^{(j)} = \sum_{i=1}^{p} \mathbf{A}_i \mathbf{y}_{n-i,s}^{(j)} + \mathbf{w}_n^{(j)} \tag{2.7}$$

$$\mathbf{y}_{n,e}^{(j)} = \sum_{i=1}^{p} \mathbf{A}_i \mathbf{y}_{n-i,e}^{(j)} + \sum_{i=0}^{\ell} \mathbf{b}_i x_{n-i}^{(j)}. \tag{2.8}$$

Note that in practice $\mathbf{y}_{n,s}^{(j)}$ and $\mathbf{y}_{n,e}^{(j)}$ are not directly observed and cannot be separated from $\mathbf{y}_n^{(j)}$ without knowledge of the MVARX model parameters. The stimulus response component $\mathbf{y}_{n,e}^{(j)}$ is a deterministic term that depends entirely on the stimulus and the model. Given the model parameters $\boldsymbol{\Theta} = [\mathbf{A}, \mathbf{B}]$, we can generate $\mathbf{y}_{n,e}^{(j)}$ by applying the stimulus sequence $\mathbf{x}_n^{(j)}$ to (2.8) with zero initial conditions. Recall that $\mathbf{w}_n^{(j)}$ is assumed to be zero mean, so $\mathbf{y}_{n,s}^{(j)}$ is a zero mean random process reflecting the spontaneous component of the recordings. It is common in MVAR modeling to subtract the mean prior to estimating MVAR model parameters [8]. This corresponds to removing the stimulus response $\mathbf{y}_{n,e}^{(j)}$ and is unnecessary with the MVARX model. We shall assume that the stimulus is repeated multiple times such that averaging $\mathbf{y}_{n,e}^{(j)}$ with respect to the stimulus onset times produces the evoked response of the system. This is not required by the model in (2.5) but is consistent with conventional electrophysiology practice.

The autoregressive parameters $\mathbf{A}$ model the inherent neural connectivity between sites - how activity at one site propagates to another site. This is evident in (2.5-2.8) by the fact that the $\mathbf{A}_i$ are applied to $\mathbf{y}_{n-i}^{(j)}$. If the spontaneous activity $\mathbf{y}_{n,s}^{(j)}$ is very weak relative to $\mathbf{y}_{n,e}^{(j)}$ then the response is described entirely by (2.8) and the measured data $\mathbf{y}_n^{(j)} \approx \mathbf{y}_{n,e}^{(j)}$. In this case there is a potential modeling ambiguity as there are many different combinations of $\mathbf{A}_i$ and $\mathbf{b}_i$ that could be used to describe $\mathbf{y}_{n,e}^{(j)}$ over a finite duration. For example, $\mathbf{y}_{n,e}^{(j)}$ can be described on $1 \leqslant n \leqslant \ell + 1$ by setting $\mathbf{A}_i = 0$ and only using $\mathbf{b}_i$. We control potential ambiguities associated with relatively weak spontaneous activity by limiting ℓ to a value commensurate with the expected duration of stimulus propagation through fibers of passage. This ensures that $\mathbf{B}$ is not able to capture long duration interactions associated with feed forward and feedback connectivity between sites. Based on previous experimental evidence [40], we set $\ell = 10$ to accommodate a 100 ms duration of propagation through fibers

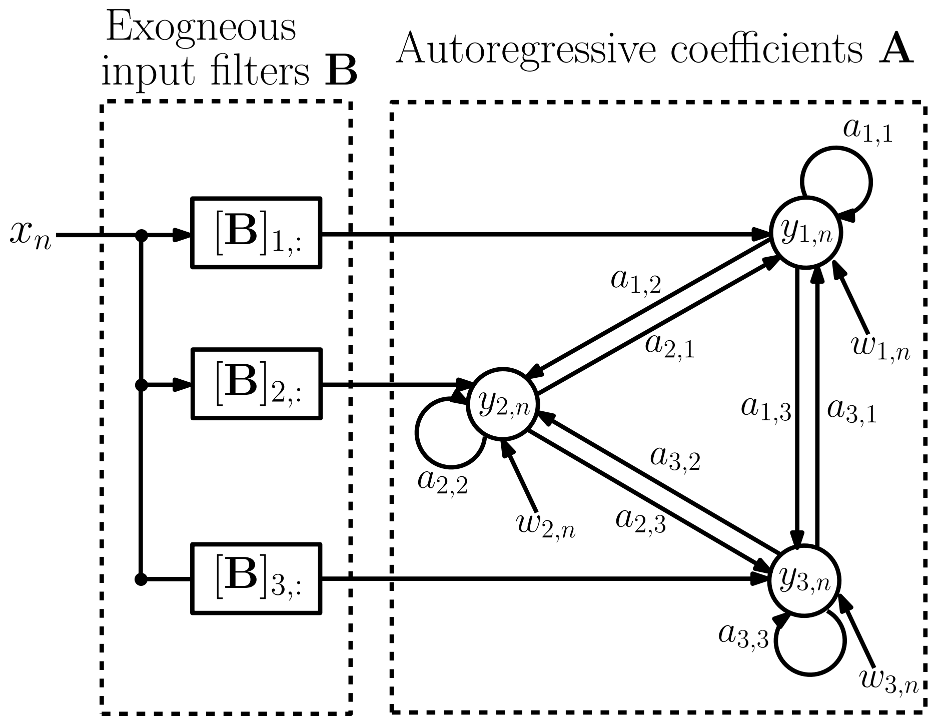


Figure 2.3: Schematic diagram of the MVARX model. $y_{i,n}$ denotes the recorded signals at the electrodes while x_n represents current stimulation and $w_{i,n}$ is model error, or equivalently, a random input that generates spontaneous activity. $a_{i,j}$ captures the a priori unknown connectivity between recording sites while $[\mathbf{B}]_{i,:}$ represents the a priori unknown transmission characteristics between the stimulus and recording sites.

of passage. We will discuss this choice more thoroughly in Section 2.6.

Estimation of MVARX model parameters

Suppose that we have the recordings and inputs $\{(\mathbf{y}_n^{(j)}, x_n^{(j)}) : j = 1, 2, \cdots, J, n = 1, 2, \cdots N_j\}$ for J epochs of N_j samples each. Denote $n_0 = \max(p, \ell)$, and suppose that $N_j \geqslant n_0 + 1$, for all j. Using the first n_0 samples as the initial values, the model in (2.5) can be rewritten in a simplified form:

$$\mathbf{y}_n^{(j)} = \mathbf{\Theta}\mathbf{z}_{n-1}^{(j)} + \mathbf{w}_n^{(j)}, \tag{2.9}$$

for $j = 1, \cdots, J, n = n_0 + 1, \cdots, N_j$, where the $d \times (dp + \ell + 1)$ matrix $\mathbf{\Theta} = [\mathbf{A}, \mathbf{B}]$ and the vector of dimension $dp + \ell + 1$, $\mathbf{z}_{n-1}^{(j)} = [(\mathbf{y}_{n-1}^{(j)})^\mathsf{T}, (\mathbf{y}_{n-2}^{(j)})^\mathsf{T}, \cdots, (\mathbf{y}_{n-p}^{(j)})^\mathsf{T}, x_n^{(j)}, x_{n-1}^{(j)},$

$\cdots, x^{(j)}_{n-\ell}]^T$. The vectors $\mathbf{y}^{(j)}_n, \mathbf{w}^{(j)}_n$, and $\mathbf{z}^{(j)}_{n-1}$ can be further concatenated as columns of the matrices $\mathbf{Y}_j, \mathbf{Z}_j$, and $\boldsymbol{W}_j$ to write:

$$\mathbf{Y}_j = \boldsymbol{\Theta}\mathbf{Z}_j + \boldsymbol{W}_j \tag{2.10}$$

where $\mathbf{Y}_j = [\mathbf{y}^{(j)}_{n_0+1}, \cdots, \mathbf{y}^{(j)}_{N_j}], \mathbf{Z}_j = [\mathbf{z}^{(j)}_{n_0}, \cdots, \mathbf{z}^{(j)}_{N_j-1}]$, and $\boldsymbol{W}_j = [\mathbf{w}^{(j)}_{n_0+1}, \cdots, \mathbf{w}^{(j)}_{N_j}]$. This expression takes the form of a linear regression model, and we can obtain an ordinary least square (OLS) estimate of $(\boldsymbol{\Theta}, \mathbf{Q})$ as ([25], chap. 10.3):

$$\hat{\boldsymbol{\Theta}} = \big(\sum_{j=1}^{J} \mathbf{Y}_j\mathbf{Z}_j^T\big)\big(\sum_{j=1}^{J} \mathbf{Z}_j\mathbf{Z}_j^T\big)^{-1}, \quad \hat{\mathbf{Q}} = \frac{1}{N_t}\sum_{j=1}^{J} \big(\mathbf{Y}_j - \hat{\boldsymbol{\Theta}}\mathbf{Z}_j\big)\big(\mathbf{Y}_j - \hat{\boldsymbol{\Theta}}\mathbf{Z}_j\big)^T, \tag{2.11}$$

where $N_t = \sum_{j=1}^{J} N_j - n_0 J$. If $\mathbf{w}^{(j)}_n$ is Gaussian, then the OLS estimate $(\hat{\boldsymbol{\Theta}}, \hat{\mathbf{Q}})$ is also the maximum-likelihood estimate of $(\boldsymbol{\Theta}, \mathbf{Q})$ [25].

Model Selection with cross-validation

In practice the order p could be chosen using numerous different model selection criteria, including Akaike information criterion and the Bayesian information criterion [25, 41]. Here we use cross-validation (CV) to determine p in a data-driven fashion (see [42] for another example of using CV to select model parameters with neurophysiological data). The data $\mathbf{y}^{(j)}_n$ and input $x^{(j)}_n$ are partitioned into training and test sets. The goal is to choose the value p that produces the best prediction of test data when the model $\boldsymbol{\Theta} = [\mathbf{A}, \mathbf{B}]$ is estimated from the training data. We consider two components in assessing model predictive capability. The first is the one-step prediction error, a measure of the model's ability to track the sample-to-sample and epoch-to-epoch fluctuations in the data. The second is the error between the average evoked response predicted by the model and the measured average response. This measures the quality of the model's response to the stimulus.

Partition the epochs of available data into training sets R_m and test sets S_m and assume there are $m = 1, 2, \cdots, M$ such partitions. Assume the sets S_m are non-overlapping and are of approximately the same size. Let $\boldsymbol{\Theta}_m$ be the model estimated from R_m as described in the preceding subsection. The one-step prediction error at time n, $\boldsymbol{e}^{(j)}_n(\boldsymbol{\Theta}_m)$ is the difference between the recording $\mathbf{y}^{(j)}_n$ and the one-step prediction made by $\boldsymbol{\Theta}_m$ using the n_0 samples prior to time n, that is, $\boldsymbol{z}_{n-1}^{(j)}$:

$$\mathbf{e}_n^{(j)}(\boldsymbol{\Theta}_m) = \mathbf{y}_n^{(j)} - \hat{\mathbf{y}}_n^{(j)}(\boldsymbol{\Theta}_m) \tag{2.12}$$

where the one-step prediction $\hat{\mathbf{y}}_n^{(j)}(\boldsymbol{\Theta}_m) = \boldsymbol{\Theta}_m \boldsymbol{z}_{n-1}^{(j)}$. Similarly we define the average response error as

$$\epsilon_n(\boldsymbol{\Theta}_m) = \overline{\mathbf{y}}_n(S_m) - \hat{\overline{\mathbf{y}}}_n(\boldsymbol{\Theta}_m, S_m) \tag{2.13}$$

where the average evoked response $\overline{\mathbf{y}}_n(S_m) = 1/|S_m| \cdot \sum_{j \in S_m} \mathbf{y}_n^{(j)}$ and the average model response $\hat{\overline{\mathbf{y}}}_n(\boldsymbol{\Theta}_m, S_m)$ over epochs in S_m, $\hat{\overline{\mathbf{y}}}_n(\boldsymbol{\Theta}_m, S_m) = 1/|S_m| \cdot \sum_{j \in S_m} \mathbf{y}_{n,e}^{(j)}(\boldsymbol{\Theta}_m)$. Here $\mathbf{y}_{n,e}^{(j)}(\boldsymbol{\Theta}_m)$ is generated using $\boldsymbol{\Theta}_m$ as described following (2.8). We define a CV score as a weighted combination of the one-step prediction and average response errors averaged over all training/test data partitions

$$CV(p) = \frac{1}{M} \sum_{m=1}^{M} \left[\frac{CV_e(p,m)}{w_e} + \frac{CV_\epsilon(p,m)}{w_\epsilon} \right] \tag{2.14}$$

where $CV_e(p,m)$ is the mean square one-step prediction error of a p-th order model $\boldsymbol{\Theta}_m(p)$ in predicting data in S_m:

$$CV_e(p,m) = \frac{1}{|S_m|} \sum_{j \in S_m} \frac{1}{N_j - n_0} \sum_{n=n_0+1}^{N_j} \|\mathbf{e}_n^{(j)}(\boldsymbol{\Theta}_m(p))\|_2^2 \tag{2.15}$$

and $CV_\epsilon(p,m)$ is the mean square value of the average response error on S_m:

$$CV_\epsilon(p,m) = \frac{1}{N} \sum_{n=1}^{N} \|\epsilon_n(\boldsymbol{\Theta}_m(p))\|_2^2. \tag{2.16}$$

Here N is the assumed duration of the average response. The weights w_e and w_ϵ vary the emphasis between the one-step prediction error and average response error. In the analysis below, we set w_e and w_ϵ to the medians of $CV_e(p,m)$ and $CV_\epsilon(p,m)$, respectively, for $m = 1, \cdots, M$ and all p considered. This approach places approximately equal emphasis on the two errors. The model order p is chosen as the p that minimizes $CV(p)$ over the range of p evaluated.

Several practical issues require attention for computing the average response error. First, use of an average evoked response assumes the stimulus is nominally identical for each epoch. Second, care must be taken in computing the average response of the model $\boldsymbol{\Theta}$ to the stimulus $x_n^{(j)}$ over epochs in S_m if the effects of preceding stimuli extend into S_m. In such a case the brain is not “at rest” upon the arrival of the new stimulus in S_m, but is still responding to the preceding stimulus. This situation occurs when the response time of the cortex is longer than the inter-stimulus interval. We mimic this aspect of the measured data when computing the average model response by presenting the entire train of stimuli to the model and averaging over the responses corresponding to epochs in S_m.

Model quality assessment

A key assumption for the consistency of the OLS estimates is that the residuals $\mathbf{w}_n^{(j)}$ be serially uncorrelated, that is, temporally white. Serial correlation in $\mathbf{w}_n^{(j)}$ may be a sign of mis-specifying the model or incorrect selection of order (p, ℓ) [43, 44]. We use a consistent test developed in [44] to validate our models. Denote $\Gamma_{\mathbf{w}}(r) = E\{\mathbf{w}_n^{(j)}(\mathbf{w}_{n-r}^{(j)})^T\}$ the covariance at lag r, the hypotheses of interest are:

$$\begin{aligned} H_0 : \quad & \Gamma_{\mathbf{w}}(r) = \mathbf{0}, \text{ for all } r \neq 0 \quad \text{vs.} \\ H_1 : \quad & \Gamma_{\mathbf{w}}(r) \neq \mathbf{0}, \text{ for some } r \neq 0. \end{aligned} \tag{2.17}$$

Let the residual at time n in epoch j be $\hat{\mathbf{w}}_n^{(j)} = \mathbf{y}_n^{(j)} - \hat{\boldsymbol{\Theta}}\mathbf{z}_{n-1}^{(j)}$. Let $q(\cdot)$ be a window function of bounded support L, that is, $q(r) > 0$, for $|r| \leqslant L$ and $q(r) = 0$ for $|r| > L$. Suppose that the last epoch is of length longer than $(J-1)L$, that is, $N_J > (J-1)L$. The test statistic derived in [44] for testing H_0 vs H_1 is

$$T_{N_c} = \frac{N_c \sum_{r=1}^{L} q^2(r) \mathrm{tr}[\mathbf{C}_{\hat{\mathbf{w}}}^T(r)\mathbf{C}_{\hat{\mathbf{w}}}^{-1}(0)\mathbf{C}_{\hat{\mathbf{w}}}(r)\mathbf{C}_{\hat{\mathbf{w}}}^{-1}(0)] - d^2 M_{N_c}(q)}{[2d^2 V_{N_c}(q)]^{1/2}} \tag{2.18}$$

where $N_c = \sum_{j=1}^{J} N_j - (J-1)L$ and

$$\mathbf{C}_{\hat{\mathbf{w}}}(r) = \frac{1}{N_c}\left[\sum_{j=1}^{J-1}\sum_{n=r+1}^{N_j} \hat{\mathbf{w}}_n^{(j)}(\hat{\mathbf{w}}_{n+r}^{(j)})^T + \sum_{n=r+1+(J-1)(L-r)}^{N_J} \hat{\mathbf{w}}_n^{(J)}(\hat{\mathbf{w}}_{n+r}^{(J)})^T\right], \tag{2.19}$$

for $r = 0, 1, \cdots, L$, are the estimated residual covariance matrices. The functionals $M_{N_c}(q)$ and $V_{N_c}(q)$ of $q(\cdot)$ and N_c are defined as [44]:

$$M_{N_c}(q) = \sum_{i=1}^{L-1} \left(1 - \frac{i}{N_c}\right) q^2(i) \tag{2.20}$$

$$V_{N_c}(q) = \sum_{i=1}^{L-2} \left(1 - \frac{i}{N_c}\right) \left(1 - \frac{(i+1)}{N_c}\right) q^4(i). \tag{2.21}$$

We use the Bartlett window defined as $q(j) = 1 - |j/L|, j \leqslant L$ and $q(j) = 0, j > L$ with a window width $L = \lceil 3N_c^{0.3} \rceil$ as suggested in [44]. For example, in our datasets the longest possible single epoch would have $N_c = 3000$ samples, which leads to the maximum value $L = 34$. Thus the test statistic (2.18) is based on estimated residual covariance matrices at lags less than or equal to 34. Under the assumption that both $\mathbf{y}_n^{(j)}$ and $x_n^{(j)}$ are stationary, the test statistic is one-sided and asymptotically standard normally distributed (see [44] Theorem 1). It declares that the residuals are serially correlated if $T_N > z_{1-\alpha}$ and are white otherwise, where $z_{1-\alpha}$ is the value of the inverse cumulative distribution function of the standard normal distribution at $1 - \alpha$ and α is the significance level of the test.

2.4 Results

Model parameters

We have varying definitions and lengths of epochs throughout our data processing procedures. For detection of outlying epochs we choose all epochs to be of length $N_j = 100$ samples based on the time between subsequent current stimuli. In model estimation and assessment of residual whiteness, the epochs are defined as the maximum contiguous segments between the time segments removed by the outlier detection process. This minimizes the impact of the initial conditions $\mathbf{z}_{n_0}^{(j)}$ required at the start of each epoch. Hence, N_j varies across epochs and conditions. In CV, the epoch lengths are set to be equal with $N_j = 100$. This, along with choosing the test sets S_m to contain approximately the same number of epochs, makes the test sets span roughly the same amount of time.

As shown in Table 2.1, the number of outlying epochs is generally larger in sleep

Table 2.2: Model order parameters for wakefulness and sleep data sets.

	Wakefulness			Sleep		
Dataset	CV Part.	MVARX p	ARX p	CV Part.	MVARX p	ARX p
Subj A, 1 mA	7	20	30	8	20	30
Subj A, 5 mA	7	26	30	11	26	26
Subj B, L1	10	30	28	8	30	24
Subj B, L2	10	18	22	7	22	30
Subj C	10	16	30	7	12	30

than in wakefulness, most likely due to the presence of slow waves during sleep. The number of partitions of the available epochs used in the CV procedure for determining model order p and the corresponding model order is shown in Table 2.2. We did not consider model orders higher than $p = 30$. We also evaluated an unconnected model consisting of d univariate ARX models to assess the importance of the coupling or connectivity between channels. The univariate models were estimated by applying the procedure described above to each channel. With the exception of subject B, stimulus location 1 (L1), the CV procedure picks a higher model order for the unconnected model and in many cases chooses the maximum order considered.

The whiteness test described in Section 2.3 was applied to the residuals from all models using a significance level $\alpha = 0.1$. Note that since exceeding the threshold implies the residuals are not white, use of a relatively large value for α leads to a more stringent test, that is, makes it easier to declare the residuals are not white. The MVARX models passed the whiteness test for every data set, while the unconnected models failed the test for every data set.

Evoked response model performance

In Figs. 2.4–2.6 we compare the average evoked response and average model response for a subset of subjects and conditions. The average responses are generated following the CV approach described in Section 2.3. Figs. 2.4 (A) and (B) show the average CV evoked responses $\overline{\mathbf{y}}_n(S) = M^{-1} \sum_{m=1}^{M} \overline{\mathbf{y}}_n(S_m)$ and average CV model responses $\hat{\overline{\mathbf{y}}}_n(\boldsymbol{\Theta}, S) = M^{-1} \sum_{m=1}^{M} \hat{\overline{\mathbf{y}}}_n(\boldsymbol{\Theta}_m, S_m)$ in channels 1, 4, 7, and 11 of subject A in wakefulness for 1 mA and 5 mA stimulation, respectively. Here 0 s on the time axis corresponds to the stimulus onset. The averaging is first done within the testing block for each CV partition, then a second phase of averaging is done over the average responses of the test blocks for all CV partitions. The average CV model response of the MVARX model (blue dashed line) follows the dynamics of the average CV evoked response (green solid line) in each channel, for both stimulus amplitudes and a range of channel response levels. In contrast, the average CV model response of the unconnected model (red dashed dot line) only tracks the average CV evoked response in channels with the largest amplitudes, even though the univariate model is fit independently to each channel. In the figures, error bars indicating one standard error are displayed every five samples. Figs. 2.4 (C)–(F) summarize the model performance on a channel-by-channel basis. Let $\overline{y}_{i,n}(S)$ and $\hat{\overline{y}}_{i,n}(\Theta, S)$ be the average CV evoked response and average CV model response at time n in the i-th channel. Figs. 2.4 (C) and (E) depict the normalized mean-squared difference (NMSD) between the average CV evoked and average CV model response for 1 mA and 5 mA stimulation, respectively, where the NMSD in channel i is defined as

$$\mathrm{NMSD}(i) = \frac{\sum_{n=1}^{N}(\overline{y}_{i,n}(S) - \hat{\overline{y}}_{i,n}(\Theta, S))^2}{\sum_{n=1}^{N}\overline{y}_{i,n}^2(S)}. \tag{2.22}$$

Figs. 2.4 (D) and (F) depict the relative root mean-squared energy (RRMS) for 1 mA and 5 mA stimulations, respectively, for each channel. The RRMS for channel i is defined as the ratio of the root mean-squared energy in channel i to that of the channel with the largest root mean-squared energy. More precisely,

$$\mathrm{RRMS}(i) = \frac{\sqrt{\sum_{n=1}^{N}\overline{y}_{i,n}^2(S)}}{\max\limits_{i'=1,\cdots,d}\sqrt{\sum_{n=1}^{N}\overline{y}_{i',n}^2(S)}}. \tag{2.23}$$

The unconnected model only gives comparable NMSD to that of full model in channel 11, which has the largest energy. The difference between the MVARX model and the unconnected model in terms of per-channel NMSD is less significant for the 1 mA stimulation, than for the 5 mA stimulation.

Figs. 2.5 (A) and (B) depict the average CV evoked and average CV model responses for subject A during NREM sleep with current stimulation of 1 mA and 5 mA, respectively. The four traces, from top to bottom, show the responses in channels 1, 7, 4, and 11, respectively. Panels (C) and (E) depict the NMSD, while (D) and (E) depict RRMS for 1 mA and 5 mA stimulation, respectively as a function of channel.

The average CV evoked responses and the average CV model responses in wakefulness for subject B, with two different stimulating sites L1 and L2, and both with current stimulus of 5 mA, are shown in panels (A) and (B) of Fig. 2.6. The four traces, from top to bottom, depict the responses in channels 1, 3, 6, and 8, respectively. The difference between the two stimulating sites lies mainly in channels with smaller energy, i.e., channels 1, 3, and 6. Panels (C) and (E) depict NMSD in each channel when the stimulating channel is L1 and L2, respectively. Panels (D) and (F) show the RRMS in each channel.

Define the normalized mean-squared response difference (NMRD) over all channels as the ratio of the mean-squared response difference to the mean-squared average CV evoked response. That is,

$$\text{NMRD} = \frac{\sum_{n=1}^{N} \|\overline{\mathbf{y}}_n(S) - \hat{\overline{\mathbf{y}}}_n(\boldsymbol{\Theta}, S)\|_2^2}{\sum_{n=1}^{N} \|\overline{\mathbf{y}}_n(S)\|_2^2}. \tag{2.24}$$

Fig. 2.7 depicts NMRD of the MVARX models for all five data sets considered. Generally the MVARX models captures the dynamics in average evoked response reasonably well with NMRD no larger than 0.25.

One-step prediction model performance

The ability of the model to predict the present recorded value of the data given past recordings reflects a different attribute than the modeling of the average evoked response. One-step prediction performance indicates the model's ability to follow spontaneous fluctuations in the data. Fig. 2.8 compares the recording $\mathbf{y}_n^{(j)}$ and one-step prediction $\hat{\mathbf{y}}_n^{(j)}(\boldsymbol{\Theta})$ of the signals recorded from subject B for 1.5 s of prestimulus data followed by two and a half epochs of evoked data, when the stimulating site is L2. The models used to perform prediction in Fig. 2.8 are trained from data excluding the data plotted. Panels (A) and (B) shows the signals in wakefulness and sleep, respectively. Similar results are obtained for the other epochs, subjects, and conditions. The traces show the signals in channels 1, 3, 6, and 8 respectively. These results indicate that the MVARX model performs accurate one-step prediction in wakefulness and sleep and for both prestimulus and evoked data segments.

Define the normalized mean-squared one-step prediction error (NMSE) as the

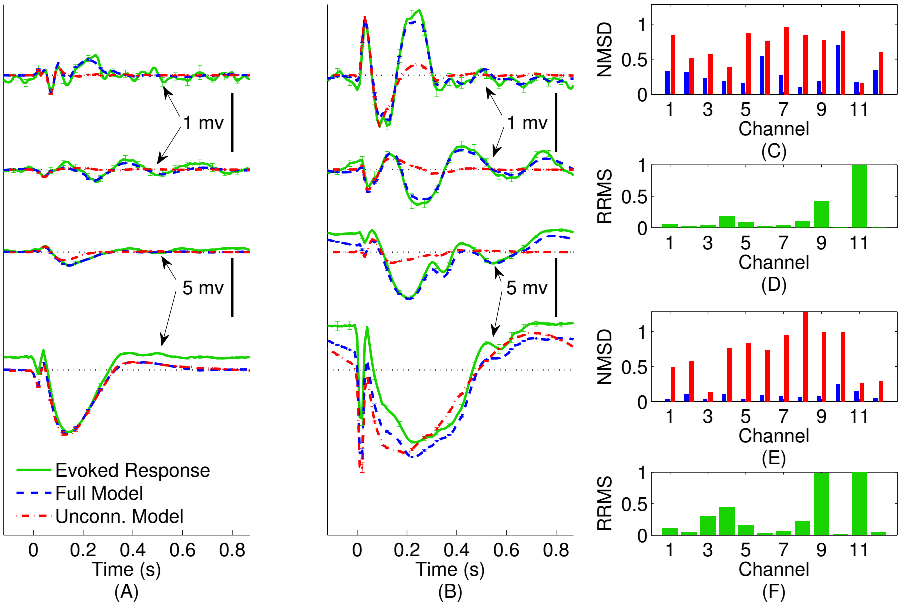


Figure 2.4: Comparison between average CV evoked and average CV model responses of subject A to two different stimulation strengths in wakefulness. In panels (A) and (B) the black dotted lines indicate the origin while the error bars denote the standard error of the mean. (A) Average CV evoked and average CV model responses of channels 1, 7, 4, and 11 with 1 mA current stimulation. (B) Average CV evoked and average CV model responses of channels 1, 7, 4, and 11 with 5 mA current stimulation. (C) Normalized mean-squared difference in each channel for 1 mA stimulation. (D) Relative root mean-squared energy in each channel for 1 mA stimulation. (E) Normalized mean-squared difference in each channel for 5 mA stimulation. (F) Relative root mean-squared energy in each channel for 5 mA stimulation.

ratio of the mean-squared prediction error over the samples to the mean squared energy. That is,

$$\mathrm{NMSE} = \frac{\frac{1}{J(N-n_0)}\sum_{j=1}^{J}\sum_{n=n_0+1}^{N}\|\mathbf{y}_n^{(j)} - \hat{\mathbf{y}}_n^{(j)}(\boldsymbol{\Theta})\|_2^2}{\frac{1}{JN}\sum_{j=1}^{J}\sum_{n=1}^{N}\|\mathbf{y}_n^{(j)}\|_2^2}. \tag{2.25}$$

As a reference, the NMSE of the model $\boldsymbol{\Theta} = \mathbf{0}$ is approximately 1. The bar diagrams in Fig. 2.9 show the NMSE of the MVARX models for all five datasets considered. Overall, our models give NMSE less than 0.06 for one-step prediction of the recordings and less than 0.02 in seven of the ten data sets studied.

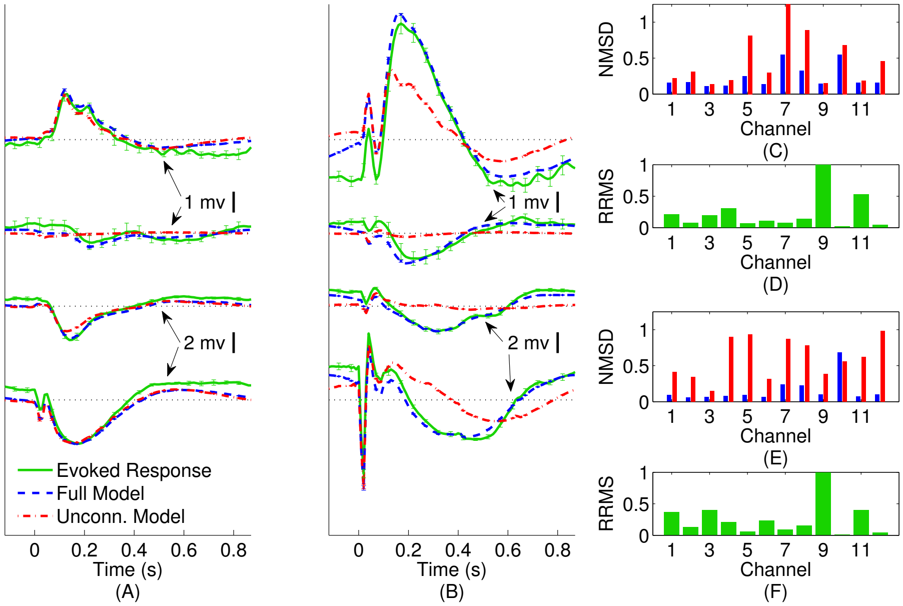


Figure 2.5: Comparison between average CV evoked and average CV model responses of subject A to two different stimulation strengths in sleep. In panels (A) and (B) the black dotted lines indicate the origin while the error bars denote the standard error of the mean. (A) Average CV evoked and average CV model responses of channels 1, 7, 4, and 11 with 1 mA current stimulation. (B) Average CV evoked and average CV model responses of channels 1, 7, 4, and 11 with 5 mA current stimulation. (C) Normalized mean-squared difference in each channel for 1 mA stimulation. (D) Relative root mean-squared energy in each channel for 1 mA stimulation. (E) Normalized mean-squared difference in each channel for 5 mA stimulation. (F) Relative root mean-squared energy in each channel for 5 mA stimulation.

B matrices

Fig. 2.10 depicts the exogenous input filters **B** matrices estimated for all 10 datasets as color plots. The i-th row of each matrix represents the FIR filter coefficients representing the path from the stimulus site to the i-th channel. Hence, rows with greater extremes of color have the strongest paths from the stimulus site.

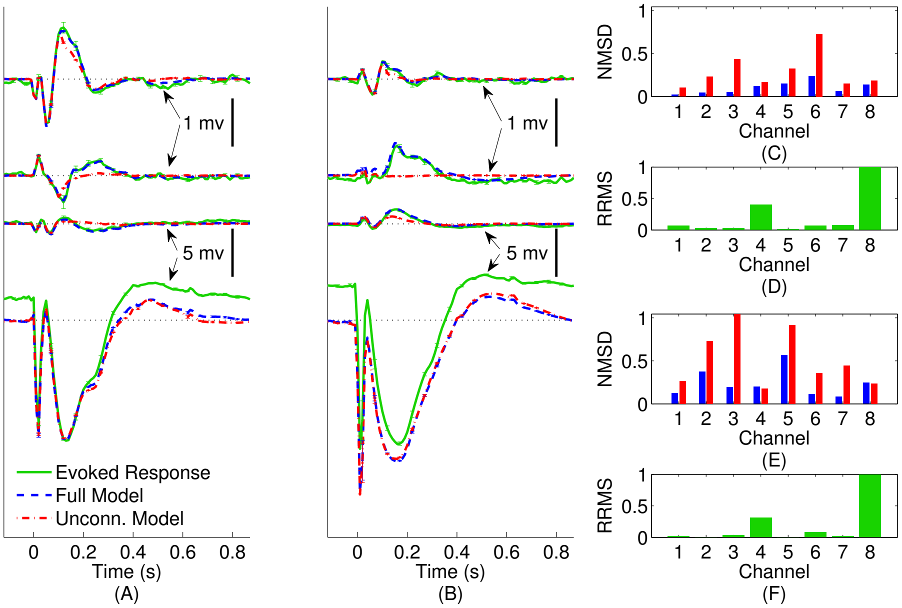


Figure 2.6: Comparison between average CV evoked responses and average CV model responses of subject B with two different stimulating locations in wakefulness. In panels (A) and (B) the black dotted lines indicate the origin while the error bars denote the standard error of the mean. (A) Average CV evoked and average CV model responses of channels 1, 3, 6, and 8 when the stimulating channel is L1. (B) Average CV evoked and average average CV model responses of channels 1, 3, 6, and 8 when the stimulating channel is L2. (C) Normalized mean-squared difference in each channel when the stimulating channel is L1. (D) Relative root mean-squared energy in each channel when the stimulating channel is L2. (E) Normalized mean-squared difference in each channel when the stimulating channel is L1. (F) Relative root mean-squared energy in each channel with the stimulating channel is L2.

2.5 Application to Consciousness Assessment

Numerous network characteristics can be obtained from an MVARX model. For example, graphs with partially directed coherence or conditional Granger causality as edges can be obtained by computing partially directed coherence or conditional Granger causality from the MVARX parameters. In this section we demonstrate the application of the model to assessment of consciousness by measuring the integrated information of the estimated MVARX model. The integrated information theory [27, 45, 46] starts from two self-evident axioms about consciousness: every

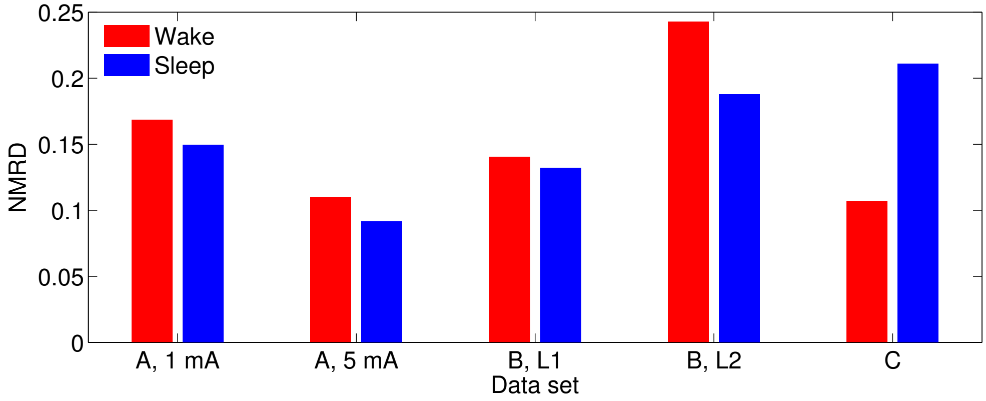


Figure 2.7: Normalized mean-squared response difference (see (2.24)) in each dataset.

experience is one out of many and generates information because it differs in its own way from the large repertoire of alternative experiences; and every experience is one, that is, integrated, because it cannot be decomposed into independent parts. The theory formalizes these notions by postulating that a physical system generates information by reducing uncertainty about which previous states could have caused its present state, and that this information is integrated to the extent that it cannot be partitioned into the information generated by parts of the system taken independently. The theory predicts that integrated information in wakefulness is higher than that in sleep. Integrated information can be measured rigorously in models such as the MVARX model presented here. The integration of information is captured by **A** and **Q** in the MVARX model — **B** only indicates how stimulation enters the network. In this section we contrast integrated information in wakefulness and sleep using a variation on the procedure introduced in [21] for obtaining a bipartition approximation to integrated information in MVAR systems. Our variation is based on use of "effective information" (Kullback-Leibler divergence) [47] in place of the difference in mutual information and ensures that integrated information is always positive [48].

Suppose y_n describes a stable MVAR(p) process:

$$y_n = \sum_{i=1}^{p} \mathbf{A}_i y_{n-i} + \mathbf{w}_n, \tag{2.26}$$

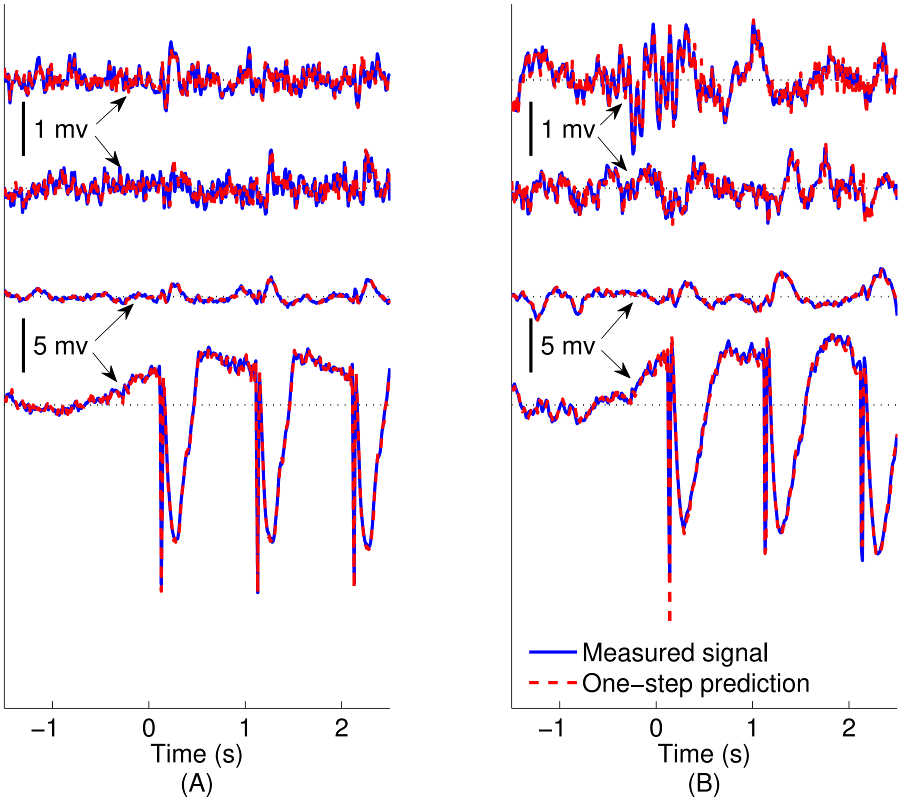


Figure 2.8: Comparison between recorded signal and one-step prediction of subject B when the stimulating site is L2. 1.5 s prestimulus is shown followed by two-and-a-half epochs of evoked data. The black dotted lines in the figures indicate the origin. The model is estimated from data beginning with the fourth epoch. (A) Wake recorded and predicted signals in channels 1, 3, 6, and 8 ordered from top to bottom. (B) Non-REM sleep recorded and predicted signals in channels 1, 3, 6, and 8 ordered from top to bottom.

where $\mathbf{w}_n$ are i.i.d. zero-mean Gaussian noise vectors with covariance $\mathbf{Q}$. Then the MVAR(p) process is wide sense stationary and $\mathbf{y}_n \sim \mathcal{N}(0, \mathbf{\Sigma}(\mathbf{y}))$ with $\mathbf{\Sigma}(\mathbf{y}) = E\{\mathbf{y}_n \mathbf{y}_n^T\}$. Given that the state at time n, $\mathbf{y}_n = \underline{\mathbf{y}}$, the conditional distribution of the state τ samples prior to sample n, $\mathbf{y}_{n-\tau}$, follows

$$\mathbf{y}_{n-\tau}|(\mathbf{y}_n = \underline{\mathbf{y}}) \sim \mathcal{N}(\mathbf{\Gamma}_\tau(\mathbf{y})\mathbf{\Sigma}(\mathbf{y})^{-1}\underline{\mathbf{y}}, \mathbf{\Sigma}(\mathbf{y}_{n-\tau}|\mathbf{y}_n)) \tag{2.27}$$

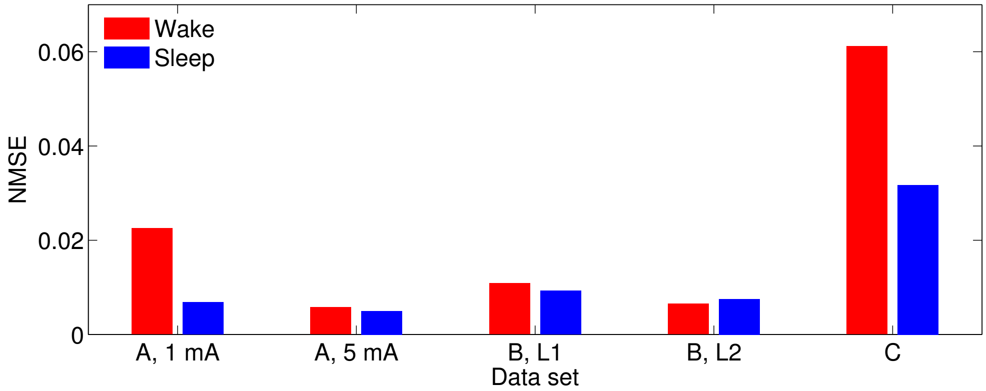


Figure 2.9: Normalized mean-squared one-step prediction error (see (2.25)) in each dataset.

where $\mathbf{\Gamma}_\tau(\mathbf{y}) = E\{\mathbf{y}_{n-\tau}\mathbf{y}_n^T\}$ and

$$\mathbf{\Sigma}(\mathbf{y}_{n-\tau}|\mathbf{y}_n) = \mathbf{\Sigma}(\mathbf{y}) - \mathbf{\Gamma}_\tau(\mathbf{y})\mathbf{\Sigma}(\mathbf{y})^{-1}\mathbf{\Gamma}_\tau(\mathbf{y})^T. \tag{2.28}$$

Given $\mathbf{A}$ and $\mathbf{Q}$, the matrices $\mathbf{\Sigma}(\mathbf{y})$ and $\mathbf{\Gamma}_\tau(\mathbf{y})$ for $\tau = 1, \cdots, \rho$, with $\rho \geqslant p - 1$, are computed as described in [21].

Let the set of the channels be $S = \{1, 2, \cdots, d\}$. A bipartiton $\mathcal{B} = \{M^1, M^2\}$, divides the channels into two mutually non-overlapping and non-empty sub-networks, $S = M^1 \bigcup M^2$. Denote two sub-systems $\mathbf{m}_n^1$ and $\mathbf{m}_n^2$ within which are the measurements in the channels corresponding to the elements in M^1 and M^2 at time n, respectively. Given $\mathbf{\Sigma}(\mathbf{y})$ and $\mathbf{\Gamma}_\tau(\mathbf{y})$, we have $\mathbf{\Sigma}(\mathbf{m}^i) = [\mathbf{\Sigma}(\mathbf{y})]_{M^i,M^i}$ and $\mathbf{\Gamma}_\tau(\mathbf{m}^i) = [\mathbf{\Gamma}_\tau(\mathbf{y})]_{M^i,M^i}$, for $i = 1, 2$. Hence, given the present state, the conditional distribution of the subsystem i at τ samples into the past is given by $\mathbf{m}_{n-\tau}^i|(\mathbf{m}_n^i = \underline{\mathbf{m}}^i) \sim \mathcal{N}(\mathbf{\Gamma}_\tau(\mathbf{m}^i)\mathbf{\Sigma}(\mathbf{m}^i)^{-1}\underline{\mathbf{m}}^i, \mathbf{\Sigma}(\mathbf{m}_{n-\tau}^i|\mathbf{m}_n^i))$, for $i = 1, 2$, where $\mathbf{\Sigma}(\mathbf{m}_{n-\tau}^i|\mathbf{m}_n^i) = \mathbf{\Sigma}(\mathbf{m}^i) - \mathbf{\Gamma}_\tau(\mathbf{m}^i)\mathbf{\Sigma}(\mathbf{m}^i)^{-1}\mathbf{\Gamma}_\tau(\mathbf{m}^i)^T$.

Define the effective information for the system $\mathbf{y}$ over a lag of τ samples under partition $\mathcal{B}$ as (see [21] (0.32))

$$\varphi(\mathbf{y}; \tau, \mathcal{B}) = \frac{1}{2}\log_2\left(\frac{\det(\mathbf{\Sigma}(\mathbf{m}_{n-\tau}^1|\mathbf{m}_n^1)) \cdot \det(\mathbf{\Sigma}(\mathbf{m}_{n-\tau}^2|\mathbf{m}_n^2))}{\det(\mathbf{\Sigma}(y_{n-\tau}|y_n))}\right) \text{ bits.} \tag{2.29}$$

The effective information is the Kullback-Leibler divergence between a system consisting of two mutually independent sub-systems $\mathbf{m}_n^1$ and $\mathbf{m}_n^2$ and the system $\mathbf{y}_n$.

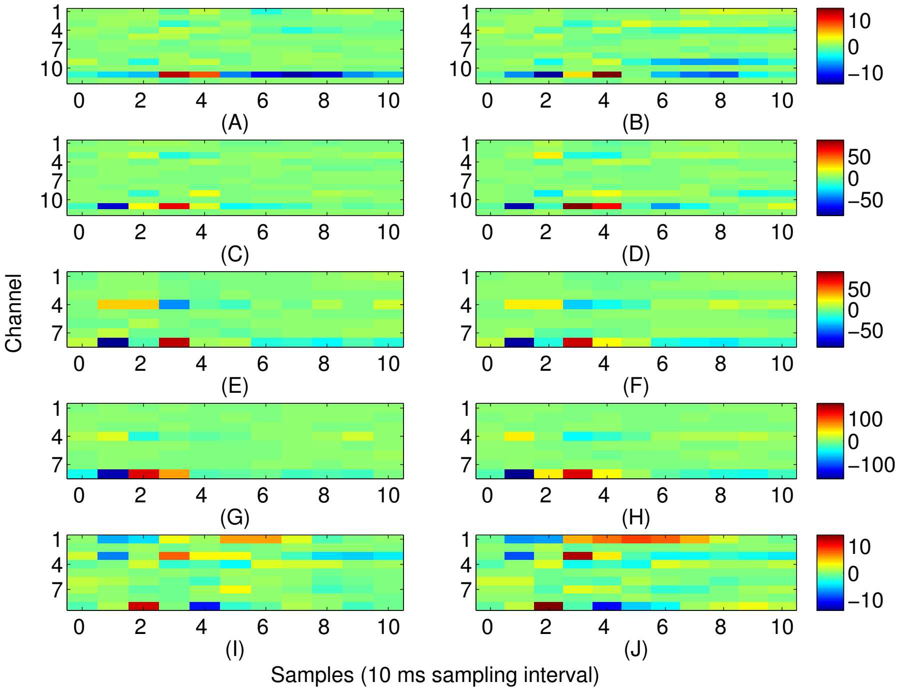


Figure 2.10: Exogenous input filters **B** for each channel as a function of time. The identical colormap is used for each row. (A) Subject A Wake, 1 mA. (B) Subject A Sleep, 1 mA. (C) Subject A Wake, 5 mA. (D) Subject A Sleep, 5 mA. (E) Subject B Wake, Stimulation site L1. (F) Subject B Sleep, Stimulation site L1. (G) Subject B Wake, Stimulation site L2. (H) Subject B Sleep, Stimulation site L2. (I) Subject C Wake. (J) Subject C Sleep.

The integrated information measured at a time difference of τ is defined as

$$\phi(\mathbf{y};\tau) = \varphi(\mathbf{y};\tau,\mathcal{B}^{\mathrm{MIB}}) \tag{2.30}$$

where the minimum information bipartion (MIB) is defined as

$$\mathcal{B}^{\mathrm{MIB}} = \arg\min_{\mathcal{B}} \left(\frac{\varphi(\mathbf{y};\tau,\mathcal{B})}{K_2(\mathcal{B})} \right) \tag{2.31}$$

with

$$K_2(\mathcal{B}) = \min(H(\mathbf{m}_n^1), H(\mathbf{m}_n^2)) \tag{2.32}$$

Table 2.3: p-values of the Wilcoxon rank sum test of whether integrated information in wakefulness and sleep are different.

	Subj A, 1 mA	Subj A, 5 mA	Subj B, L1	Subj B, L2	Subj C
p-value	3.18e-4	0.0012	2.06e-4	0.0068	0.0553

and the differential entropy of $\mathbf{m}_n^i$, $H(\mathbf{m}_n^i)$ is given by

$$H(\mathbf{m}_n^i) = \frac{1}{2}\log_2\left((2\pi e)^{|M^i|}\det(\mathbf{\Sigma}(\mathbf{m}^i))\right). \tag{2.33}$$

Figure 2.11 depicts the integrated information of subject A for stimulus of 5 mA, as the time difference τ varies from 10 ms to 300 ms. The integrated information in wakefulness is higher than that in sleep. In both wakefulness and sleep, the integrated information increases until the time difference is approximately 100 ms and then remains approximately constant. We further used the CV procedures described in Section 2.3 to study the difference between integrated information in wakefulness and sleep. Specifically, we estimated a model from the training set of each CV partition and compute integrated information for each CV partition. This provides M different estimates of integrated information for each data set, where M is the number of CV partitions. We compare the maximum values of the estimates of integrated information for each CV partition in wakefulness and sleep using the Wilcoxon rank sum test, which tests the null (H_0) hypothesis that the measured maximum integrated information values in wakefulness and sleep for all CV partitions are samples from continuous distributions with equal medians, against H_1 that they are not. The p-values of the rank sum test for each conditions are shown in Table 2.3. With the exception of Subject C, all of the cases have p-values below 0.05, and Subject C is only slightly above 0.05. Figure 2.12 depicts the average maximum value of integrated information and average time delay τ at which the maximum value is achieved, where the averaging is done across CV results, and error bars indicates one standard error.

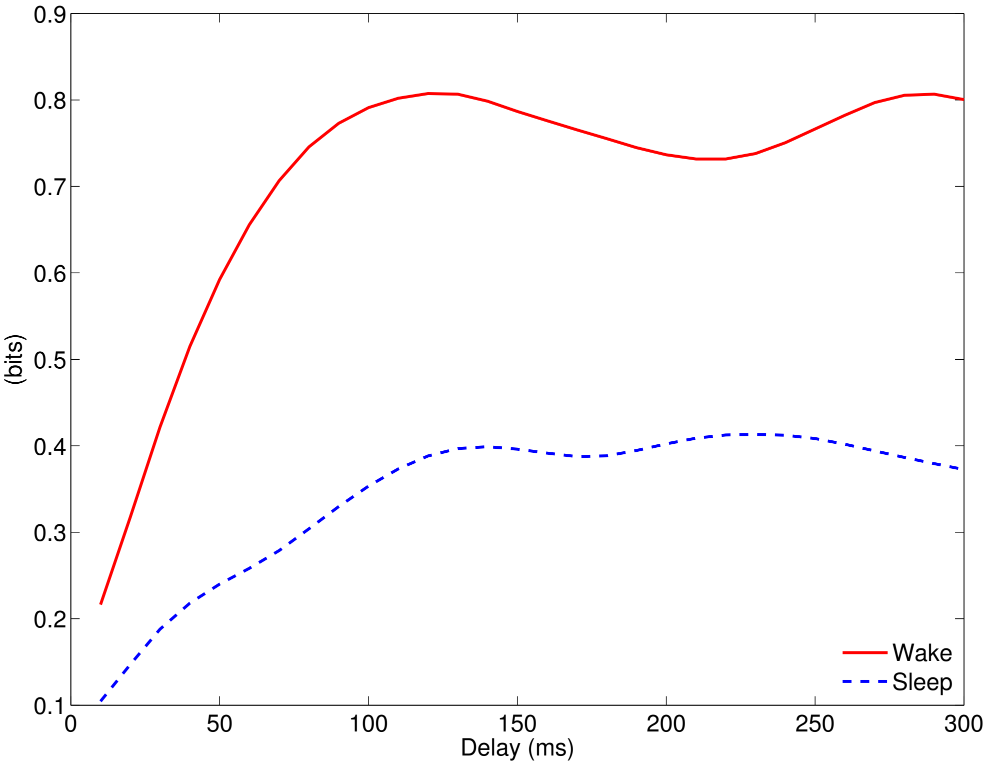


Figure 2.11: Integrated information of subject A, when the stimulation current is of 5 mA.

2.6 Discussion

The results demonstrate the effectiveness of the MVARX model for intracerebral electrical stimulation data. Excellent agreement between measured and modeled evoked responses is found across channels, two stimulus amplitudes, vigilance states, stimulus sites, and subjects (Figs. 2.4–2.7). One-step prediction is used to show that the MVARX model also accurately captures the spontaneous fluctuations in the measured signals (Figs. 2.8 and 2.9). We contrasted the MVARX models with a series of univariate ARX models, one for each channel, to illustrate the importance of accounting for the interaction between cortical signals (Figs. 2.4–2.6). In some channels for some subjects/conditions the univariate ARX model describes the evoked response as well as the MVARX model. However, in general modeling interactions between cortical signals is necessary to capture the measured response.

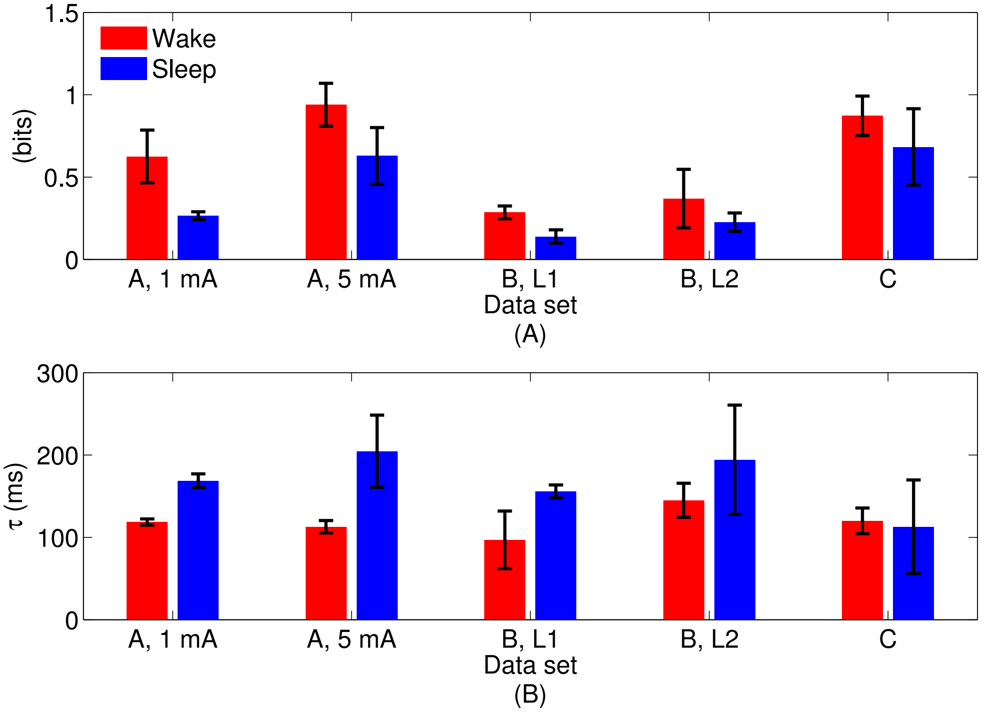


Figure 2.12: (A) Average maximum values of integrated information with error bars indicating one standard error. (B) Average lag at which maximum integrated information is achieved, with error bars indicating one standard error.

For example, in Fig. 2.4 (B) the univariate model fails to model the responses in channels 1, 4, and 7 beyond 200 ms after the stimulation.

The MVARX model explicitly represents both evoked and spontaneous (or background) brain activity using a deterministic input term to capture the effect of stimuli and a random input term to generate spontaneous activity. Stimuli generally give rise to a non-zero mean component in the response that varies with time, i.e., is nonstationary. Conventional approaches to MVAR modeling of cortical event-related potentials (e.g.,[8]), subtract the ensemble mean of the data before processing to avoid the negative effects of the nonstationary mean on the MVAR model. However, subtraction of the ensemble mean significantly reduces the SNR of the data and is not necessary if the exogenous input is properly accounted for in the modeling procedure.

The effect of the stimulus on each recording channel is addressed by applying a separate filter in each channel to the stimulus signal. The filter coefficients are estimated jointly with the autoregressive model parameters from the measured evoked data. This approach accounts for the generally unknown and different characteristics of the transmission paths from the stimulation to each measurement site. The length of the filters (ℓ samples in (2.5)) should be limited based on physiological expectations for the stimulus paradigm. Indeed, the autoregressive coefficients $\mathbf{A}_i$ and filters $\mathbf{b}_i$ are estimated simultaneously and the evoked response ($\mathbf{y}_{n,e}^{(j)}$ in (2.8)) is often much larger than the spontaneous component ($\mathbf{y}_{n,s}^{(j)}$ in (2.7)). If ℓ is set equal to the duration of one epoch of $\mathbf{y}_{n,e}^{(j)}$, then it is possible to perfectly model $\mathbf{y}_{n,e}^{(j)}$ using only the $\mathbf{b}_i$ while setting the $\mathbf{A}_i = 0$. We have shown that the MVARX models are capable of characterizing $\mathbf{y}_{n,s}$ by one-step prediction of data not used to estimate the model (see Fig. 2.8). Moreover, the model describes the dynamics in $\mathbf{y}_{n,e}$, as was shown in Figs 2.4 – 2.6.

In order to define a practical value for ℓ we refer to previous electrophysiological studies on intracerebral evoked potentials [40, 49, 50]. In these studies from Matsumoto and colleagues the possible generator mechanisms of intracerebral potentials evoked by direct electrical stimulation are thoroughly discussed. In all of these studies it has been shown that the duration of the "purely evoked" response expires within 100 ms. Based on these results and our 100 Hz sampling frequency we set $\ell = 10$. The 100 ms value is also consistent with our data. Indeed, the first 100 ms post-stimulus of the evoked waveforms exhibit quite different character than later portions. Typically the initial 100 ms of the measured response contain relatively sharp, high frequency waveforms, while later portions of the response have a smoother, lower frequency behavior. This suggests two regimes in the modeling process. The exogenous input filters account for the sharp initial response, as evident by the filter impulse responses shown in Fig. 2.10. Channels having relatively large impulse response tend to rapidly transition from negative to positive maxima over one or two samples, consistent with the sharp features in the early portions of the evoked response. These sharp inputs to the channels are smoothed by the autoregressive component of the model to obtain the later portions of the response. The filter responses depicted in Fig. 2.10 decay to relatively small values by the 10-th lag (100 ms) and generally contain most of their energy in the first through sixth lags, that is between 10 and 60 ms. This further supports the choice of $\ell = 10$.

The energy transmission characteristics shown in Fig. 2.10 are consistent with physiological expectations for modeling stimulation of fibers of passage. There is general consistency between wakefulness and sleep in all subjects (Fig. 2.10, left column vs. right column) even though the evoked responses differ markedly (Fig. 2.4 vs. Fig. 2.5); channels with strong and weak responses are the same in wakefulness and sleep, and the shape of the responses in each channel are generally very similar. The subtle differences between wakefulness and sleep may be due to changes in neural excitability. Comparing 1 mA and 5 mA stimulation in subject A (Fig. 2.10 (A–B) and (C–D)) reveals that channel 11 has the strongest response in both stimulation levels and the strength of the response increases roughly by a factor of five, consistent with the factor of five change in the stimulation level. This is because we used the trigger signal to represent the exogenous input without adjusting its amplitude. However, the shape of the response in channel 11 differs slightly, with the 5 mA case having reduced latency by approximately 10 ms and a higher frequency response reflected by the sharper, shorter duration of the filter. This suggests that the higher stimulus level is associated with a faster response. The two stimulation sites L1 and L2 in subject B (Fig. 2.10 (E–F) and (G–H)) both involve channels 8 and 4 as the strongest response, suggesting similar fibers of passage are excited at the two sites. However, the overall gain differs by a factor of two and the shape of the response in channel 8 and 4 differ, especially in wakefulness. Subject C (Fig. 2.10 (I–J)) exhibits multiple channels with strong linkage to the stimulus site.

Our MVARX approach assumes the dynamic interactions between evoked and spontaneous cortical signals follow the same model, that is, both evoked and spontaneous activity are described by one set of $\mathbf{A}_i$. The excellent one-step prediction performance in the pre-stimulus interval of Fig. 2.8 combined with the high quality fitting of the evoked responses suggests this is a reasonable assumption, at least for these particular data sets. This approach also assumes that the measured signal is the sum of the evoked and spontaneous activity.

The windowed median filtering procedure successfully eliminated the volume conduction artifact without changing the measured signal at and beyond 20 ms post-stimulus. The outlier detection strategy only eliminates epochs that have significant deviation from the average evoked response. Both of these strategies significantly improve model fidelity to the measured data. Seven times as many outlier epochs were identified in sleep than in wakefulness, likely due to the presence of occasional slow waves during an epoch. However, in seven of the 10 data sets we analyzed 28 or more of the 30 available epochs, which indicates our artifact detection procedure is not overly restrictive. Subject A had the most outlier epochs and in the worst case (5 mA, sleep) our procedure eliminated eight of the possible 30 epochs. The CV strategy for choosing MVAR model order is effective, as demonstrated by the fidelity of the model evoked responses (Figs. 2.4–2.7) and the ability of the models to accurately perform one-step prediction on pre-stimulus data (Fig. 2.8). Outlier rejection helps the data meet the stationarity assumption of the MVARX model. While it is unlikely that the data is truly stationary, the accuracy with which the model describes the data and the whiteness of the residuals suggests that the stationarity assumption is reasonable.

As a proof of concept application, we used the MVARX model to assess changes in the level of information integration between wakefulness and deep sleep in human subjects. Using a simple, bipartition approximation we found that, as predicted by theoretical considerations [27, 30], integrated information is higher in wakefulness than sleep for each subject/condition, supporting the notion that integrated information reflects the capacity for consciousness. We note that the integrated information results presented here only apply to the recordings analyzed. Analysis of the dependence of integrated information on recording coverage is beyond the scope of this paper. Our findings indicate that the human cerebral cortex is better suited at information integration — being both functionally specialized and functionally integrated — when awake and conscious. In contrast, when consciousness fades in deep sleep, the parameters of the system change in such a way that information integration is diminished, in line with theoretical predictions [27] and consistent with qualitative evidence obtained from experiments employing transcranial magnetic stimulation and high density EEG [51]. We also found that the lag at which the maximum level of information integration is attained is consistently longer in sleep than wakefulness. Maximum information integration in wakefulness occurred at lags of 30 to 110 ms, while those in sleep were from 70 to 140 ms longer, consistent with the increased low frequency activity of sleep.

3 SPARSE MULTIVARIATE AUTOREGRESSIVE MODELS WITH EXOGENOUS INPUTS FOR MODELING INTRACEREBRAL RESPONSES TO DIRECT ELECTRICAL STIMULATION OF THE HUMAN BRAIN

3.1 Introduction

The remarkable power of the human brain is widely believed to be due to a delicate balance between functional segregation and integration of cortical systems, that is, network properties. Multivariable autoregressive (MVAR) models have been widely applied to magneto-/electroencephalography (M/EEG) to study effective connectivity [52] between cortical regions from scalp EEG [53] or between recording sites from intracranial data [54] [1]. In this paper we employ MVAR-based Granger causality [55] as metric for assessing effective connectivity associated with direct electrical stimulation of the cortex. An overview of other models and metrics for accessing effective connectivity is given in [52].

Evidence that brain networks have the small world property and hub topologies [56] suggests that brain networks are likely not fully connected and motivates the sparse MVAR time series model in [57]. In the model, we partition the autoregressive coefficients into groups with each group containing parameters associated with the individual connections between nodes. An ℓ_1 penalty on the groups is added to the squared error to encourage MVAR models with a sparse number of connections between nodes. The ℓ_1 regularization not only provides information about the structure of the brain network model but also facilitates estimation of large-scale network models from limited data. The group penalty on the MVAR coefficients leads to a group lasso [58] procedure for identifying the MVAR model. The self-connected group lasso (SCGL) [57] is the focus of this paper. It assumes each node is driven by its own past and does not penalize self connections.

During pre-surgical evaluation of drug-resistant epileptic patients, direct electrical stimulation of the brain is systematically performed for diagnostic purposes to identify the epileptogenic zone [59]. Electrical stimulation generates a time-varying

[1] ©2013 IEEE. Reprinted, with permission, from [J.-Y. Chang, A. Pigorini, F. Seregni, M. Massimini, L. Nobili and B. Van Veen, "Sparse multivariate autoregressive models with exogenous inputs for modeling intracerebral responses to direct electrical stimulation of the human brain," 2013 Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, 2013, pp. 803-807.].

evoked response at the recording sites. This violates the stationary MVAR model assumption that the process is of constant mean. In this paper we model the response of the brain to electrical stimulation using stationary MVAR models with an exogenous input (MVARX) derived from the stimulus characteristics. The advantage of the MVARX model is that it does not require subtraction of the mean (as in [60]) and consequent reduction in signal-to-noise ratio (SNR) or the complication of time-varying models for describing the evoked response [9]. The MVARX model captures both the mean evoked response and the background activity present during the recordings.

In this paper we adapt the SCGL to the the MVARX model by adding groups of coefficients associated with the exogenous stimulation to the set of penalized groups. We use this modified form of SCGL to find a sparse MVARX model for 31 channels of intracranial measurements from a human subject. The sparse model is then compared to a full MVARX model in terms of the model predicted responses and one-step prediction errors. We compare sparse MVARX models for the subject in either wakefulness or non-rapid eye movement (NREM) sleep using two different stimulation sites. Common and differing characteristics of the networks are then discussed in terms of the set of active connections and Granger causality.

This paper is organized as follows. Section 2 describes the sparse MVARX model and the SCGL problem. Section 3 describes the data. Section 4 presents results for sparse network modelling under different conditions. This paper then concludes with a discussion of the results in Section 5. For notation, boldface lower and upper case symbols represent vectors and matrices, respectively, while superscript T denotes matrix transpose and superscript -1 denotes matrix inverse.

3.2 Methods

Suppose that the considered process is stationary and can be described by the MVARX model of order (p, ℓ) [55]:

$$\mathbf{y}_t^{(j)} = \boldsymbol{\nu} + \sum_{i=1}^{p} \mathbf{A}_i \mathbf{y}_{t-i}^{(j)} + \sum_{i=0}^{\ell} \mathbf{b}_i x_{t-i}^{(j)} + \mathbf{w}_t^{(j)} \tag{3.1}$$

for time $t = t_0 + 1, \cdots, T_s$, and epoch $j = 1, \cdots, J$, where $t_0 = \max(p, \ell)$, $\mathbf{y}_t^{(j)} \in \mathbb{R}^K$ is the recorded data from K different sites at time $t = 1, \cdots, T_s$ and epoch j, $x_t^{(j)}$ is the input at time t and epoch j, $\boldsymbol{\nu} \in \mathbb{R}^K$ is a constant and $\mathbf{w}_t^{(j)} \in \mathbb{R}^K$ is the unmodeled residual. We assume that in each epoch, samples $\mathbf{y}_1^{(j)}, \mathbf{y}_2^{(j)}, \cdots, \mathbf{y}_{t_0}^{(j)}$ and inputs $x_1^{(j)}, x_2^{(j)}, \cdots, x_{t_0}^{(j)}$ are also available. The residuals at different time samples are assumed to be uncorrelated and identically distributed with mean 0 and covariance matrix $\mathbf{Q}$. The matrices $\mathbf{A}_i = [a_{m,n}(i)]_{m=1,n=1}^{K,K}$ contain autoregressive coefficients describing the influence of channel n on channel m at lag i, and the vectors $\mathbf{b}_i = [b_m(i)]_{m=1}^K$ contain filter coefficients from the exogenous stimulus to channel m at lag i. Let $\mathbf{z}_{t-1}^{(j)} = [1, (\mathbf{y}_{t-1}^{(j)})^T, (\mathbf{y}_{t-2}^{(j)})^T, \cdots, (\mathbf{y}_{t-p}^{(j)})^T, x_t^{(j)}, x_{t-1}^{(j)}, \cdots, x_{t-\ell}^{(j)}]^T$, $\mathbf{A} = [\mathbf{A}_1, \cdots, \mathbf{A}_p]$, $\mathbf{B} = [\mathbf{b}_0, \cdots, \mathbf{b}_\ell]$, $\mathbf{Y}_j = [\mathbf{y}_{t_0+1}^{(j)}, \cdots, \mathbf{y}_{T_s}^{(j)}]$, $\mathbf{Z}_j = [\mathbf{z}_{t_0}^{(j)}, \cdots, \mathbf{z}_{t_0+T_s-1}^{(j)}]$, $\boldsymbol{W}_j = [\mathbf{w}_{t_0+1}^{(j)}, \cdots, \mathbf{w}_{T_s}^{(j)}]$ and $\boldsymbol{\Theta} = [\boldsymbol{\nu}, \mathbf{A}, \mathbf{B}]$. The MVARX model then can be recast as $\mathbf{Y} = \boldsymbol{\Theta}\mathbf{Z} + \boldsymbol{W}$ with $\mathbf{Y} = [\mathbf{Y}_1, \cdots, \mathbf{Y}_J]$, $\mathbf{Z} = [\mathbf{Z}_1, \cdots, \mathbf{Z}_J]$, and $\boldsymbol{W} = [\boldsymbol{W}_1, \cdots, \boldsymbol{W}_J]$.

The least square estimate of the m-th row of the full MVARX model can obtained by solving: $\hat{\boldsymbol{\Theta}}_{m,:}^T = \arg\min_{\mathbf{c}_m} \frac{1}{T} \|\mathbf{Y}_{m,:}^T - \mathbf{Z}^T \mathbf{c}_m\|_2^2$ where $\mathbf{Y}_{m,:}$ is the m-th row of $\mathbf{Y}$ and $T = \sum_{j=1}^J T_s - Jt_0$. Then $\hat{\boldsymbol{\Theta}} = [\hat{\boldsymbol{\Theta}}_{1,:}^T, \cdots, \hat{\boldsymbol{\Theta}}_{K,:}^T]^T$ and $\hat{\mathbf{Q}} = (\mathbf{Y} - \hat{\boldsymbol{\Theta}})(\mathbf{Y} - \hat{\boldsymbol{\Theta}})^T / T$. The SCGL procedure adds an ℓ_1/ℓ_2 penalty to the least squares problem and takes the form

$$\begin{aligned}(\hat{\boldsymbol{\Theta}}_{m,:}^{\text{SCGL}}(\lambda))^T = &\arg\min_{\mathbf{c}_m} \frac{1}{T} \|\mathbf{Y}_{m,:}^T - \mathbf{Z}^T \mathbf{c}_m\|_2^2 \\ &+ \lambda \Big(\sum_{n=1, n \neq m}^{K} \sqrt{p} \|\mathbf{a}_{m,n}\|_{\mathbf{D}_n} + \sqrt{\ell + 1} \|\mathbf{b}_{m,:}^T\|_{\mathbf{D}_x} \Big) \end{aligned} \tag{3.2}$$

where $\mathbf{a}_{m,n} = [a_{m,n}(1), \cdots, a_{m,n}(p)]^T$ is the group of autoregressive coefficients connecting channel n to channel m, $\mathbf{b}_{m,:}$ is the m-th row of $\mathbf{B}$ and $\|\mathbf{u}\|_{\mathbf{D}} = \sqrt{\mathbf{u}^T \mathbf{D} \mathbf{u}}$. Note that here $\mathbf{a}_{m,n}$ and $\mathbf{b}_{m,:}$ are not separate variables, but are subsets of c_m. The notation in (3.2) is intended to explicitly indicate the nature of the penalty on the model parameters. We have $\mathbf{D}_n = \sigma_n^2 \mathbf{I}_p$ where $\sigma_n^2 = \|\mathbf{Y}_{n,:}\|_2^2 / T$ is the measurement power in recording n and $\mathbf{D}_x = \sigma_x^2 \mathbf{I}_{\ell+1}$ with $\sigma_x^2 = \sum_{j=1}^J \sum_{t=t_0+1}^{T_s} (x_t^{(j)})^2 / T$ being the stimulation power. In the group lasso literature, the signals are often assumed to be normalized to have equal power before solving the group lasso problem. The parameters solved with the group lasso procedure are then rescaled to account for the normalization (see [57] and references therein). This procedure is equivalent to substituting the norm $\|\mathbf{u}\|_{\mathbf{D}}$ for the ℓ_2-norm in the classical group lasso objective, as is done here in (3.2) [61]. The terms $\sqrt{p}$ and $\sqrt{\ell+1}$ are included to adjust for the group sizes [61]. The penalty parameter λ determines the level of sparsity of the estimated model – as λ increases, more groups of coefficients are set to zero. The SCGL problem can be solved with group lasso algorithms. Discussion of appropriate algorithms can be found in [61] and references in [57].

It is well-known that the group lasso penalty shrinks the nonzero coefficients toward zero. This could lead to biased inference of network properties so we introduce a debiasing step in our analysis. Once the active set is identified by the SCGL estimate, we solve a least squares problem involving only the coefficients belonging to the active set. This gives us the sparse model parameterized by λ, $\hat{\boldsymbol{\Theta}}(\lambda)$ and $\hat{\mathbf{Q}}(\lambda)$.

In practice, (p, ℓ) and λ are not given and are determined with model selection procedures or physiological knowledge. We postpone the rationale for our selection of ℓ to Section III. p and λ are chosen with cross-validation (CV). Our CV criterion focuses both on the minimization of the one-step prediction error – a measure of the model's ability to track the sample-to-sample and epoch-to-epoch fluctuations in the data, and the average response error – a measure of the quality of the model's response to the stimulus. In performing CV, we partition the epochs of available data into I non-overlapping sets $S_1, \cdots, S_I$ of approximately the same duration. For each subset S_i, we "train" a model from all subsets except for S_i and obtain a model denoted as $\hat{\boldsymbol{\Theta}}_i(p, \lambda)$. This model is then tested on the set S_i to obtain the two error measures. The one-step prediction error at time t, $\boldsymbol{e}_t^{(j)}(i, p, \lambda)$, is the difference between the recorded data $\mathbf{y}_t^{(j)}$ and the one-step prediction made by $\hat{\boldsymbol{\Theta}}_i(p, \lambda)$ using the t_0 samples prior to time t, that is, $\boldsymbol{z}_{t-1}^{(j)}$ so we have $\boldsymbol{e}_t^{(j)}(i, p, \lambda) = \mathbf{y}_t^{(j)} - \hat{\boldsymbol{\Theta}}_i(p, \lambda)\boldsymbol{z}_{t-1}^{(j)}$.

In our data, the stimulation is applied every T_s samples. We define one epoch as the T_s samples corresponding to a single stimulation interval. The average evoked response over set S_i is given by $\bar{\mathbf{y}}_t(i) = M_i^{-1} \sum_{j \in S_i} \mathbf{y}_t^{(j)}$ for $t = 1, \cdots, T_s$, where M_i is the number of epochs in S_i. The average model response over set S_i is given by $\hat{\bar{\mathbf{y}}}_t(i, p, \lambda) = M_i^{-1} \sum_{j \in S_i} \mathbf{y}_{t,e}^{(j)}(i, p, \lambda)$ where $\mathbf{y}_{t,e}^{(j)}(i, p, \lambda)$ is the output of model $\hat{\boldsymbol{\Theta}}_i(p, \lambda)$ in response to the stimulus alone, that is, when $\mathbf{w}_t^{(j)} = 0$. The average response error is given as $\boldsymbol{\epsilon}_t(i, p, \lambda) = \bar{\mathbf{y}}_t(i) - \hat{\bar{\mathbf{y}}}_t(i, p, \lambda)$.

The CV score is defined as a weighted combination of the one-step prediction and average response errors averaged over all I subsets

$$\mathrm{CV}(p,\lambda) = \frac{1}{I}\sum_{i=1}^{I}\left[\frac{\mathrm{CV}_e(i,p,\lambda)}{w_e} + \frac{\mathrm{CV}_\epsilon(i,p,\lambda)}{w_\epsilon}\right] \tag{3.3}$$

where the mean square one-step prediction error component is given as $\mathrm{CV}_e(i,p,\lambda) = {M_i}^{-1}\sum_{j\in S_i}\frac{1}{T_s-t_0}\sum_{t=t_0+1}^{T_s}\|\mathbf{e}_t^{(j)}(i,p,\lambda)\|_2^2$ and the mean square response error component is defined as $\mathrm{CV}_\epsilon(i,p,\lambda) = \frac{1}{T_s}\sum_{t=1}^{T_s}\|\boldsymbol{\epsilon}_t(i,p,\lambda)\|_2^2$. The weights w_e and w_ϵ vary the emphasis between the one-step prediction error and average response error. In the analysis below, we set w_e and w_ϵ to the medians of $\mathrm{CV}_e(i,p,\lambda)$ and $\mathrm{CV}_\epsilon(i,p,\lambda)$, respectively, for $i = 1,\cdots,I$ and all p and λ considered. This approach places approximately equal emphasis on the two errors. p and λ are chosen as values that minimize $\mathrm{CV}(p,\lambda)$.

The MVARX model assumes the residuals are uncorrelated and identically distributed, so we use the whiteness test proposed in [62] to check model consistency. A model is selected by the CV criterion only if it passes the whiteness test.

3.3 Data

The data we analyze in this paper is collected from a subject with long-standing drug resistant focal epilepsy. The patient was a candidate for surgical removal of the epileptic focus. The decision on stimulation sites and duration of implantation was made entirely on clinical needs. A train of 30 electrical pulses of strength 5 mA are applied at intervals of 1 s. The subject is in either wakefulness or stage 4 of NREM sleep. The 31 channels analyzed are located in 19 different cortical areas. We have no access to the stimulation channel due to physiological and recording technology limitations. The raw data is sampled at 1000 Hz and is low pass filtered by an FIR filter with passband-edge of 48 Hz and stopband-edge of 49.9 Hz to eliminate 50 Hz powerline contamination, and downsampled by 10 to a sampling rate of 100 Hz. Nonphysiological artifacts and outliers are excluded from our analysis based on the windowed median filter and outlier rejection procedures discussed in [63].

We choose ℓ to be 10, that is, 100 ms, based on the discussion of the possible generator mechanisms of intracerebral potentials evoked by direct electrical stimulation [64]. Their results show that the duration of the “purely evoked” responses last no longer than 100 ms. This choice is further discussed in [63].

3.4 Results

In this section we present the results of the sparse MVARX model for wakefulness and sleep and for two different stimulation sites labeled L15 and R14. The number of epochs rejected are 7 for L15 wakefulness, 7 for R14 in wakefulness, 4 for L15 in sleep, and 7 for R14 in sleep.

We search over model orders $p = 4, 8, \cdots, 32$ and penalty parameters $-7.5 \geqslant \lambda \geqslant -36$ dB. Use of a logarithmic scale is typical for λ [65]. The largest value of λ is chosen as the one for which all coefficient groups are zero, and smallest value as that for which none of the coefficient groups are zero. The selected p and λ correspond to the model that gives the least CV score, as discussed in Section II, among the models passing the whiteness test. The value λ = -19.5 dB is selected in each of the four cases. The order of the models are $p = 16$ for L15 in wakefulness, $p = 16$ for R14 in wakefulness, $p = 12$ for L15 in sleep and $p = 16$ for R14 in sleep, respectively. Both full and sparse models use the same order p.

We compare the prediction performance of the sparse model to that of the full model to show its effectiveness. Panels (A) and (B) of Fig. 1 show the CV evoked responses (green dash-dot lines) and the CV model responses of the sparse model (blue solid lines) and the full model (red dashed lines). That is, the curves indicate the average over $\overline{\mathbf{y}}_t(i)$ or $\hat{\overline{\mathbf{y}}}_t(i, p, \lambda)$ for different subsets S_i and the error bars indicate the standard error of the mean. The upper two traces are shown using a scale of 1 mv and the bottom two are using a 5-mv scale. In panels (C) and (D) we present the channel-wise normalized squared one-step prediction error (NMSE), which, for subset S_i, is defined as $M_i^{-1} \sum_{j \in S_i} \frac{1}{T_s - t_0} \sum_{t=t_0+1}^{T_s} (e_{m,t}^{(j)}(i, p, \lambda))^2 / \sigma_m^2$. Here $e_{m,t}^{(j)}(i, p, \lambda)$ is the m-th component of $\mathbf{e}_t^{(j)}(i, p, \lambda)$ and σ_m^2 is the average power in channel m. The NMSE of a zero model would be approximately 1, hence the NMSE quantifies the relative mis-predicted portion of the measured signals. The bars show the average NMSE over CV subsets and error bars show the standard error of the mean. The channel-wise NMSE varies from 0.02 to 0.4. The overall NMSE as a ratio of the mean squared one step prediction error in all channels to the power in all channels is 0.08 for the sparse model in wakefulness, 0.10 for the sparse model in sleep, and 0.09 for the full model in wakefulness, and 0.11 for the full model in sleep.

Panels (E) and (F) show the channel-wise normalized squared response difference (NMRD), which is defined as $\sum_{t=1}^{T_s}(\epsilon_{m,t}(i,p,\lambda))^2/\sum_{t=1}^{T_s}(\overline{y}_{m,t}(i))^2$, for CV subset S_i, where $\epsilon_{m,t}(i,p,\lambda)$ and $\overline{y}_{m,t}(i)$ are respectively the m-th components of $\epsilon_t(i,p,\lambda)$ and $\overline{\mathbf{y}}_t(i)$. The bars indicate the average NMRD over CV subsets and error bars show standard error of the mean. The channel-wise NMRD varies from 0.2 to 0.75. The overall ratio of the mean squared response difference in all channels to the power of the evoked response in all channels is 0.34 for the sparse model in wakefulness, 0.43 for the sparse model in sleep, 0.39 for the full model in wakefulness, and 0.41 for the full model in sleep.

In Fig. 3.2, we show the active connections in all four sparse MVARX models. The red-edged blocks depict channels located in a common cortical area. For instance, channels 6 to 9 are located in the same area. Table 1 presents the ratio of the number of active connections to the total number of connections (excluding self connections), the ratio of the number of active connections within a cortical regions to the total number possible, and the ratio of the number of active connections between cortical regions to the total number possible. The results suggest that sleep involves a sparser model and that a greater percentage of connections are active within cortical regions than between them.

Fig. 3.3 depicts the Granger causality connectivity matrix between the 19 cortical areas sampled by the 31 electrodes. Node i is said to Granger cause node j if the inclusion of node i in the model decreases the prediction error in node j, compared to the model that excludes node i. This notion is sometimes called conditional Granger causality for networks with more than three nodes. The Granger causality measures are computed with the procedure described in [66] from the sparse MVARX models. For example the Granger causality from region 2 to region 4 reflects the extent to which past values in channels 2 and 3 decrease the prediction errors in channels 6-9. We note that Granger causality is computed using only $\hat{\mathbf{A}}$ and $\hat{\mathbf{Q}}$; the exogenous input parameters $\hat{\mathbf{B}}$ do not play a role. Also, the stimulation is closest to region 19 (channel 31) in the L15 case and to region 12 (channel 22) in the R14 case.

3.5 Discussion

We have modified the SCGL of [57] to estimate sparse MVARX models of the cortical interactions excited by direct current stimulation of the cortex. The parameters in the MVARX model are partitioned into non-overlapping groups such that each group contains parameters belonging to a particular connection between channels. The SCGL penalizes all connections except for the self connections, and produces an MVARX model that is sparsely connected, yet best fits the observed data. The level of sparsity is controlled by a penalty parameter λ. Increasing λ produces a sparser model. Sparse models of brain activity are motivated by small world and other properties [56] suggesting a balance between integration and segregation.

The sparse MVARX model is determined by the data and three parameters: λ, which controls the level of sparsity, ℓ, which controls the duration of the stimulus effects, and p, the memory of the model. We choose ℓ based on physiological considerations reported in the literature and use a CV procedure to select the p and λ that attain the lowest CV score among the models with residuals passing a whiteness test. We compare both the one-step prediction error performance and the mean response error to that of a fully connected MVARX model. We consider two different stimulation locations for both wakefulness and NREM sleep in an epilepsy patient. We observe that NMSE for sparse models are smaller than those of the full models, both on a channel-by-channel basis and overall. In some cases the sparse models produce slightly worse NMRD than the full models, although the differences in the mean responses are subtle as shown in panels (A) and (B) of Fig. 1.

It is not possible to know the true underlying network. However, the sparse MVARX models are consistent with physiological evidence on several levels. First of all, we expect channels within a given cortical region to be more densely connected than those between regions. This property is evident in all cases as shown in Table 3.1. While the percentage of active within region connections is larger overall, some cortical regions are more densely connected within themselves than others. For example, region 11 (channels 18-21) has many more active internal connections than region 4 (channels 6-9) in all four scenarios. A second physiological property evident in our sparse MVARX models is reduced connectivity in sleep [4]. The results in Table 3.1 and Figs. 1 and 2 indicate that the MVARX models for sleep are less well connected in terms of the number of active connections and the richness of the Granger causality representations.

Finally, there is a reasonable level of consistency across the four cases in the nature of the networks as illustrated in Figs. 2 and 3. Clearly the anatomical connectivity is the same in all cases, which supports a common network structure.

Table 3.1: Ratio of Number of Active Connections to Total Number of Possible Connections for Different Scenarios and Combinations.

	Overall	Within common regions	Between different regions
L15 in Wake	0.45	0.67	0.44
R14 in Wake	0.41	0.75	0.40
L15 in Sleep	0.29	0.53	0.28
R14 in Sleep	0.34	0.75	0.33

However, we expect significant differences in effective connectivity between wake and sleep based on previous research [4, 63]. Also the similarity between networks identified from different stimulation sites has not yet been studied as best we know. In spite of these caveats, there are a number of common features. For example, Fig. 2 suggests that there are relative few active connections to channels 8-11. In Fig. 3 we see connections from region 7 to 8 and 8 to 7 in all four cases. There is also a cluster of connections between regions 9-11 that is similar across all four cases. There appear to be stimulation dependent effects in the networks identified using this procedure. Stimulus location L15 is very close to region 19 while stimulus location R14 is very close to region 12. We see elevated connectivity from regions 19 and 12 and suppressed connectivity to regions 19 and 12 for stimuli L15 and R14, respectively. This could reflect either physiological effects of the stimulation changing the nature of the network or be an artifact of the modeling procedure. From a model perspective, signals recorded in regions adjacent to the stimulus generally have a very large evoked response component that is well described using only self-connections, hence the absence of connections to these channels. Furthermore, the large nature of the evoked response makes it well suited to predicting the evoked responses in other channels, which leads to the strong apparent connections to other regions.

The SCGL sparse MVARX model shows significant promise for network discovery using intracranial EEG with direct electrical current stimulation and warrants further exploration.

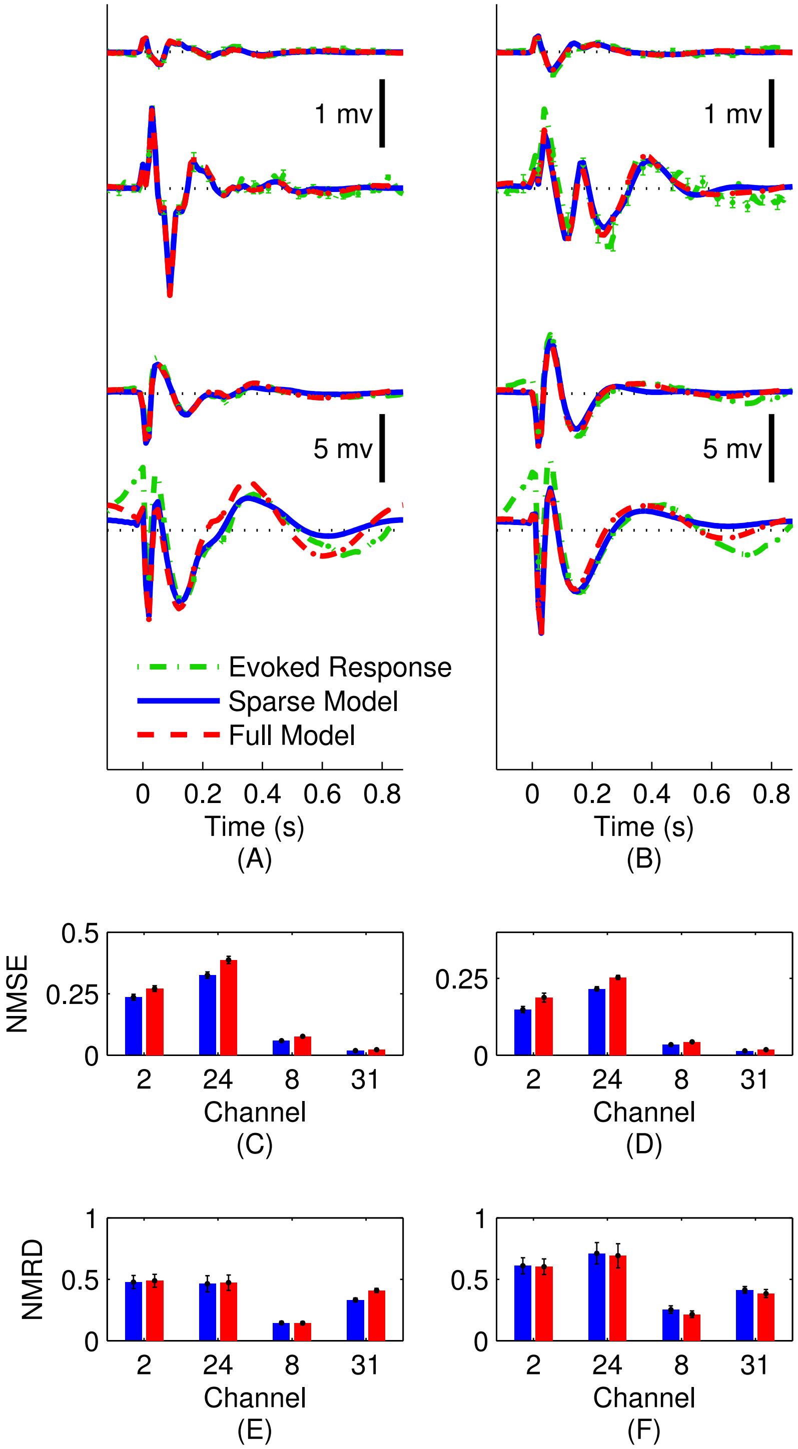


Figure 3.1: (A) Average evoked and model responses of channels 1, 24, 8, and 31 in wakefulness when stimulating from L15, (B) Average evoked and model responses of channels 1, 24, 8, and 31 in sleep when stimulating from L15, (C) Channel-wise normalized mean squared prediction error (NMSE) in wakefulness when stimulating from L15, (D) Channel-wise NMSE in sleep when stimulating from L15, (E) Channel-wise normalized mean squared response difference (NMRD) in wakefulness when stimulating from L15, (F) Channel-wise NMRD in sleep when stimulating from L15.

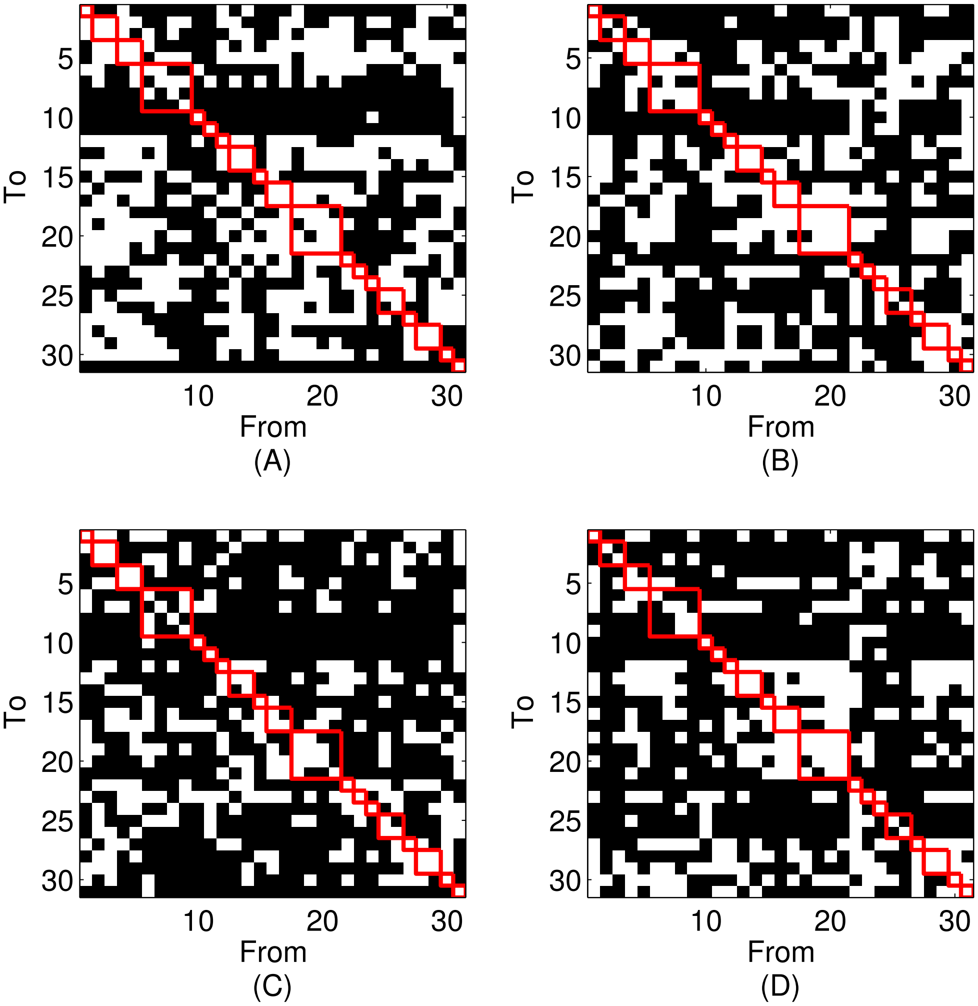
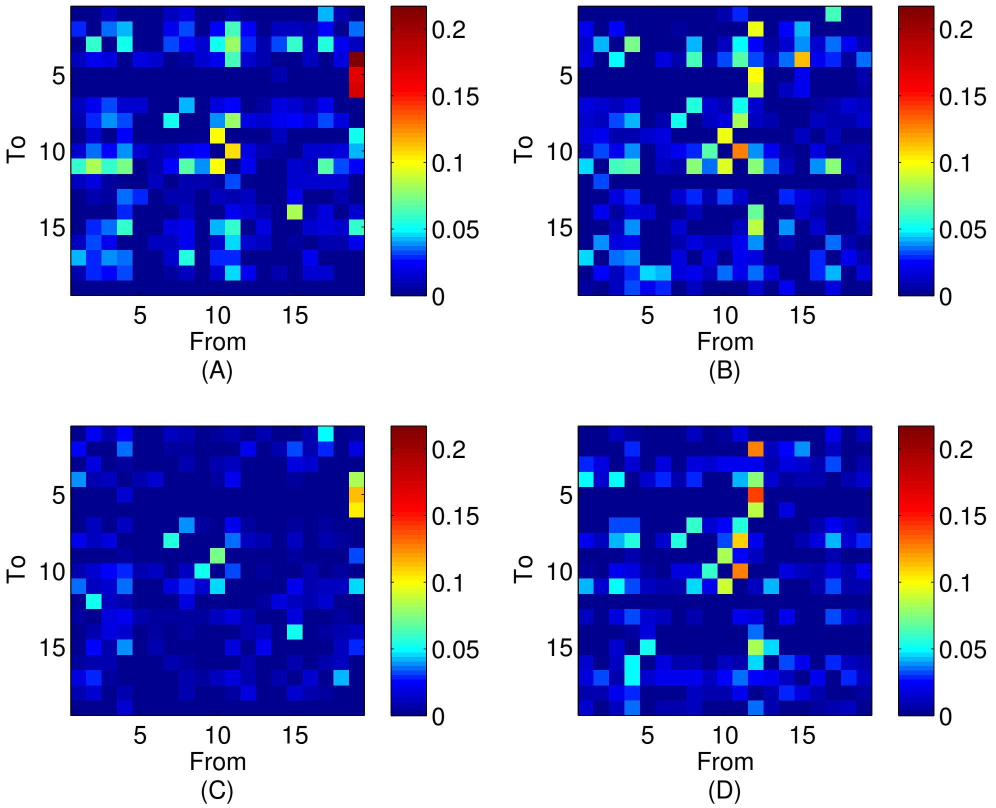


Figure 3.2: Connectivity matrices for the sparse models. Black squares denote that the connection between the specified channels is turned off. Red borders indicate channels located in the same cortical region. (A) L15 in wakefulness, (B) R14 in wakefulness, (C) L15 in sleep, (D) R14 in sleep.



(A)

(B)

(C)

(D)

Figure 3.3: Granger causality connectivity matrices between 19 cortical regions. (A) L15 in wakefulness, (B) R14 in wakefulness, (C) L15 in sleep, (D) R14 in sleep.

4 ASSESSING RECURRENT INTERACTIONS IN CORTICAL NETWORKS: MODELING EEG RESPONSE TO TRANSCRANIAL MAGNETIC STIMULATION

4.1 Introduction

The development of multichannel transcranial magnetic stimulation (TMS)-compatible electroencephalography (EEG) amplifiers [67, 68, 69] has recently opened the possibility of recording the electrical response of the human brain to a direct cortical stimulation [1]. TMS/high-density electroencephalography (hd-EEG) stimulates and records directly from the cerebral cortex, while by-passing sensory pathways and motor pathways. Unlike traditional sensory-evoked potentials and TMS-evoked muscle potentials, this approach does not depend on the integrity/status of sensory and motor systems and can be applied to directly assess changes in cortical reactivity and cortico-cortical connectivity in physiological and pathological conditions. For example, perturbing different cortical targets with TMS in healthy, awake humans always triggers a complex, compound response which involves a distributed set of cortical areas and lasts for about 300 ms. In contrast, during non-rapid eye movement (NREM) sleep the same stimulus elicits a much simpler response of comparable duration [4]. This simpler response to TMS has also been observed in other conditions in which consciousness is lost, such as general anesthesia [70] and the vegetative state [71]. Based on these observations, the perturbational complexity index (PCI) was developed to quantify the complexity of the overall EEG response to TMS [72]. PCI has proven to be a reliable index of consciousness [72, 73]. The slow-wave-like response typical of NREM sleep has always been found associated with low values of PCI.

Although these empirical findings have practical implications, the basic mechanisms underlying the physiological EEG response to TMS are still largely unknown. Previous *in computo* works suggest that the specificity of the complex EEG response to TMS elicited during wakefulness may be associated with long-range connections [74] and a combination of intrinsic neuronal properties and cortico-cortical

[1] This work was supported in part by the National Institutes of Health under awards R21EB015542 and R01GM116916, the European Union under awards H2020-FETOPEN-2014-2015-RIA and HBP-SP3-T32-2 and the Swiss National Science Foundation under award CRSII3_160803/1.

circuits interactions [75]. Intracranial EEG recordings in humans also point to the importance of long-range recurrent connections [76]. However, the extent to which these factors contribute to the physiological EEG response to TMS during wakefulness and to the breakdown of effective connectivity and complexity during loss of consciousness still remains to be clarified.

Here we address the role of recurrent connections by interpreting TMS-evoked potentials (TEPs) using a modeling perspective. Specifically, we estimate models from single-trial EEG responses to TMS while exploring two different hypotheses for the complex, long-lasting compound response observed in wakefulness and sleep: (1) a segregated model (Figure 4.1 (a)) in which TMS results in a feedforward sweep that engages a number of cortical regions with different properties whose independent responses result in the TEP; (2) an integrated model (Figure 4.1 (b)) in which the initial feedforward sweep due to TMS is followed by recurrent interactions among cortical regions to produce the TEP. The identical set of cortical regions are used in both models. A semi-data-driven procedure that is independent of the models is used to select the regions for each subject across both wakefulness and sleep conditions. Hence, the only difference between the integrated and segregated models is the presence/absence of recurrent interactions between the selected cortical regions.

We examined the integrated and segregated hypotheses by extending the linear state-space model (SSM) framework developed in [53] for spontaneous EEG to incorporate the feedforward pathways engaged by exogenous stimulation such as TMS. Our method estimates the coefficients associated with the feedforward pathways to the cortical regions and the (possible) recurrent interactions between regions. We use this new method to compare the two different models for TMS/EEG recordings during both wakefulness and NREM sleep. The integrated model (Figure 4.1 (b)) assumes the cortical regions involved in the response are fully connected and interacting. In contrast, the segregated model (Figure 4.1 (a)) assumes the cortical regions involved in the response do not interact with one another. The segregated and integrated models mimic the absence or presence of long-range connections, respectively.

The estimated models are compared using the TMS evoked response and cross-validation of one-step prediction errors on single trials. Our results show that the brain dynamics evoked by TMS during NREM sleep can be described equally well

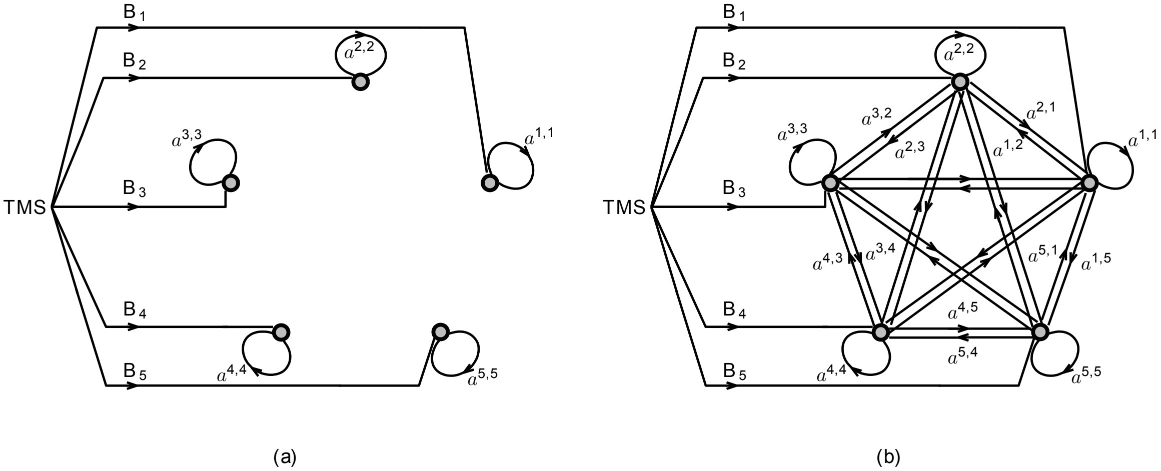


Figure 4.1: Schematic diagram of multivariate autoregressive with exogenous stimulation (MVARX) network models. Each gray circle represents a cortical ROI, and directed edges denote non-zero MVARX coefficients. The $\mathbf{B}_i$ represent the direct or indirect feedforward effect of TMS on each ROI. (a) Segregated model assumes each cortical region acts independently of all others. (b) Integrated model represents network interactions between cortical regions.

by both segregated and integrated models over the entire response duration. In contrast, integrated and segregated models provide comparable fit to the actual TEP in wakefulness only for the assumed feedforward path duration. After feedforward effects subside the integrated model provides a much better fit to the actual TEP. Similarly, the integrated model has lower one-step prediction error than the segregated model on single trials not used to train the models. The results strongly suggest that the high levels of complexity typical of TMS/EEG responses during wakefulness, as also assessed by PCI, requires the presence of effective recurrent interconnections.

4.2 Methods

Data

Subjects Seven healthy volunteers participated in the study. All subjects gave written informed consent, and the experiment was approved by the Comitato Etico Interaziendale Milano Area A, Milan, Italy. A clinical examination was performed before the experiment to exclude potential adverse effects of TMS. TMS was performed in accordance with current safety guidelines. TMS/EEG data was initially collected during wakefulness when subjects were alert and relaxed, with eyes open, and then the same stimulation was performed after subjects entered a consolidated period of NREM sleep.

TMS Targeting Stimulation was performed by a focal figure-of-eight coil (mean/outer winding diameter 50/70 mm, biphasic pulse shape, pulse length 280 μs, and focal area of the stimulation 0.68 cm^2) driven by a Mobile Stimulator Unit (eXimia TMS Stimulator, Nexstim Ltd.). Cortical TMS targets were identified on MRI (magnetic resonance imaging) scans acquired on a 3T magnetic resonance scanner (Trio Tim, Siemens, Germany) using a T1-weighted MP-RAGE (magnetization-prepared rapid acquisition gradient echo) sequence. We controlled TMS parameters by means of a Navigated Brain Stimulation (NBS) system (Nexstim, Helsinki, Finland). A 3D infrared tracking position sensor unit was employed to locate the relative positions of the coil and subject's head referenced to the individual MRI scan with an error tolerance of 3 mm. The NBS system also calculated the distribution and the intensity of the intracranial electric field induced by TMS on the cortical surface in real time. The output of the stimulating unit was adjusted to induce an electric field of 90 V/m on the cortical surface, which is above the threshold (50 V/m) for a significant EEG response [77, 78]. The stimulation coordinates were passed to a software aiming tool that ensured the reproducibility of position, direction, and angle of the stimulator throughout the session. At least 200 trials were collected. TMS was delivered with an interstimulus interval jittering randomly between 2000 and 2300 ms (0.4–0.5 Hz).

EEG Recordings During TMS We recorded TMS-evoked potentials by means of a TMS-compatible 60-channel amplifier (Nexstim), a device that prevents amplifier saturation by means of a proprietary sample-and-hold circuit [67]. The analog output of the amplifier is held constant from 100 μs before stimulus to 2 ms after stimulus. The impedance at all electrodes was kept below 5 kΩ. The EEG signals were referenced to an additional electrode on the forehead, bandpass filtered (0.1–350 Hz) and sampled at 1450 Hz with 16-bit resolution. Two extra sensors were used to record the electro-oculogram (EOG). Well established procedures were followed for collecting TMS/EEG. Subjects wore earplugs and a sound was played continuously to avoid contamination of TMS-evoked potential by auditory potentials evoked by the click associated with the TMS discharge [4, 79, 80]. Bone conduction was attenuated by placing a thin layer of foam between coil and scalp [81]. These precautions ensure the measured EEG is due to direct cortical stimulation [82].

General Experimental Procedures During the experiment each subject was lying on a reclining chair with a head-rest that allowed a comfortable and stable head position. The navigation system was calibrated with a muscle artifact free target location in the left/right premotor (Brodmann area 6 or BA6) or the left/right posterior parietal (BA7) cortex identified prior to TMS/EEG data collection in wakefulness. These areas were stimulated along the midline, thus reducing the possibility of inducing muscular activation [83] and/or any possible secondary cortical response due to somatosensory perception [84]. A second TMS/EEG data collection session using identical stimulation parameters was initiated after subjects entered a consolidated period (>5 min) of NREM sleep stage 3.

TMS evoked potential In order to measure the duration of TMS induced response, we calculated the evoked potential and the duration of significant response following procedures in [4]. Figure 4.2 (a) and (b) show the TMS evoked potentials and the temporal extent of significant response (in red) during wakefulness and sleep for a single subject. In panel (c), we show the average and standard deviation of the maximum extent of the TMS-induced response among the seven subjects in wakefulness and sleep.

Data Preprocessing Data analysis was performed using MATLAB (MathWorks). First, TMS/EEG trials were visually inspected to detect and reject trials containing excessive noise, muscle activity, or eye movements. Next, trials were segmented into windows of ±800 ms around the TMS stimulus. Channels with large residual artifacts or bad signal quality were excluded from further analysis. All sessions analyzed used a minimum of 52 channels. The EEG data were average referenced, baseline corrected and independent component analysis (ICA) was applied in order to remove residual artifacts.

The data was further downsampled by a factor of 15 in two stages using the MATLAB function `resample` to obtain an effective sampling rate of 96.67 Hz. The downsampled data was then zero-phase high-pass filtered with a Butterworth

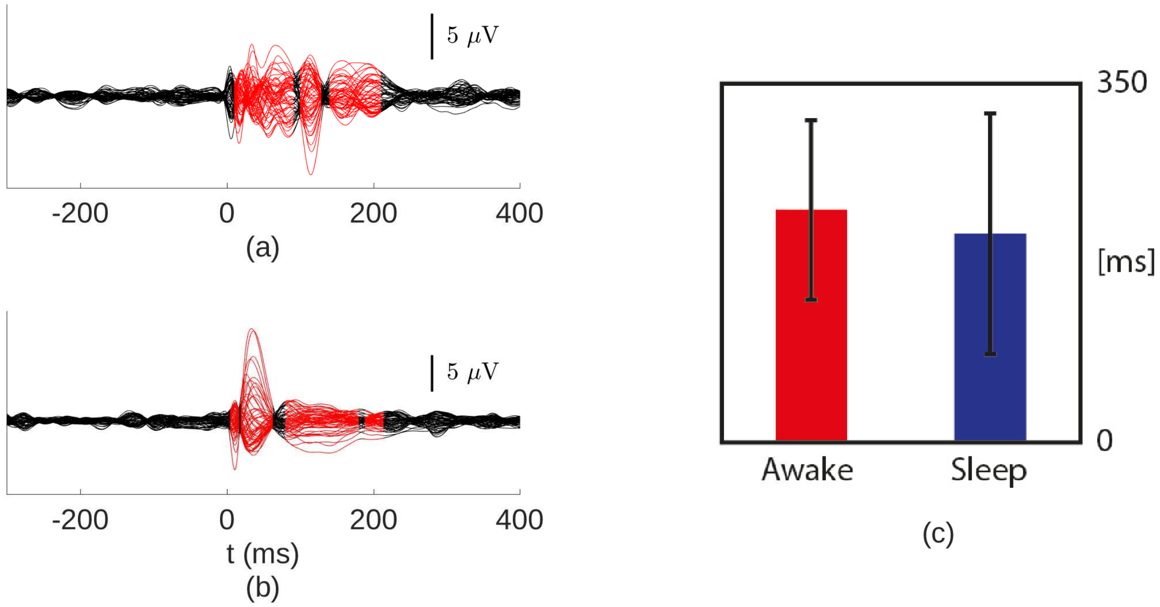


Figure 4.2: TMS-evoked potential and significant response durations. (a) and (b) Butterfly plots of TMS evoked potentials from 60 channels, recorded from the same subject during wakefulness and sleep, respectively. The red portions of the traces indicate the duration during which TMS induced a statistically significant response (calculated as in [4] - bootstrap statistics, $p < 0.01$). (c) Mean and standard deviation of extent of TMS-induced response for the seven subjects in wakefulness (red) and sleep (blue). There is no statistically significant difference between wakefulness and sleep (Wilcoxon rank-sum test, $p < 0.01$).

filter with passband edge frequency of 2 Hz. The downsampled, filtered data was baseline-corrected once more to make each channel of each trial zero-mean. We then followed a procedure similar to that described in [63] to identify outlying TEPs. The Mahalanobis distance [85] between the data in the trial being tested as an outlier and the remaining trials was computed. Trials with Mahalanobis distances having probability less than 0.1 were excluded from the subsequent analysis. A minimum of 115 trials were available for each session. Each trial of the data used for analysis included 310.3 ms prior to stimulation onset, 403.4 ms post stimulation and contains 70 samples. The TMS onset was at the 31st sample.

Perturbational Complexity Index (PCI) PCI [72] is a nonparametric measure of the data's spatio-temporal complexity. It is hypothesized to reflect the ability of many functionally specialized thalamo-cortical modules to interact producing a complex response. PCI was calculated by first band-pass filtering TEPs with a 0.1–45 Hz passband and then down-sampling to 362.5 Hz. Second, the cortical current density was estimated using a three-sphere head model [86, 87] and an empirical Bayes solution with weighted minimum norm constraint [88, 89, 90]. Next, significant cortical activations representing the deterministic pattern of TMS-evoked responses at the source level were obtained by applying a non-parametric bootstrap-based statistic [91, 92], leading to a 2D binary space-time matrix. The normalized Lempel-Ziv complexity [93] of this matrix was finally accumulated over space to obtain the temporal evolution of PCI, PCI(t).

Figure 4.3 depicts the temporal evolution of PCI (cumulative PCI) and the rate of complexity divergence in the difference between cumulative PCI in wakefulness and sleep.

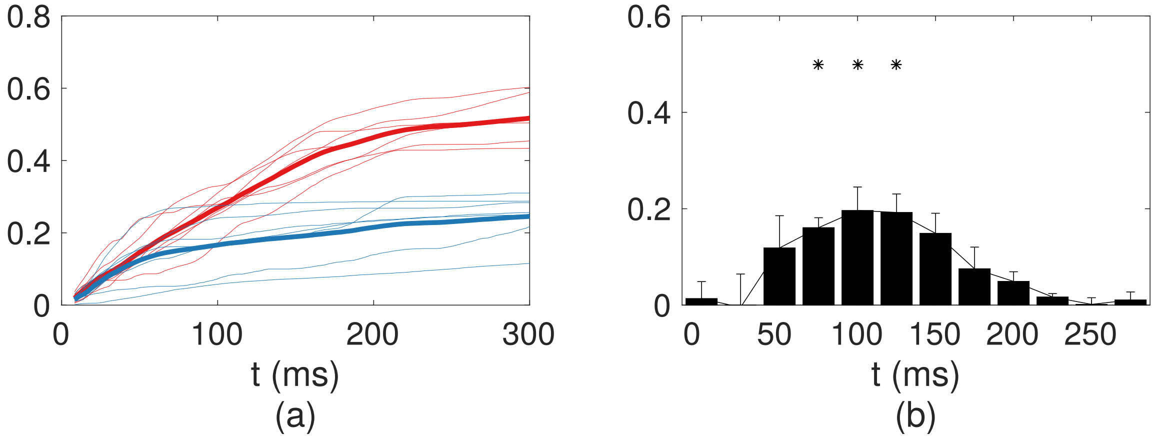


Figure 4.3: (a) Cumulative PCI is shown for every subject (thin lines) and at the group level (thick lines) for both wakefulness (red) and sleep (blue). (b) The rate of divergence of difference in cumulative PCI between wakefulness and sleep, ΔPCI, calculated from single-subject differences between cumulative PCI during wakefulness and sleep, with 25-ms time bins. Statistical significance with respect to zero across bins is indicated in asterisks (significance level $\alpha = 0.01$, Mann-Whitney).

Model

The linear SSM framework developed in [53] is extended to TMS/EEG recordings by explicitly modeling the feedforward effects of TMS stimulation as illustrated in Figure 4.1. The SSM model for TMS/EEG consists of two linear equations. A state equation describes the evolution of cortical activity as a multivariate autoregressive process with exogenous stimulation (MVARX), where the stimulation is TMS. An observation equation characterizes the measured single-trial EEG recordings as a weighted sum of cortical activity and noise. The parameters of the SSM are unknown, but assumed to be constant during the measurement time. An expectation-maximization (EM) algorithm is employed to find the maximum likelihood estimates (MLE) of the unknown parameters.

State Equation The K cortical signals representing activity in the K cortical ROIs at time n and trial j are denoted by the K by 1 vector $\mathbf{x}_{n,j} = [x^1_{n,j}, x^2_{n,j}, \cdots, x^K_{n,j}]^\top$ where $x^k_{n,j}$ is the cortical signal in ROI k at time n and trial j. The cortical activity is modeled as an MVARX-(p, ℓ) process [55, 63]:

$$\mathbf{x}_{n,j} = \sum_{i=1}^{p} \mathbf{A}_i \mathbf{x}_{n-i,j} + \sum_{i=0}^{\ell-1} \mathbf{b}_i u_{n-i,j} + \mathbf{w}_{n,j} \tag{4.1}$$

for $n = 1, 2, \cdots, N$ and $j = 1, 2, \cdots, J$. TMS is represented by the input $u_{n,j}$. We have $u_{n_o,j} = 1$ if TMS is applied at time n_o in trial j, and $u_{n,j} = 0$ for $n \neq n_o$. The residual error $\mathbf{w}_{n,j}$ is modeled as a zero-mean normally distributed random variable with K by K covariance matrix $\mathbf{Q}$. The K by K matrix $\mathbf{A}_i$ characterizes how cortical signals from i time samples in the past influence present cortical signals. The $(m, k)^{\text{th}}$ element of $\mathbf{A}_i$, $a_i^{m,k}$, is a weight that models the contribution of the signal from ROI k at i time samples in the past to the prediction of the signal from ROI m at the current time. The segregated model (Figure 4.1 (a)) constrains the $\mathbf{A}_i$ to be diagonal while the integrated model has no constraints on the $\mathbf{A}_i$. The K by 1 vector $\mathbf{b}_i$ captures the influence of the TMS on each of the K ROIs i samples following stimulation onset. Thus, the $\mathbf{b}_i$ coefficients model the feedforward volley of activity induced by TMS in the i^{th} ROI. This feedforward volley is assumed to include indirect effects of TMS through brain regions not included in the cortical regions being modeled. The number of $\mathbf{b}_i$, ℓ, models the maximum latency in significant feedforward connections from the stimulation site directly or indirectly to the K ROIs.

We collect the $\mathbf{b}_i$ into a K by ℓ matrix $\mathbf{B} = [\mathbf{b}_0, \mathbf{b}_1, \cdots, \mathbf{b}_{\ell-1}]$ for notational convenience. Similarly, we collect the $\mathbf{A}_i$ into a K by Kp matrix $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, \cdots, \mathbf{A}_p]$. We assume the initial state for each trial j, $[\mathbf{x}^\top_{0,j}, \mathbf{x}^\top_{-1,j}, \cdots, \mathbf{x}^\top_{-p+1,j}]^\top$, is normally distributed with identical, but unknown mean $\boldsymbol{\mu}_0$ and covariance matrix $\sigma_0^2 \cdot \mathbf{I}$. **A** and $\mathbf{Q}$ may be used to compute functional and effective connectivity and measures of integrated information [53, 94].

Observation Equation Let the M by 1 vector $\mathbf{y}_{n,j}$ denote the measured data in M EEG channels at time n and trial j. The observation equation models $\mathbf{y}_{n,j}$ as the sum of activity due to each ROI, $\mathbf{H}_k \boldsymbol{\lambda}_k x^k_{n,j}$, plus observation noise $\mathbf{v}_{n,j}$:

$$\mathbf{y}_{n,j} = \sum_{k=1}^{K} \mathbf{H}_k \boldsymbol{\lambda}_k x^k_{n,j} + \mathbf{v}_{n,j}. \tag{4.2}$$

The M by 1 vector of observation noise $\mathbf{v}_{n,j}$ is assumed to be zero-mean normally distributed with covariance matrix $\mathbf{R} = \sigma^2_R \cdot \mathbf{I}$. Here $\mathbf{H}_k$ is an M by 3 matrix describing the forward model for the k^{th} cortical ROI. $\mathbf{H}_k$ is a cortical patch basis [95]. The 3 by 1 vector $\boldsymbol{\lambda}_k$ specifies the orientation of the source with respect to the basis formed by the columns of $\mathbf{H}_k$ and is assumed to be unit norm but unknown.

We assume the leadfield matrix is based on known dipole orientations and thus $\mathbf{H}_k$ is a rank-3 approximation to the space spanned by the columns of the leadfield vectors associated with all dipole sources within ROI k. More detailed discussion of the observation model is available in [53].

EM Algorithm

The parameters to be estimated in the SSM are $\theta = \{\mathbf{A}, \mathbf{B}, \mathbf{Q}, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \cdots, \boldsymbol{\lambda}_K, \mathbf{R}, \boldsymbol{\mu}_0, \sigma^2_0\}$. Our goal is to find the MLE of θ. MLEs have the least variance of all unbiased estimates for sufficiently large data sets [96]. We write the log likelihood function as

$$\begin{aligned} L(\theta) &= \log p(\mathbf{Y}|\mathbf{U}, \theta) = \log \int p(\mathbf{Y}, \mathbf{X}|\mathbf{U}, \theta) d\mathbf{X} \\ &= \sum_{j=1}^{J} \log p(\mathbf{y}_{1,j}, \mathbf{y}_{2,j}, \cdots, \mathbf{y}_{N,j}|\mathbf{U}, \theta) \end{aligned} \tag{4.3}$$

where $\mathbf{Y}$ denotes the collection of measured data from all trials, $\mathbf{U}$ denotes the TMS input and $\mathbf{X}$ denotes the collection of cortical signals. The MLE is the solution to the optimization problem: $\max_\theta L(\theta)$. This optimization problem does not have a closed-form solution and in general is not convex as it involves latent variables $\mathbf{X}$. The EM algorithm is an iterative coordinate ascent algorithm for finding MLEs

[97, 53]. The algorithm starts with an initial guess $\theta^{(0)}$ and iterates the E- and M-steps:

- E-step: Evaluate the probability distribution $p(\mathbf{X}|\mathbf{Y}, \mathbf{U}, \theta^{(k)})$ and the conditional expectation $\mathcal{Q}(\theta, \theta^{(k)}) = \mathbb{E}_{\mathbf{X}|\mathbf{Y},\mathbf{U},\theta^{(k)}}[\log p(\mathbf{Y}, \mathbf{X}|\mathbf{U}, \theta)]$
- M-step: Find $\theta^{(k+1)} = \arg\max_{\theta} \mathcal{Q}(\theta, \theta^{(k)})$.

A convergence criterion is employed to decide whether or not the algorithm should terminate at iteration k. The convergence criterion involves comparing $L(\theta^{(k)})$ and $L(\theta^{(k-1)})$[98] in the E-step. The EM algorithm is guaranteed to monotonically increase the objective function $L(\theta)$. Thus, at worst the EM algorithm finds a local maximum of the log likelihood function. We start our EM algorithm at multiple initial guesses and choose the solution with largest log likelihood to improve the chances of finding the global maximum. The EM algorithm has been shown to give more accurate model estimates for spontaneous data than two-step methods that employ source reconstruction followed by estimation of the cortical multivariate autoregressive model [53]. Additional details of the modeling procedure and the EM algorithm are presented in A.

Region Selection

We choose the subject-specific ROIs using a semi-data-driven sequential method. First, a minimum norm estimate of source activity at each dipole is reconstructed. Second, patches with the largest power for both wakefulness and sleep are sequentially added into the subject's ROI set.

The cerebral cortex of each subject's brain was modeled as a three-dimensional grid of 3004 fixed dipoles oriented normally to cortical surface. This model was adapted to the anatomy of each subject using the Statistical Parametric Mapping software package (SPM5, http://www.fil.ion.ucl.ac.uk/spm/) using the same parameters used in [77]. Finally, the inverse transformation was applied to the Montreal Neurological Institute (MNI) canonical mesh of the cortex for approximating to real anatomy.

Cortical patches are defined as collections of dipoles. We exclude deep dipoles from the analysis to avoid source activity with very low signal-to-noise ratio. Deep dipoles are defined based on the distances from the dipole to all electrodes. Our criteria results in 2295 dipoles being included in the construction of patches. We generated 617 patches of geodesic radius 2 cm that are overlapping and approximately uniformly distributed.

The power in the measured data associated with each candidate patch is computed using the minimum norm method for each subject's wake and sleep data set. We select the ROI set according to patch signal power using a sequential approach. The first ROI is chosen as the patch with the largest normalized average power over wake and sleep from the Brodmann area targeted by TMS. In each subsequent iteration, the patch with the largest normalized average power over wake and sleep is added to the ROI set, under the constraint that this new patch does not physically overlap with any of the already selected patches. This procedure is repeated until $\lfloor M/3 \rfloor$ patches are selected where M is the smaller of the numbers of artifact free channels in wakefulness and sleep. $\lfloor \cdot \rfloor$ denotes the greatest integer less than operation. Note that $\lfloor M/3 \rfloor$ is the maximum number of regions that can be modeled without introducing linearly dependent components into the observation equation. A detailed description of the region selection procedure is described in B.

Model Selection

The number of regions included in the model, K, and the memory, p, are selected using the Akaike information criterion (AIC) [55]. We considered $K = 3, 6, 9, \cdots, \lfloor M/3 \rfloor$ and $p = 5, 10, 15, \cdots, 30$. The model parameter ℓ represents the approximate duration the feedforward volley of TMS-induced activity to the ROIs in samples. The actual duration may vary from one region to another; however, we chose a single value for all regions as a modeling compromise. We choose $\ell = 5$, $\ell = 10$, and $\ell = 15$ corresponding to maximum feedforward volley durations of 50 ms, 100 ms, and 150 ms, for the results presented in the paper. These choices are motivated by previous research that suggests feedforward effects are approximately limited to within 100 ms post-stimulus [99, 100, 101].

Model Evoked Response

The evoked response of the model is obtained by setting $\mathbf{w}_{n,j} = \mathbf{0}$ in Eq. (4.1), $\mathbf{v}_{n,j} = \mathbf{0}$ in Eq. (4.2), and

$$u_{n,j} = \begin{cases} 1, & n = n_o \\ 0, & n \neq n_o \end{cases}$$

where n_o corresponds to the time that the stimulus is applied. Note that setting $\mathbf{w}_{n,j} = \mathbf{0}$ and $\mathbf{v}_{n,j} = \mathbf{0}$ eliminates the dependence of the model on j, so no averaging is utilized in obtaining the model evoked response.

Insight into the nature of the two models is obtained by expressing the evoked response in terms of the feedforward and feedback model parameters. For simplicity of presentation we assume $p = 1$ and drop the trial subscript j. Denote the evoked response of the model in the cortical regions as $\hat{\mathbf{x}}_n$. Using $\hat{\mathbf{x}}_n = \mathbf{0}$ for $n < n_o$ we have

$$\begin{aligned}
\hat{\mathbf{x}}_{n_o} &= \mathbf{b}_0; \\
\hat{\mathbf{x}}_{n_o+1} &= \mathbf{A}_1\hat{\mathbf{x}}_{n_o} + \mathbf{b}_1; \\
\hat{\mathbf{x}}_{n_o+2} &= \mathbf{A}_1\hat{\mathbf{x}}_{n_o+1} + \mathbf{b}_2; \\
&\vdots \\
\hat{\mathbf{x}}_{n_o+\ell-1} &= \mathbf{A}_1\hat{\mathbf{x}}_{n_o+\ell-2} + \mathbf{b}_{\ell-1}; \\
\hat{\mathbf{x}}_{n_o+\ell} &= \mathbf{A}_1\hat{\mathbf{x}}_{n_o+\ell-1}; \\
&\vdots \\
\hat{\mathbf{x}}_k &= \mathbf{A}_1\hat{\mathbf{x}}_{k-1}, \text{ for } k \geqslant n_o + \ell;
\end{aligned}$$

These expressions indicate that it is possible to exactly fit the first ℓ values of a measured evoked response $\mathbf{x}_n$ *for any* $\mathbf{A}_1$ by choosing the $\mathbf{b}_i$ appropriately. If we choose $\mathbf{b}_0 = \mathbf{x}_{n_o}$, and set $\mathbf{b}_i = \mathbf{x}_{n_o+i} - \mathbf{A}_1\mathbf{x}_{n_o+i-1}$ for $i = 1, 2, \ldots, \ell - 1$, then $\hat{\mathbf{x}}_n = \mathbf{x}_n$ for $n_o \leqslant n \leqslant n_o + \ell - 1$. That is, structural constraints on $\mathbf{A}_1$ do not necessarily manifest in the first ℓ values of the model evoked response. The model evoked response after the first ℓ time steps evolves according to only $\mathbf{A}_1$, and this is when we expect differences between integrated and segregated model evoked responses to be most evident.

4.3 Results

Model Performance in Capturing TEPs

A detailed description of results for one subject and $\ell = 10$ is given first and is followed by a summary of results for all seven subjects and values of ℓ studied.

Figure 4.4 depicts the ROIs selected for the analysis of the subject (see 4.2). The numerical label on each ROI indicates the order in which the patch was chosen during the region selection process. The TMS target was in the left parietal cortex (BA7), which is shown in red. The TMS-evoked and model-evoked responses in a representative set of six channels are shown in Figure 4.5 during both wakefulness

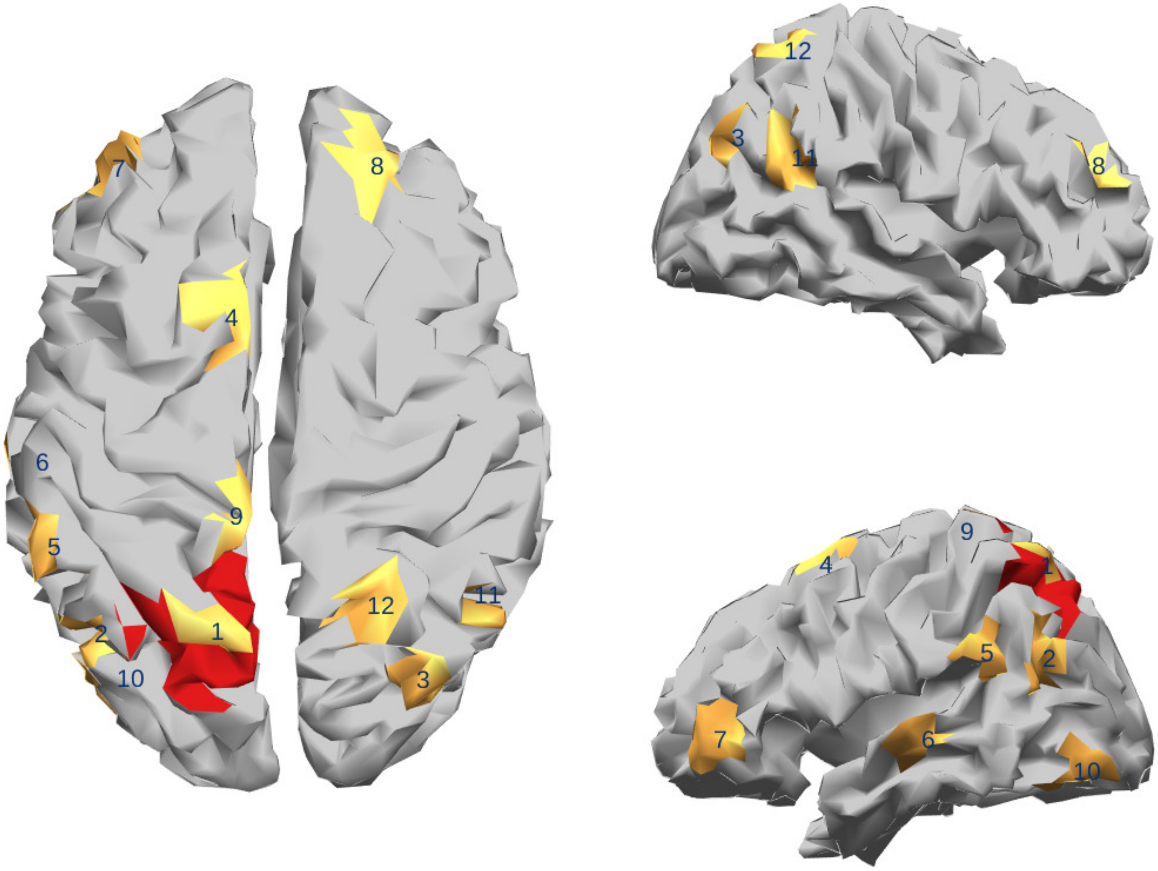


Figure 4.4: ROIs and TMS stimulation site in the analysis of Subject 1. The number associated with each ROI (in yellow) represents the order in which the ROIs were identified by the ROI selection procedure. TMS stimulation was applied to the left parietal cortex, which is shown in red.

and sleep. The number of ROIs K and order p of the segregated model were set equal to the values selected by AIC for the integrated model (see 4.2). Figures 4.6 and 4.7

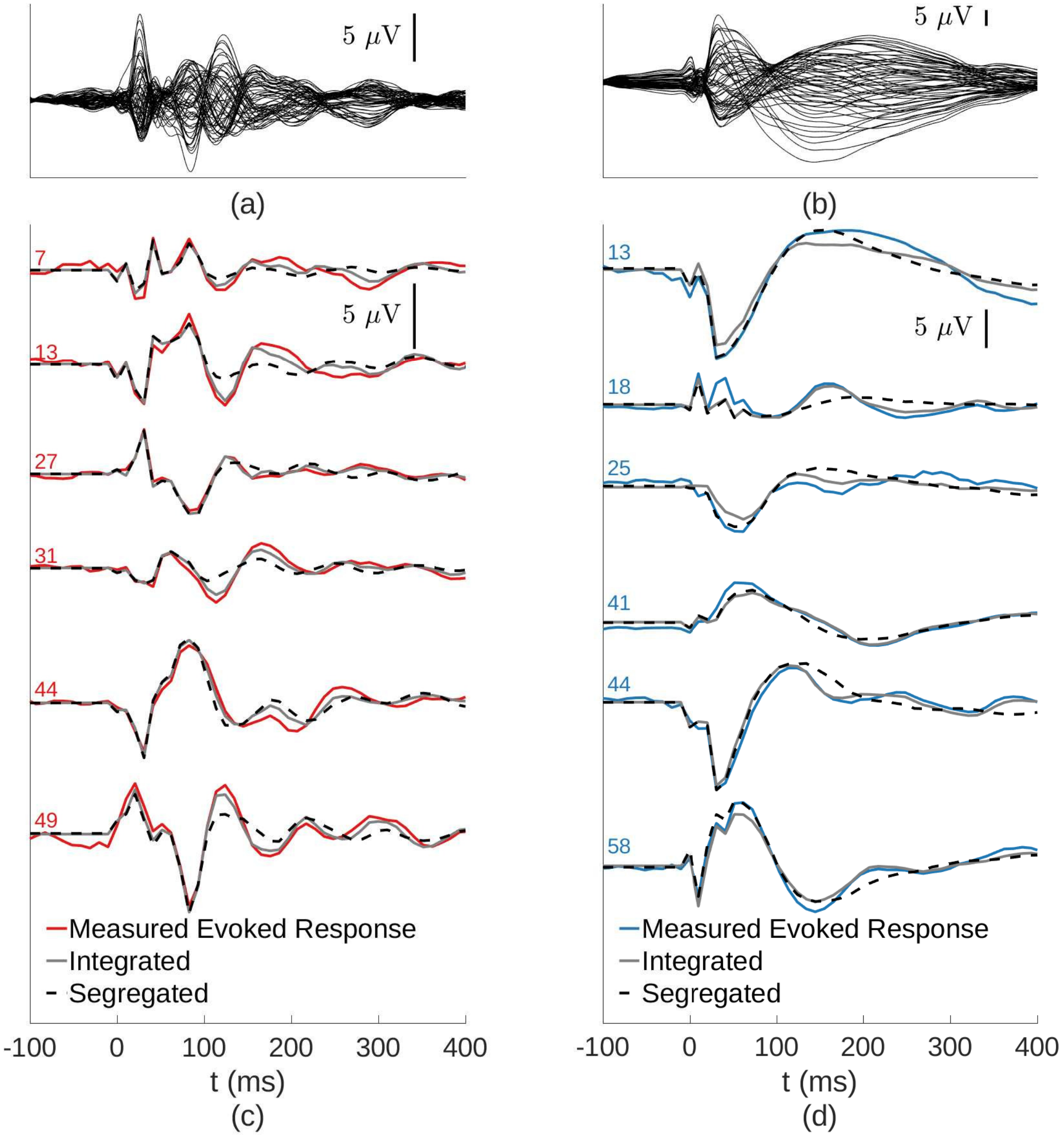


Figure 4.5: Measured and modeled TMS-evoke response of subject 1. (a) and (b) Butterfly display of 60-channel TMS-evoked response of subject 1 in wakefulness and sleep, respectively. (c) and (d) Measured and modeled TMS-evoked response in selected channels for Subject 1 in wakefulness and sleep, respectively.

show per-channel normalized squared response error (NSRE) measured over the interval of 100 to 400 ms post stimulus for each channel. Let $\bar{\mathbf{y}}_n$ be the TMS-evoked response and $\hat{\mathbf{y}}_n(\hat{\theta})$ be the model evoked response where $(\hat{\theta})$ denotes the functional dependence of the model evoked response on the estimated parameters $\hat{\theta}$. The per-channel NSRE in the m^{th} channel is defined as

$$\text{per-channel NSRE}_m = \frac{\sum_{n=41}^{70}(\bar{y}_{n,m} - \hat{y}_{n,m}(\hat{\theta}))^2}{\sum_{n=41}^{70}\bar{y}_{n,m}^2}$$

where $\bar{y}_{n,m}$ is the TMS-evoked response from the m^{th} channel at time n and $\hat{y}_{n,m}(\hat{\theta})$ is the model evoked response of the m^{th} channel at time n. The interval 100 to 400 ms post stimulus was chosen to isolate the effect of feedback interactions from the feedforward TMS distribution modeled by the $\mathbf{B}_i$ (Figure 4.1). The channels depicted in Figure 4.5 are marked with colored bars. Channel indices that are not shown were identified as artifactual during data preprocessing.

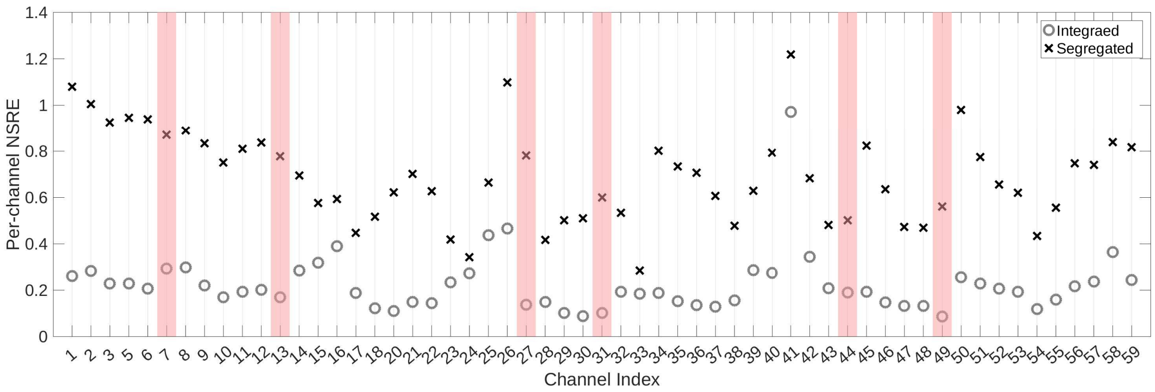


Figure 4.6: Per-channel normalized squared response error (NSRE) between the measured and modeled TMS-evoked response for Subject 1 in wakefulness assuming $\ell = 10$. The light red vertical bars indicate channels whose waveforms are displayed in Figure 4.5.

Scatter plots of global NSRE for models with different ℓ's are depicted in Figure 4.8 as a function of the global mean field power (GMFP) [102] ratio. The global NSRE is measured over all channels and the time interval of $[10\ell, 400]$ ms poststimulus. More specifically, it is defined as

$$\text{global NSRE}(\ell) = \frac{\sum_{n=\ell+31}^{70}||\bar{\mathbf{y}}_n - \hat{\mathbf{y}}_n(\hat{\theta})||_2^2}{\sum_{n=\ell+31}^{70}||\bar{\mathbf{y}}_n||_2^2}$$

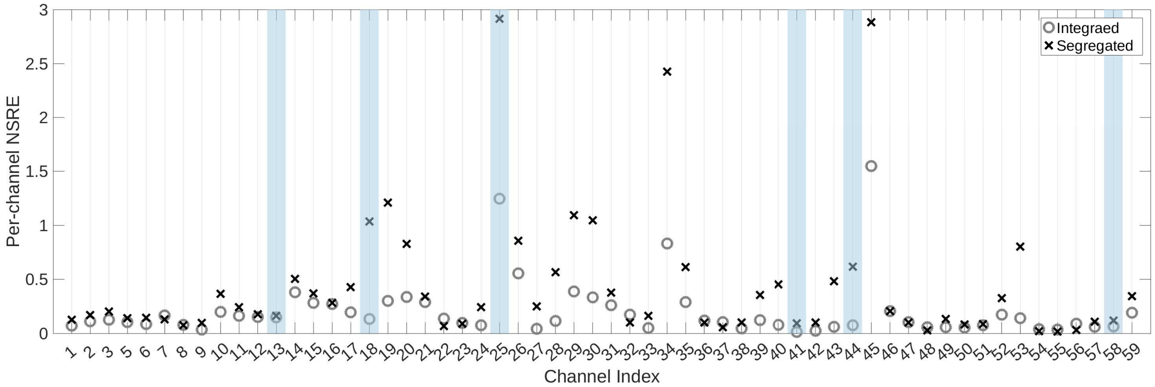


Figure 4.7: Per-channel NSRE between the measured and modeled TMS-evoked response for Subject 1 in sleep assuming $\ell = 10$. The light blue vertical bars indicate channels whose waveforms are displayed in Figure 4.5.

where $||\mathbf{y}||_2$ denotes the Euclidean norm of the vector $\mathbf{y}$. The GMFP ratio is formed as the ratio of the average GMFP over the interval of 0 ms to 400 ms post stimulation to a constant representing a bootstrapped estimate [103] of the maximum GMFP over the pre-stimulation interval -300 ms to -50 ms. Panels (a) and (b) depict global NSRE measured from 50 to 400 ms for models with $\ell = 5$. Panels (c) and (d) show global NSRE measured from 100 to 400 ms for models with $\ell = 10$. Panels (e) and (f) depict global NSRE measured from 150 to 400 ms for models with $\ell = 15$. Robust linear fits of global NSRE as a function of GMFP ratio are also shown in Figure 4.8. The robust linear fits are estimated with the iteratively reweighted least squares algorithm.

The cumulative NSRE was computed as a function of time by summing the SRE over all channels from stimulus onset to the current time

$$\text{cumulative NSRE}(n) = \frac{\sum_{n'=31}^{n} ||\overline{\mathbf{y}}_{n'} - \hat{\mathbf{y}}_{n'}(\hat{\theta})||_2^2}{\sum_{n'=31}^{n} ||\overline{\mathbf{y}}_{n'}||_2^2}.$$

Figure 4.9 depicts cumulative NSRE for all subjects with models of $\ell = 5$, $\ell = 10$, and $\ell = 15$ in wakefulness and sleep. There is variability in the cumulative NSRE as a function of time across subjects in both wakefulness and sleep. A clear difference is apparent between integrated and segregated models in wakefulness when feedforward effects subside.

We also compared the integrated and segregated models using cross-validated mean squared prediction error (MSPE). The MSPE evaluates the models' ability

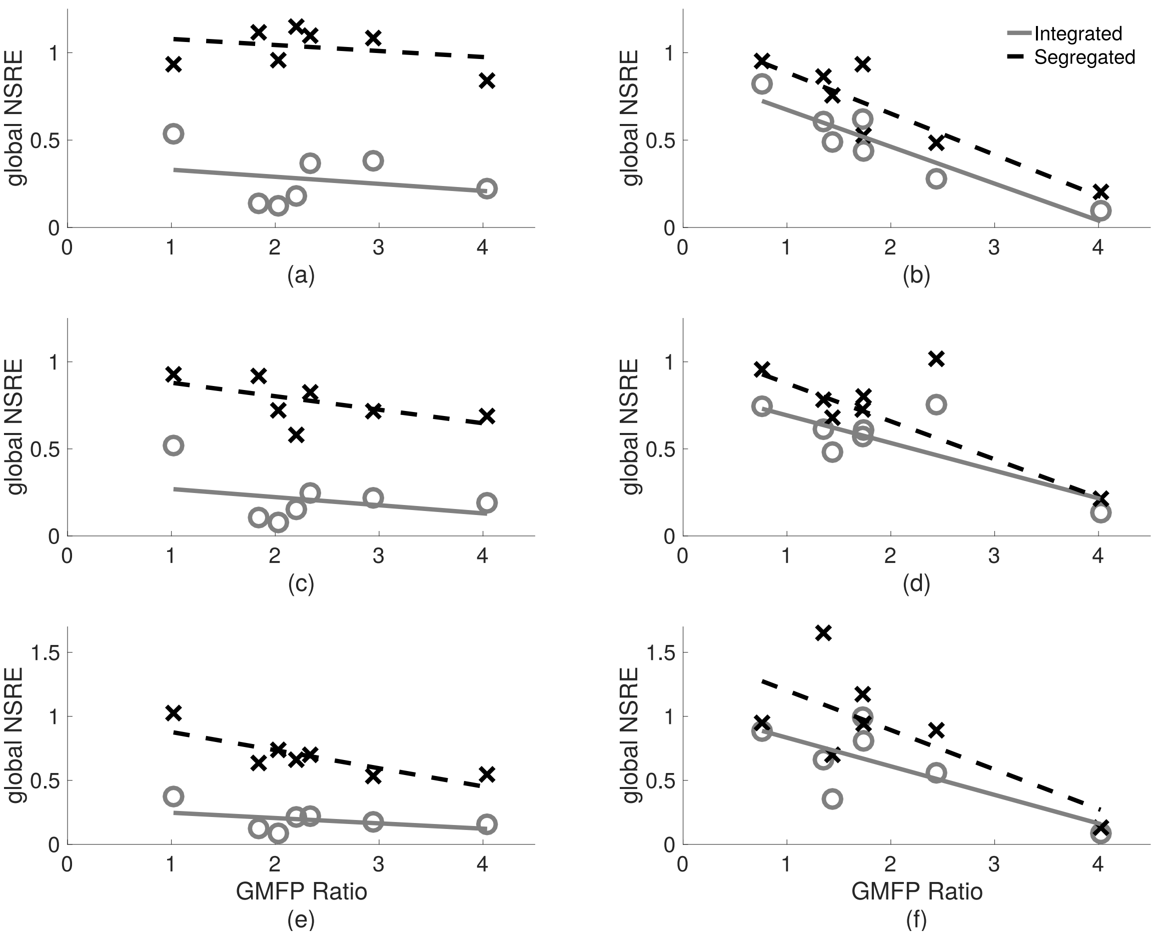


Figure 4.8: Global NSRE between measured and modeled TMS-evoked responses summed over all channels as a function of global mean field power (GMFP) ratio for all seven subjects. (a) Wakefulness, $\ell = 5$ (b) Sleep, $\ell = 5$ (c) Wakefulness, $\ell = 10$. (d) Sleep, $\ell = 10$. (e) Wakefulness, $\ell = 15$. (f) Sleep, $\ell = 15$.

to predict single trials of the scalp measurements one step into the future. The model fitting procedure (see 4.2) minimizes the error between model predictions and measured single trials in a probabilistic sense. This motivates the MSPE as an intuitive measure for the model fit to single trials.

Cross validation is used to control for the possibility of overfitting by the integrated model and involves splitting the single trials into test and training sets. The training set is used to estimate the model parameters while the test set is used to evaluate the models' ability to generalize to new data. We used ten-fold cross validation, so the data is partitioned into ten groups, with each group containing

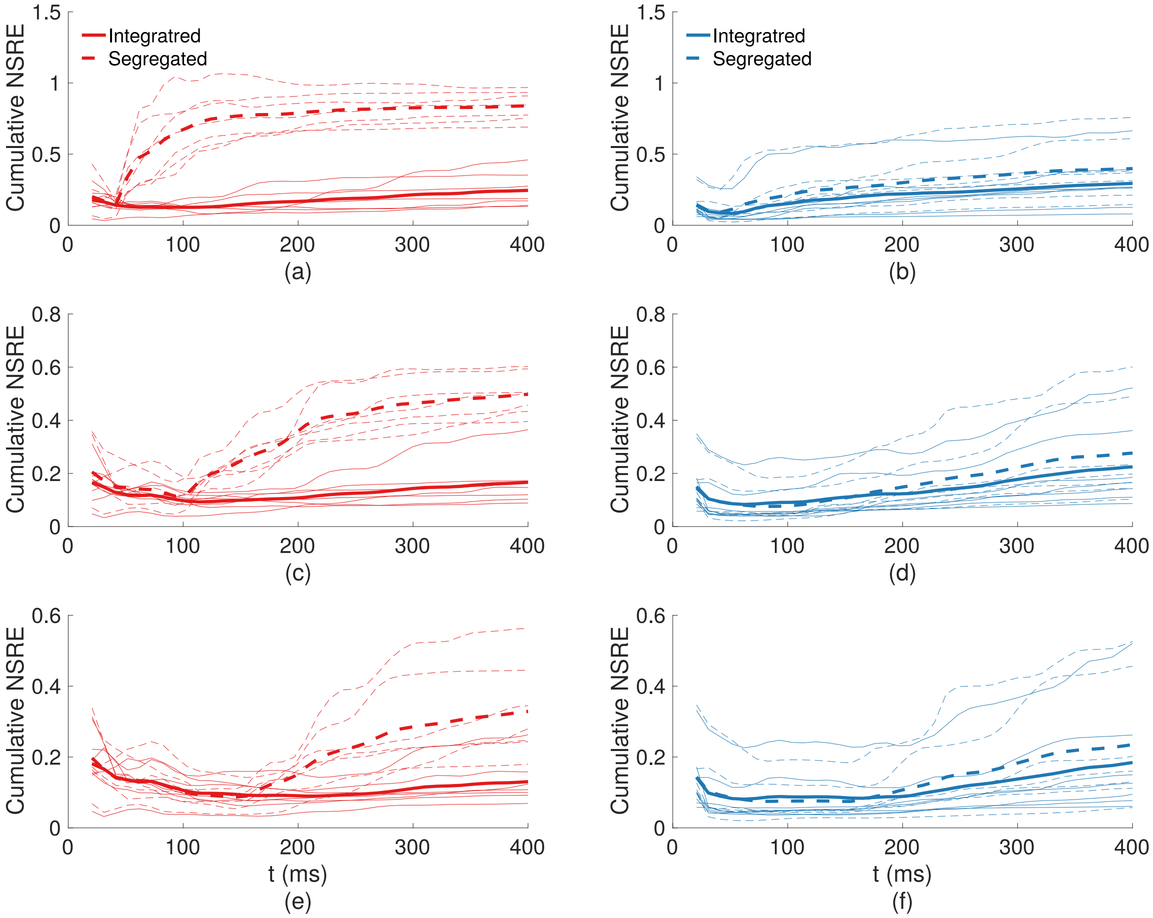


Figure 4.9: Cumulative normalized squared response error (NSRE). Thin lines indicate individual subject values while think lines indicate averages across seven subjects. (a) Wakefulness, $\ell = 5$. (b) Sleep, $\ell = 5$. (c) Wakefulness, $\ell = 10$. (d) Sleep, $\ell = 10$. (e) Wakefulness, $\ell = 15$. (f) Sleep, $\ell = 15$.

roughly the same number of trials. For the i^{th} fold of validation, the trials in the i^{th} group are used as the test set, and the other nine groups of trials are used as the training set. The model parameters are first estimated from the training set. Then the test trails withheld from training are used to evaluate one-step MSPE of the model estimated from the training set. The one-step MSPE of fold i is defined as

$$\text{MSPE}_i = \frac{1/|\text{Group}_i| \cdot \sum_{j \in \text{Group}_i} \sum_{n=1}^{70} \|\mathbf{y}_{n,j} - \tilde{\mathbf{y}}_{n,j}(n-1, \hat{\theta}_{-i})\|_2^2}{1/J \cdot \sum_{j \in J} \sum_{n=1}^{70} \|\mathbf{y}_{n,j} - \tilde{\mathbf{y}}_{n,j}(n-1, \hat{\theta}_{-i})\|_2^2}$$

where Group_i denotes the set of trials assigned to group i, $\tilde{\mathbf{y}}_{n,j}(n-1, \hat{\theta}_{-i})$ denotes the one-step prediction of $\mathbf{y}_{n,j}$ using the measurements up to time point $n - 1$ in trial j and the estimated parameters $\hat{\theta}_{-i}$. Here subscript $-i$ denotes that the model parameters are estimated from all trials except for those in the i^{th} group. The one-step prediction error is computed with Kalman-filtering procedures (see Appendix A and [104]).

The difference between segregated and integrated models in MSPE for each group of trials and subject are depicted in Figure 4.10. The differences are greater in wake than sleep for all subjects and choices of ℓ. A nonparametric Wilcoxon signed-rank test was used to test the hypothesis that the mean of the difference is zero. This hypothesis is rejected at an $\alpha = 0.01$ significance level in all subjects in wake and four of the seven subjects in sleep, for all three choices of ℓ. We also performed a two-way ANOVA and obtained a significant ($p < 0.01$) difference between integrated and segregated models in wakefulness but not in sleep.

4.4 Discussion

We have presented a method for assessing recurrent interactions in the cortex based on an MVARX model for the cortical activity induced by TMS and patch-based forward models for mapping the cortical activity to the scalp EEG. The MVARX model describes the cortical activity in each ROI as a weighted combination of past activity in all ROIs plus feedforward activation by TMS. The MVARX model approximates the complex interactions in the cortex with a linear model—the simplest causal model that can account for the rich temporal dynamics associated with EEG. Linear models benefit from reduced computational complexity, simpler parameter estimation approaches, and increased robustness to noise. We previously reported on MVARX models for intracranial EEG with direct electrical current stimulation [63]. Here we extend the MVARX approach to modeling cortical activity due to TMS using scalp EEG.

In contrast, while nonlinear models such as DCM [105] offer the potential of parameters with physiological meaning and higher degrees of freedom in modeling complex responses, they are rarely applied to more than six or seven ROIs due to the computational complexity of estimating model parameters and the difficulty of ensuring convergence of estimated parameters to good solutions. We did not consider DCM or other nonlinear models in this study because of this effective limitation to a

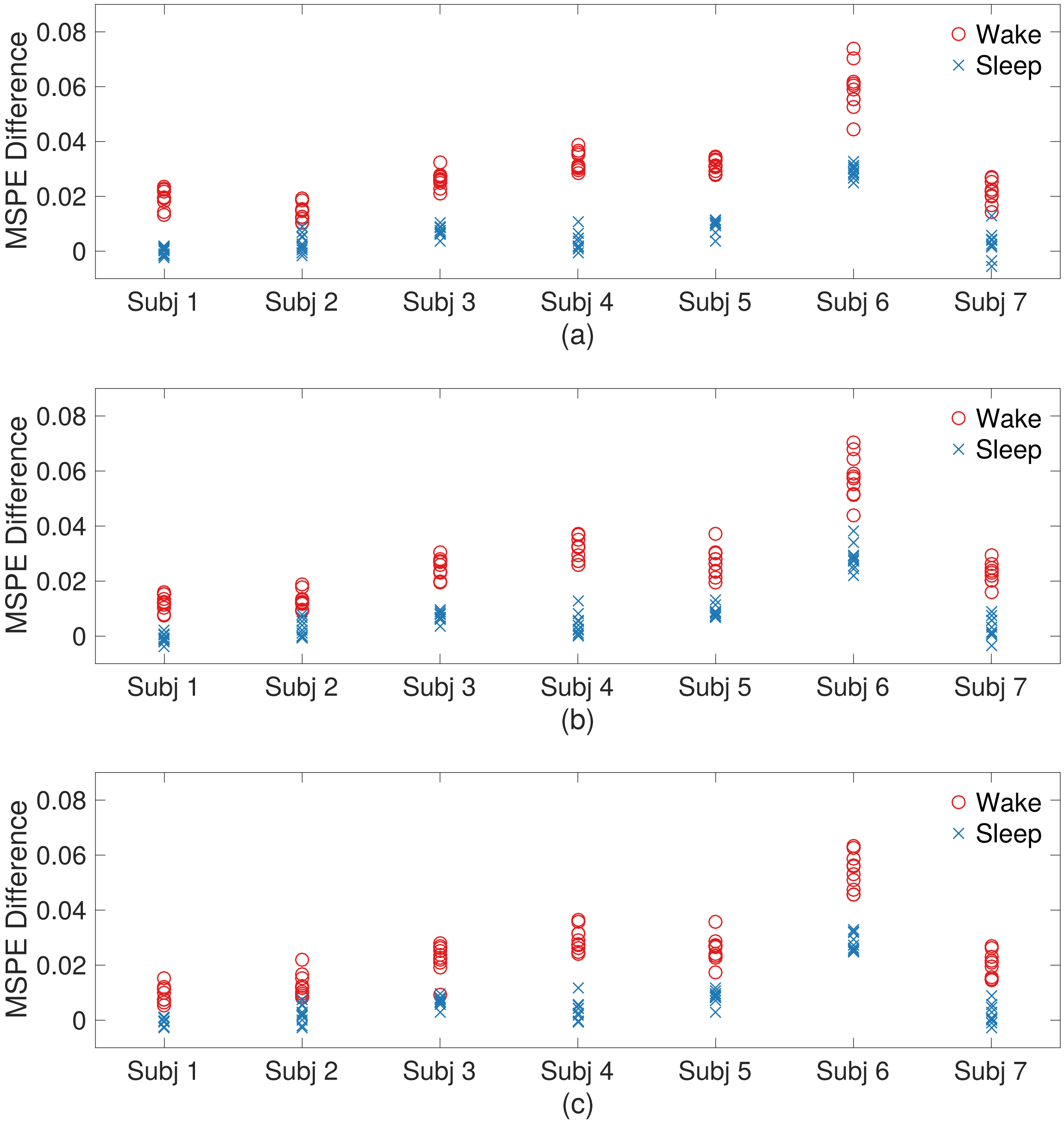


Figure 4.10: Difference between segregated and integrated model cross-validated MSPE for ten partitions of the trials. (a) $\ell = 5$. (b) $\ell = 10$. (c) $\ell = 15$.

smaller number of ROIs. The TMS-evoked response is relatively widely distributed throughout the cortex, especially in wakefulness [4]. Consequently in this paper we routinely use twelve or more ROIs to capture as many sources of activity as possible. The number of ROIs considered is constrained to be smaller than the number of measurement channels divided by three, otherwise the observation equation would become degenerate. Using the maximum possible number of ROIs also helps reduce sensitivity to the ever present hidden node problem with network models.

An EM algorithm is presented for estimating the MVARX model parameters from single-trial scalp EEG. Our EM approach directly estimates the cortical model from the scalp EEG using the maximum likelihood criterion. This avoids the suboptimal nature of two-step methods that first attempt to solve the inverse problem to obtain cortical activity, and then solve a second problem to fit a model to the estimated cortical activity. Our one-step approach has the potential for significantly better performance at modest and low SNR than two-step approaches. This is because the inverse problem is ill-posed and its solution amplifies noise. Noise in the estimated cortical activity contaminates and biases the model parameter estimates in the second step. While the cortical signals and models are unknown in human data, a comparison between EM and two-step approaches in a related problem [53] provides clear support for this reasoning. The results in this paper show that our approach leads to models with good fidelity to TEP.

The role of recurrent interactions in cortical networks is assessed by comparing and contrasting the performance of two MVARX models—an integrated model with all possible interactions between ROIs (Figure 4.1 (b)) and a segregated model with no interactions between ROIs (Figure 4.1 (a))—in both wakefulness and sleep. We evaluated the effect of reentrant connectivity on a given set of ROIs using a model independent, semi-data-driven procedure to chose a common set of ROIs across both conditions for each subject. Using different numbers or choices of ROIs across models or conditions would confound the respective choices with the effect of network structure.

Our results show that the integrated model provides significantly better visual agreement with the measured TEP than the segregated model for the more complex responses associated with wakefulness (Figures 4.5 (a), 4.6, 4.8), especially later than 10ℓ ms post stimulus for all three values of ℓ studied. Consistent with previous intracranial recordings in monkeys [99] and humans [106], a previous non-invasive TMS/EEG experiment showed that a maximum spread of activation, possibly reflecting the first feedforward sweep, can occur 80-100 ms after the stimulation [77]. In line with these studies, we considered durations for the exogenous parts of both integrated and segregated models of 50 ms ($\ell = 5$), 100 ms ($\ell = 10$), and 150 ms ($\ell = 15$).

Modeling of the feedforward effects accounts for the fidelity of the model responses during the first 10ℓ ms post stimulus (Figure 4.9). The EM algorithm chooses the model parameters using the single trials and does not explicitly fit the TEP. However, the TEP is a large component of the response right after the stimulus, and thus it is not surprising that both integrated and segregated models choose their feedforward parameters to closely fit the TEP in the first 10ℓ ms as discussed in Section 4.2. However, after the first 10ℓ ms the interactions between ROIs provided by the integrated model make a significant difference in modeling the more complex responses of wakefulness. The TEP in sleep is of comparable duration (Figure 4.2), but is not as complex and thus is nearly equally well described by the integrated and segregated models.

The NSRE is computed over the interval 10ℓ ms - 400 ms post stimulus to quantitatively assess the impact of the recurrent interactions in the model on the TEP. The per-channel NSRE for Subject 1 (Figure 4.6) shows that the integrated model performs better in every single channel during wakefulness. Normalization enables us to see how the error compares to the amplitude of the signal. This avoids the deceptive implications of very small errors that are associated with very low amplitude channels. In contrast, the largest normalized errors in Figure 4.6 correspond to channels with very weak amplitude signals. The global NSRE across all channels shows that the integrated model performs significantly better for all seven subjects in wakefulness (Figure 4.8 (a), (c), (e)). These results highlight the important role of recurrent interactions in producing the complex TMS response patterns associated with wakefulness. In contrast, the difference between integrated and segregated models is much less in sleep (Figures 4.5 (b), 4.7, 4.8 (b), (d), (f)) than in wakefulness.

The cumulative NSRE (Figure 4.9) shows that the performance of integrated and segregated models is nearly identical in the first 10ℓ ms post TMS in both wakefulness and sleep. This is due to the identical duration assumed for the feedforward volley in both models. In wakefulness the integrated and segregated model NSREs diverge after 10ℓ ms post TMS and continue to diverge as time increases (Figure 4.9 (a), (c), (e)). Interestingly, these results are paralleled by the time course of the PCI metric shown in Figure 4.3. The population cumulative PCI for wakefulness and sleep is similar up to 75-100 ms, maximally diverge between 100 and 125 ms and then the cumulative PCI for wakefulness increases significantly, while that for sleep does not. The cumulative PCI thus suggests that a more complex model is required to describe the TEP in wakefulness than sleep. Overall, our findings seem to suggest that the build-up of the complex responses observed during wakefulness after feedforward effects subside requires the engagement of recurrent interactions among cortical nodes and that these interactions are impaired during NREM sleep. This mechanism and its timeframe are generally consistent with the results of a recent intracranial study employing single-pulse electrical stimulations and stereo EEG recordings [76]. This study showed that during wakefulness electrical stimulation triggers a sequence of deterministic phase-locked activations in its cortical targets. In contrast, during NREM sleep cortical neurons have the tendency to fall into a silent down-state upon receiving a input due to underlying bistability. The downstate occurs as early as 100 ms after the pulse and obliterates the deterministic effects of the initial input, as indicated by a sharp drop of phase-locked activity. Thus, one may hypothesize that while during wakefulness the initial feedforward activation triggered by cortical stimulation evolves after 100 ms into a chain of deterministic interactions among cortical ROIs leading to a complex response, during NREM sleep the same feedforward sweep simply triggers down-state in target neurons which blocks the build up of complex interactions.

The integrated model has more parameters than the segregated model. We used cross validation on the MSPE [104] to assess whether the improved performance of the integrated model in wakefulness could be due to overfitting. The reduced MSPE for the integrated model is statistically significant in all of the seven subjects in wakefulness. Figure 4.10 shows the ten differences between segregated and integrated model cross-validated MSPE are positive for every subject in wakefulness. This strongly suggests that the improved performance of the integrated model in wakefulness is not due to overfitting the data—it shows that the integrated model generalizes better to data not used to train the model. Note that cross validation provides a very robust approach to model selection that naturally controls for model complexity. If the more complex model is fitting noise, then it will have poor performance describing data not used to train the model. In contrast, if the cross-validated MSPE is lower for the more complex model, then additional parameters are genuinely contributing to better modeling the data. Cross validation in principle also provides a more robust approach to selecting other model parameters, such as the number of regions and the memory p. However, cross validation has a very high computational cost and thus we chose to limit its use here to a binary comparison of integrated and segregated models.

It is possible that an MVARX model with a subset of the feedback connections in the integrated model—that is, some connections between ROIs constrained to zero—could improve upon the performance of the fully integrated model. Reducing model degrees of freedom potentially reduces the variance associated with model parameter estimation. We did not explore limiting the number of feedback connections due to the very large number of such models possible for even twelve ROIs. Furthermore, it would not change the conclusion that effective feedback connections between ROIs appear to be required to produce the TMS/EEG in wakefulness. The contrast between fully integrated and segregated models provides evidence for the importance of at least some feedback connections. It does not imply that a fully connected model is necessary.

Modeling involves a compromise between computation and parameter estimation considerations and the faithfulness of the model to the underlying, typically unknown, phenomenon. This tradeoff is manifest in our decisions to employ a linear model versus a nonlinear model, only consider a fully connected model versus partially connected models, and use the same values of p in all connections.

A challenging aspect of applying the modeling methodology described here is identification of the ROIs to include in the cortical network model. Any prior information of brain regions involved in the paradigm being studied should be used to select ROIs as illustrated in the spontaneous EEG studies of [107, 108]. Semi-data-driven ROI selection, such as used in the present study, is closely related to the source localization problem and thus, any source localization method may be used. However, we strongly recommend that the ROI selection process be based on mapping source localization results onto the bases used to represent the ROIs in the observation equation. This ensures ROI selection maps directly to ROI representation in the model.

Source localization methods are unlikely to identify ROIs that are relatively silent in the scalp EEG due to depth or weak electrical activity. Similarly, our method for MVARX model estimation will have difficulty identifying network interactions involving ROIs that do not contribute measurable activity to the scalp EEG.

The problem of hidden nodes is endemic to any network model. If an important ROI is not included in the model, then it is possible to draw false inferences about causal influences and connectivity within the true underlying network. Our procedure for selecting ROIs includes those with the most significant contributions to the measured scalp EEG. It is possible that an ROI with very weak activity, or a deep ROI, plays a significant role in network interactions in this study. However, the difference we found between integrated and segregated models is likely insensitive to hidden nodes since a hidden node would only change the connectivity of the integrated model. In contrast, the potential impact of hidden nodes must be considered in studies that rely on comparing connectivity within a single model.

Theoretical considerations have been used to argue complex network interactions as a basis for consciousness [109]. This study concurs with others [63] that highlight the role of network interactions for supporting consciousness. The TEP in sleep, when the subjects are unconscious, is explained almost as well by the segregated model as it is by the integrated model. Including network interactions provides marginal benefit in describing the measured TEP or to the model's one-step prediction performance of single trials. In contrast, during wakefulness the network interactions of the integrated model result in substantially better fit to the TEP and improved one-step prediction of single trials.

5 VARIATIONAL BAYESIAN INFERENCE FOR STATE-SPACE MODELS

5.1 Introduction

In signal processing, two recurring questions are model order selection and learning of the structural pattern of the parameter. To answer these questions with point estimation frameworks we have been concerned with thus far, the estimation procedure has to be done multiple times with different model specifications as in model selection and with different sets of data as in bootstrapping or cross-validation.

In Bayesian inference, instead of point estimation, we learn posterior distributions of the parameters. With the posterior distribution, we can address the question of model complexity from a different perspective. The posterior covariance of the parameters provide a measure of uncertainty of the parameters in credible interval. In addition, when making predictions or learning model evoked responses, we are not limited to working with a single estimate. We can infer statistics of interest by averaging over the posterior with Bayesian integration.

In this chapter we will be concerned with an inference algorithm for the posterior distribution of the parameters from the full probability model of the TMS-EEG data, which extends the model introduced in Chapter 4 by imposing prior distribution over the parameters. Inferencing the posterior distributions involves marginalization over the cortical activity and parameters, which is analytically intractable [110, 111]. In the literature, two classes of approximation inference methods – Markov chain Monte Carlo (MCMC) [112, 113] and variational inference [110, 114, 115, 111] – have been considered. The former methods are asymptotically exact but can be computationally demanding. The latter are based on analytical approximations of the posterior distribution and have the advantage of explicit objective function, less computational overhead and hence scalability to large data applications. We consider the variational inference methods in this chapter for its lower computational expense.

A major challenge in performing variational inference in the state space model (SSM) is to learn the posterior distribution of the cortical activities. Previous work in the literature have proposed using message passing algorithm [110] and Kalman filter with augmented observation [114] to infer the distribution of the cortical activities. However, both considered square state transition matrix $\mathbf{A}$ and none of the work considered autoregressive order $p > 1$. As we have seen in the previous chapters, autoregressive order $p > 1$ is essential in modeling M/EEG signal. Another limitation in the previous work is that they assumed no constraint over the orientations of the states, which in turn leads to their assumptions of identity state noise matrix. Our model assumes that the cortical signals lie in subspaces determined by the regions of interest and the lead field. In this chapter we adapt the augmented Kalman filter procedure [114] by expressing the transition matrix as $\mathbf{A}_s$ as in Chapter 4 to learn the posterior distribution from processes with autoregressive order $p > 1$. We also consider that the cortical signal of fixed orientation and we estimate the orientations $\mathbf{\Lambda}$ and assume that the state covariance matrix is a scaled identity matrix with an unknown scaling factor.

With Bayesian inference, no only can we learn the credible interval of each parameter, we can also impose different priors over the parameters and hence our beliefs for the working mechanism of the system can be deployed to guide our reasoning. In this chapter, we consider two classes of prior over the state transition matrices $\mathbf{A}$ and $\mathbf{B}$. We will show how the inference can be done with both classes of priors and compare their performance over different simulated cases.

The rest of the chapter is structured as follows. Section 2 introduces the full probability model in the Bayesian SSM. Section 3 presents variational inference procedures. Section 4 demonstrates the evaluation of the evidence lower bound and optimization procedures of the hyperparameters. Section 5 demonstrates the use of the proposed methods with several simulated cases and compares the two classes of priors.

5.2 Bayesian Linear State Space Model

In the linear state space model, the measurements from M electrodes at time n from trial j, $\mathbf{y}_{n,j}$ is modeled as

$$\begin{aligned}\mathbf{x}_{n,j} &= \mathbf{A}\mathbf{z}_{n-1,j} + \mathbf{B}\mathbf{u}_{n,j} + \mathbf{w}_{n,j}\\ \mathbf{y}_{n,j} &= \mathbf{H}\mathbf{\Lambda}\mathbf{x}_{n,j} + \mathbf{v}_{n,j}\end{aligned}$$

for time $n = 1, 2, \cdots, N$ and epoch $j = 1, 2, \cdots, J$, where the K by 1 vector $\mathbf{x}_{n,j}$ is the cortical signals from the K regions of interest (ROIs) at time n and trial j, the Kp by 1 vector $\mathbf{z}_{n,j} = [\mathbf{x}_{n,j}^\top, \mathbf{x}_{n-1,j}^\top, \cdots, \mathbf{x}_{n-p+1,j}^\top]^\top$ denotes the system state, the ℓ by 1 vector $\mathbf{u}_{n,j} = [u_{n,j}, u_{n-1,j}, \cdots, u_{n-\ell+1,j}]^\top$ is the exogenous input vector, $u_{n,j}$ is 1 if the stimulation is applied at time n and epoch j and 0 otherwise. The autoregressive matrix $\mathbf{A}$ and the exogenous input matrix $\mathbf{B}$ follow the same formulation and interpretation as in Chapter 4. The K by 1 vector $\mathbf{w}_{n,j}$ is zero-mean Gaussian system noise with covariance matrix $\mathbf{Q}$ and the M by 1 vector $\mathbf{v}_{n,j}$ is zero-mean Gaussian observation noise with covariance matrix $\mathbf{R}$. The initial state $\mathbf{z}_{0,j}$ of each trial is assumed to be Gaussian with mean $\boldsymbol{\pi}_0$ and covariance matrix $\boldsymbol{\Sigma}_0$. The matrix $\mathbf{H}$ can be formed by low-dimensional approximation of the lead field sub-matrix for each ROI, following the descriptions in the previous chapter or [53].

In the following we will consider diagonal state and observation noise covariance matrices. More specifically, we parameterize the state and observation noise precision matrices with scalars ρ_q and ρ_r and we have the precision matrices given as $\boldsymbol{\Lambda}_{\mathbf{Q}} = \mathbf{Q}^{-1} = \rho_q \cdot \mathbf{I}$ and $\boldsymbol{\Lambda}_{\mathbf{R}} = \mathbf{R}^{-1} = \rho_r \cdot \mathbf{I}$.

The complete log likelihood function for the linear state space model is given as

$$\begin{aligned}
&\log p(\mathbf{Y}, \mathbf{Z} | \mathbf{A}, \mathbf{B}, \rho_q, \rho_r, \boldsymbol{\Sigma}_0) \\
= \; & -\frac{1}{2} \sum_{j=1}^{J} \Big[\sum_{n=1}^{N} \big(\mathbf{x}_{n,j} - \mathbf{A}\mathbf{z}_{n-1,j} - \mathbf{B}\mathbf{u}_{n,j}\big)^\top \boldsymbol{\Lambda}_{\mathbf{Q}} \big(\mathbf{x}_{n,j} - \mathbf{A}\mathbf{z}_{n-1,j} - \mathbf{B}\mathbf{u}_{n,j}\big) \\
& + \big(\mathbf{y}_{n,j} - \mathbf{H}\boldsymbol{\Lambda}\mathbf{x}_{n,j}\big)^\top \boldsymbol{\Lambda}_{\mathbf{R}} \big(\mathbf{y}_{n,j} - \mathbf{H}\boldsymbol{\Lambda}\mathbf{x}_{n,j}\big) \\
& + \big(\mathbf{z}_{0,j} - \boldsymbol{\pi}_0\big)^\top \boldsymbol{\Sigma}_0^{-1} \big(\mathbf{z}_{0,j} - \boldsymbol{\pi}_0\big) \Big] + \text{const.} \\
= \; & -\frac{1}{2} \sum_{j=1}^{J} \Big[\sum_{n=1}^{N} \big(\mathbf{z}_{n,j} - \mathbf{A}_s\mathbf{z}_{n-1,j} - \mathbf{B}_s\mathbf{u}_{n,j}\big)^\top \boldsymbol{\Lambda}_{\mathbf{Q},s} \big(\mathbf{z}_{n,j} - \mathbf{A}_s\mathbf{z}_{n-1,j} - \mathbf{B}_s\mathbf{u}_{n,j}\big) \\
& + \big(\mathbf{y}_{n,j} - \mathbf{C}\mathbf{z}_{n,j}\big)^\top \boldsymbol{\Lambda}_{\mathbf{R}} \big(\mathbf{y}_{n,j} - \mathbf{C}\mathbf{z}_{n,j}\big) \\
& + \big(\mathbf{z}_{0,j} - \boldsymbol{\pi}_0\big)^\top \boldsymbol{\Sigma}_0^{-1} \big(\mathbf{z}_{0,j} - \boldsymbol{\pi}_0\big) \Big] + \text{const.}
\end{aligned}$$

where $\mathbf{Y} = [\mathbf{y}_{1,1}, \cdots, \mathbf{y}_{N,1}, \mathbf{y}_{1,2}, \cdots, \mathbf{y}_{N,J}]$ and $\mathbf{Z} = [\mathbf{z}_{1,1}, \cdots, \mathbf{z}_{N,1}, \mathbf{z}_{1,2}, \cdots, \mathbf{z}_{N,J}]$. In the second equality, instead of $\mathbf{x}_{n,j}$, we work with the $\mathbf{z}_{n,j}$ in the state equation, the state transition matrix $\mathbf{A}_s$ and exogenous input matrix $\mathbf{B}_s$ are given as

$$\mathbf{A}_s = \begin{bmatrix} \mathbf{A} & \\ \mathbf{I}_{K(p-1)} & \mathbf{0}_{K(p-1)\times K} \end{bmatrix} \mathbf{B}_s = \begin{bmatrix} \mathbf{B} \\ \mathbf{0}_{K(p-1)\times \ell} \end{bmatrix}.$$

The block-diagonal matrix $\mathbf{\Lambda}_{\mathbf{Q},s} = \text{block\_diag}(\mathbf{Q}^{-1}, \mathbf{0}_{K(p-1)})$ is not a proper precision matrix and serves the purpose for presenting our results hereafter. The observation matrix $\mathbf{C} = [\mathbf{H\Lambda},\ \mathbf{0}_{M\times K(p-1)}]$.

We did not explicitly express the dependence of the complete log likelihood function $\log p(\mathbf{Y}, \mathbf{Z}|\mathbf{A}, \mathbf{B}, \rho_q, \rho_r, \mathbf{\Sigma}_0)$ on $\mathbf{\Lambda}$ and $\boldsymbol{\pi}_0$, as we are only considering priors over $\mathbf{A}, \mathbf{B}, \rho_q, \rho_r$, and $\mathbf{\Sigma}_0$. We will sometimes represent the collection of the parameters $\mathbf{A}, \mathbf{B}, \rho_q, \rho_r$, and $\mathbf{\Sigma}_0$ as θ. Both θ and the individual parameters will be used interchangeably in the following. In our analysis we consider $\mathbf{\Lambda}$ and $\boldsymbol{\pi}_0$ as fixed and unknown parameters. We will discuss how the inference can be done with priors specified over $\mathbf{\Lambda}$ and $\boldsymbol{\pi}_0$ as future works in the last chapter.

Priors over the Parameters

We consider two classes of conjugate priors over the parameters with different structural preference for **A** and **B**. The priors over the rest of the parameters are defined identically for both classes of the priors. In addition to the priors over the parameters, we also specify hierarchical priors over some parameters. The hierarchical priors decrease the number of hyperparameters in our priors. They also encourage sharing of information among different groups of parameters and shrinkage different parameters that depends on common hierarchical priors to common means [113].

Matrix Normal Prior In the matrix normal prior, the priors are given by

$$\begin{aligned} \mathbf{K}_\kappa &\sim \mathcal{W}(\cdot|\nu_\kappa, \mathbf{S}_\kappa) \\ \rho_q &\sim \mathcal{G}(\cdot|\alpha_q, \beta_q) \\ \mathbf{\Xi} &\sim \mathcal{MN}(\cdot|\mathbf{0}, 1/\rho_q \cdot \mathbf{I}, \mathbf{K}_\kappa^{-1}) \\ \rho_r &\sim \mathcal{G}(\cdot|\alpha_r, \beta_r) \\ \rho_{s,j} &\sim \mathcal{G}(\cdot|\alpha_s, \beta_s) \quad j = 1, 2, \cdots, Kp \end{aligned}$$

where $\Xi = [\mathbf{A}, \mathbf{B}]$ and the ρ_{s_j}'s parameterize $\mathbf{\Sigma}_0$ as $\mathbf{\Sigma}_0 = \mathrm{diag}(1/\rho_{s_1}, 1/\rho_{s_2}, \cdots, 1/\rho_{s_{Kp}})$. $\mathbf{K}_\kappa$ is a K-by-K precision matrix specifies the covariance among the columns of Ξ. ν_κ, $\mathbf{S}_\kappa$, α_q, β_q,α_r, β_r, α_s, and β_s are hyperparameters of the prior. The matrix normal distribution, the gamma distribution and the Wishart distribution are given by

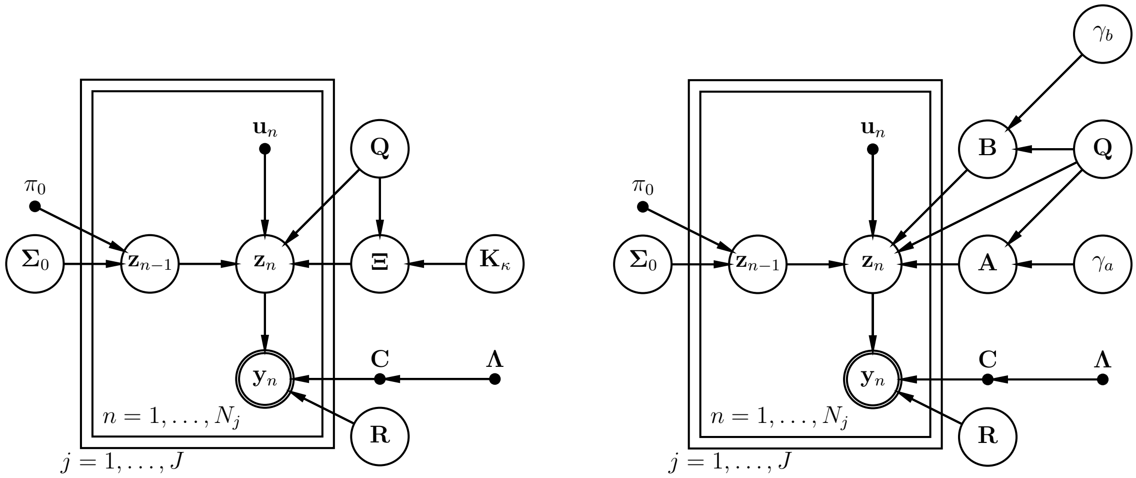
$$\begin{aligned}
\mathcal{MN}(\Xi|\Xi_\mu, \mathbf{\Lambda}_\mathbf{Q}^{-1}, \mathbf{K}_\kappa^{-1}) &= (2\pi)^{-K(Kp+\ell)/2}|\mathbf{K}_\kappa|^{K/2}|\mathbf{\Lambda}_\mathbf{Q}|^{(Kp+\ell)/2} \\
&\quad \cdot \exp\left\{-\frac{1}{2}\mathrm{tr}\left[(\Xi - \Xi_\mu)^\top \mathbf{\Lambda}_\mathbf{Q}(\Xi - \Xi_\mu)\mathbf{K}_\kappa\right]\right\} \\
\mathcal{G}(\rho|a, b) &= \frac{1}{\Gamma(a)} b^a \rho^{a-1} \exp(-b\rho) \\
\mathcal{W}(\mathbf{W}|\nu, \mathbf{S}) &= \left(2^{\nu K/2}\Gamma_K(\nu/2)\right)^{-1}|\mathbf{S}|^{-\nu/2}|\mathbf{W}|^{(\nu-K-1)/2} \\
&\quad \cdot \exp\left(-\frac{1}{2}\mathrm{tr}(\mathbf{S}^{-1}\mathbf{W})\right)
\end{aligned}$$

where $\Gamma(x)$ is the gamma function, $\Gamma_K(x) = \pi^{K(K-1)/4} \prod_{j=1}^K \Gamma(x + (1-j)/2)$ is the multivariate gamma function and $\mathbf{S}$ is a K-by-K positive definite matrix [116, 117]. A graphical model for the prior and the data generative model is depicted in Figure 5.1(A). Combining the priors, we have the matrix normal prior specified as

$$\begin{aligned}
p(\mathbf{K}_\kappa, \rho_q, \Xi, \rho_r, \boldsymbol{\rho}_s) &= p(\mathbf{K}_\kappa)p(\rho_q)p(\Xi|\rho_q, \mathbf{K}_\kappa)p(\rho_r)p(\boldsymbol{\rho}_s) \\
&= p(\mathbf{K}_\kappa)p(\rho_q)p(\Xi|\rho_q, \mathbf{K}_\kappa)p(\rho_r)\prod_{j=1}^{Kp} p(\rho_{s_j}).
\end{aligned}$$

ARD Prior In the automatic relevance determination (ARD) prior [118, 119, 120], the priors over the parameters are given by

$$\begin{aligned}
\gamma_{a,i,j} &\sim \mathcal{G}(\cdot|\alpha_a, \beta_a) \quad i = 1, 2, \cdots, K, j = 1, 2, \cdots, Kp \\
\gamma_{b,i,j} &\sim \mathcal{G}(\cdot|\alpha_b, \beta_b) \quad i = 1, 2, \cdots, K, j = 1, 2, \cdots, \ell \\
\rho_q &\sim \mathcal{G}(\cdot|\alpha_q, \beta_q) \\
\boldsymbol{\xi} &\sim \mathcal{N}(\cdot|\mathbf{0}, (\rho_q \Lambda_\xi)^{-1}) \\
\rho_r &\sim \mathcal{G}(\cdot|\alpha_r, \beta_r) \\
\rho_{s,j} &\sim \mathcal{G}(\cdot|\alpha_s, \beta_s) \quad j = 1, 2, \cdots, Kp
\end{aligned}$$



(A) Matrix normal prior

(B) ARD prior

Figure 5.1: Graphical model for the linear state space model with the two classes of priors. Circular nodes are random and each follows a distribution. Nodes with double circles are observed and those with single circles are latent variables .The directed edges indicate dependence. Nodes represented by single dots are fixed variables. The panels represent repetitions. (A) Matrix normal prior. (B) ARD prior.

where $\xi = \text{vec}(\Xi) = \text{vec}([\mathbf{A},\ \mathbf{B}])$, $\Lambda_\xi = \text{diag}(\boldsymbol{\gamma}_A, \boldsymbol{\gamma}_B)$, $\boldsymbol{\gamma}_A = [\gamma_{a_{1,1}}, \gamma_{a_{2,1}}, \cdots, \gamma_{a_{K,1}}, \gamma_{a_{1,2}}, \cdots, \gamma_{a_{K,Kp}}]^\top$, $\boldsymbol{\gamma}_B = [\gamma_{b_{1,1}}, \gamma_{b_{2,1}}, \cdots, \gamma_{b_{K,\ell}}]^\top$. α_a, β_a, α_b, β_b, α_q, β_q, α_r, β_r, α_s, and β_s are hyperparameters of the prior. $\mathcal{N}(x|\nu, \boldsymbol{\Sigma})$ denotes the multivariate normal distribution. The ARD prior assumes that each system coefficient is independent and Gaussian with precision $\rho_q \gamma_{a_{i,j}}$ (or $\rho_q \gamma_{b_{i,j}}$). The precision coefficients are also learned in our analysis and have the property of pruning individual system coefficient to zero when the corresponding precision $\rho_q \gamma_{a_{i,j}}$ (or $\rho_q \gamma_{b_{i,j}}$) is large. The prior and the data generative model is illustrated in Figure 5.1(B). The product of different factors in the prior is given as

$$\begin{aligned} p(\boldsymbol{\gamma}_A, \boldsymbol{\gamma}_B, \rho_q, \xi, \rho_r, \boldsymbol{\rho}_s) &= \prod_{i=1}^{K} \prod_{j=1}^{Kp} p(\gamma_{a_{i,j}}) \prod_{i'=1}^{K} \prod_{j'=1}^{\ell} p(\gamma_{b_{i',j'}}) p(\rho_q) p(\xi|\rho_q, \boldsymbol{\gamma}_A, \boldsymbol{\gamma}_B) \\ &\cdot p(\rho_r) \prod_{j=1}^{Kp} p(\rho_{s_j}). \end{aligned}$$

We can modify the aforementioned ARD prior – which we will refer to as the independent ARD prior henceforth – by adding structural preference [121]. We follow the same grouping idea as in the self-connected group lasso [57, 122] and define the self-connected ARD prior as

$$\begin{aligned}\gamma_{a_{i,j}} &\sim \mathcal{G}(\cdot|\alpha_a, \beta_a) \quad i = 1, 2, \cdots, K, j = 1, 2, \cdots, K \\ \gamma_{b_i} &\sim \mathcal{G}(\cdot|\alpha_b, \beta_b) \quad i = 1, 2, \cdots, \ell \\ \boldsymbol{\xi} &\sim \mathcal{N}(\cdot|\mathbf{0}, (\rho_q \Lambda_\xi)^{-1})\end{aligned}$$

where $\Lambda_\xi = \text{diag}(\boldsymbol{\gamma}_a, \boldsymbol{\gamma}_b)$, $\boldsymbol{\gamma}_a = \mathbf{1}_p \otimes [\gamma_{a_{1,1}}, \gamma_{a_{2,1}}, \cdots, \gamma_{a_{1,2}}, \cdots, \gamma_{a_{K,K}}]^\top$, $\mathbf{1}_p$ is a p-by-1 all-one vector, $\otimes$ denotes the Kronecker product and $\boldsymbol{\gamma}_b = [\gamma_{b_1}, \gamma_{b_2}, \cdots, \gamma_{b_\ell}]^\top \otimes \mathbf{1}_K$. In the prior we group the coefficients by the two ROIs each coefficient connects and the direction of the connection. Each group comprises the autoregressive coefficients from ROI i to ROI j with delays 1 to p, and the coefficients for ROI j to ROI i belong to a separate group. The self-connected ARD prior will encourage each directed link with different delays be pruned off together in a data driven way. We also group the coefficients in the **B** matrix by the delay in time, which can help in understanding the order of **B**. The self-connected ARD prior is put together as

$$\begin{aligned}p(\boldsymbol{\gamma}_a, \boldsymbol{\gamma}_b, \rho_q, \boldsymbol{\xi}, \rho_r, \boldsymbol{\rho}_s) \;=\; & \prod_{i=1}^{K}\prod_{j=1}^{K} p(\gamma_{a_{i,j}}) \prod_{i'=1}^{\ell} p(\gamma_{b_{i'}}) p(\rho_q) p(\boldsymbol{\xi}|\rho_q, \boldsymbol{\gamma}_a, \boldsymbol{\gamma}_b) \\ & \cdot p(\rho_r) \prod_{j=1}^{Kp} p(\rho_{s_j}).\end{aligned}$$

5.3 Variational Inference

The joint probability of **Y** and θ of our model is given as

$$p(\mathbf{Y}, \theta) = \int p(\theta) p(\mathbf{Y}, \mathbf{Z}|\theta) d\mathbf{Z}.$$

Recall that θ denotes the collections of parameters in the prior. Our goal is to learn $p(\theta|\mathbf{Y})$, which would require marginalization over θ and is intractable as the cortical activities **Z** and the parameters θ are both unknown and coupled in the integral.

Instead of finding the exact posterior distribution, we resort to an approximate function $q(\theta, \mathbf{Z})$ to the true posterior $p(\theta, \mathbf{Z}|\mathbf{Y})$ that minimizes the KL divergence

$\mathrm{KL}(q||p) = \mathrm{E}_{q(\theta,\mathbf{Z})}\big[\log q(\theta,\mathbf{Z}) - \log p(\theta,\mathbf{Z}|\mathbf{Y})\big]$. The KL divergence can be related to the model evidence $p(\mathbf{Y})$ with the equality:

$$\log p(\mathbf{Y}) = \mathcal{L} + \mathrm{KL}(q||p)$$

where the evidence lower bound (ELOB) $\mathcal{L}$ [111] is given by

$$\mathcal{L} = \mathrm{E}_{q(\theta,\mathbf{Z})}[\log p(\mathbf{Y},\mathbf{Z},\theta)] + \mathrm{H}(q)$$

and $\mathrm{H}(q)$ is the differential entropy of q. The ELOB is a lower bound to the logarithm of the model evidence $p(\mathbf{Y})$, which expresses the preference shown by the data for the model [123, 124].

If no constraint is imposed over $q(\cdot)$, the optimum $q(\cdot)$ that minimizes the KL divergence (or maximizes the ELOB) is $p(\theta,\mathbf{Z}|\mathbf{Y})$, which is intractable to find. To solve the problem analytically, we consider approximate posterior distributions that assume a mean field form

$$q(\theta,\mathbf{Z}) = \prod_{i=1}^{N} q(\theta_i)q(\mathbf{Z})$$

where $\{\theta_i\}$ is a partition over θ. With this factorized form, the approximate posterior distribution is iteratively optimized. For each iteration, the update procedure cycles through each factor $q(\theta_i)$ (or $q(\mathbf{Z})$) by maximizing ELOB with respect to $q(\theta_i)$ (or $q(\mathbf{Z})$), with the rest factors fixed [124, 110]:

$$\begin{aligned} q(\theta_i) &\propto \exp\Big\{\mathrm{E}_{q(\mathbf{Z})\prod_{j=1;j\neq i}^{N} q(\theta_j)}\big[\log p(\theta_i|\theta_{A(i)}) + \log p(\mathbf{Y},\mathbf{Z}|\theta)\big]\Big\} \\ &\quad \text{for } i = 1,2,\cdots,N \\ q(\mathbf{Z}) &\propto \exp\Big\{\mathrm{E}_{\prod_{i=1}^{N} q(\theta_i)}\big[\log p(\mathbf{Y},\mathbf{Z}|\theta)\big]\Big\}. \end{aligned}$$

until the convergence of ELOB or a pre-specified maximum number of iterations. In the expression, $\theta_{A(i)}$ denotes the set of the parameters in θ that depends on θ_i, as specified in the prior.

In the following we show how the approximate posterior distribution can be found for each prior. We constrain the search space of the approximate posterior within the mean-field factorized form of $q(\theta,\mathbf{Z})$. For each prior we assume different factorization over $q(\theta)$, but we always assume that $q(\mathbf{Z})$ is independent of the approximate posterior over the parameters. By assuming that $q(\mathbf{Z})$ is independent of the rest factors, we can learn $q(\mathbf{Z})$ following the same procedure for all three priors. We start by showing the procedures for finding the $q(\theta_i)$'s for each prior. Then we demonstrate how $q(\mathbf{Z})$ can be learned by Kalman filtering with augmented observed variables [114].

Update of $q(\theta_i)$

Matrix Normal Prior We assume that the approximate posterior over $\mathbf{K}_\kappa$, ρ_q, Ξ, ρ_r, and $\boldsymbol{\rho}_s$ can be factored as

$$q(\theta) = q(\mathbf{K}_\kappa)q(\rho_q, \Xi)q(\rho_r)\prod_{i=1}^{Kp} q(\rho_{s_i}).$$

We start by finding the updating procedure for $q(\rho_q, \Xi)$. Following the updating equation for $q(\theta_i)$ we have in last subsection, we write out $\log q(\rho_q, \Xi)$ as

$$\begin{aligned}
\log q(\rho_q, \Xi) &= \log p(\rho_q) + \mathsf{E}_{q(\mathbf{K}_\kappa)}\Big[\log p(\Xi|\rho_q, \mathbf{K}_\kappa)\Big] + \mathsf{E}_{q(\mathbf{Z})}\Big[\log p(\mathbf{Z}|\rho_q, \Xi)\Big] + \text{const.} \\
&= (\alpha_q - 1)\log \rho_q - \beta_q \rho_q - \frac{1}{2}\text{tr}\Big[\Xi^\top \boldsymbol{\Lambda}_{\mathbf{Q}} \Xi \cdot \mathsf{E}_{q(\mathbf{K}_\kappa)}\big[\mathbf{K}_\kappa\big]\Big] \\
&\quad + \frac{Kp + \ell}{2}\log|\boldsymbol{\Lambda}_{\mathbf{Q}}| + \mathsf{E}_{q(\mathbf{Z})}\big[\log p(\mathbf{Z}|\rho_q, \Xi)\big] + \text{const.}
\end{aligned}$$

Following similar steps as in [124, 125], we can find $q(\Xi|\rho_q)$ as

$$\begin{aligned}
\log q(\Xi|\rho_q) &= -\frac{\rho_q}{2}\text{tr}\Big[\Xi^\top \Xi \cdot \mathsf{E}_{q(\mathbf{K}_\kappa)}\big[\mathbf{K}_\kappa\big]\Big] \\
&\quad - \frac{\rho_q}{2}\sum_{j=1}^{J}\sum_{n=1}^{N}\text{tr}\Big\{\mathsf{E}_{q(\mathbf{Z})}\big[\mathbf{x}_{n,j}\mathbf{x}_{n,j}^\top\big] - \mathsf{E}_{q(\mathbf{Z})}\big[\mathbf{x}_{n,j}\mathbf{m}_{n,j}^\top\big]\Xi^\top \\
&\quad - \Xi \cdot \mathsf{E}_{q(\mathbf{Z})}\big[\mathbf{m}_{n,j}\mathbf{x}_{n,j}^\top\big] + \Xi \cdot \mathsf{E}_{q(\mathbf{Z})}\big[\mathbf{m}_{n,j}\mathbf{m}_{n,j}^\top\big]\Xi^\top\Big\} + \text{const.} \\
&= -\frac{\rho_q}{2}\text{tr}\Big[\Xi \cdot \mathsf{E}_{q(\mathbf{K}_\kappa)}\big[\mathbf{K}_\kappa\big]\Xi^\top\Big] \\
&\quad - \frac{\rho_q}{2}\text{tr}\big[\mathbf{R}_{\mathbf{XX}} - \mathbf{R}_{\mathbf{XM}}\Xi^\top - \Xi\mathbf{R}_{\mathbf{MX}} + \Xi\mathbf{R}_{\mathbf{MM}}\Xi^\top\big] + \text{const.} \\
&= -\frac{\rho_q}{2}\text{tr}\left\{\Big[\left(\Xi - \mathbf{R}_{\mathbf{XM}}\mathbf{R}_{\bar{\mathbf{M}}\bar{\mathbf{M}}}^{-1}\right)\mathbf{R}_{\bar{\mathbf{M}}\bar{\mathbf{M}}}\left(\Xi - \mathbf{R}_{\mathbf{XM}}\mathbf{R}_{\bar{\mathbf{M}}\bar{\mathbf{M}}}^{-1}\right)^\top + \mathbf{R}_{\mathbf{X}|\bar{\mathbf{M}}}\Big]\right\} + \text{const.}
\end{aligned}$$

where $\mathbf{m}_{n,j} = [\mathbf{z}_{n-1,j}^\top, \mathbf{u}_{n,j}^\top]^\top$,

$$\begin{aligned}
\mathbf{R}_{\mathbf{XX}} &= \sum_{j=1}^{J}\sum_{n=1}^{N} \mathrm{E}_{q(\mathbf{Z})}\left[\mathbf{x}_{n,j}\mathbf{x}_{n,j}^\top\right] \\
\mathbf{R}_{\mathbf{XM}} &= \sum_{j=1}^{J}\sum_{n=1}^{N} \mathrm{E}_{q(\mathbf{Z})}\left[\mathbf{x}_{n,j}\mathbf{m}_{n,j}^\top\right] \\
\mathbf{R}_{\bar{\mathbf{M}}\bar{\mathbf{M}}} &= \mathbf{R}_{\mathbf{MM}} + \mathrm{E}_{q(\mathbf{K}_\kappa)}\left[\mathbf{K}_\kappa\right] = \sum_{j=1}^{J}\sum_{n=1}^{N} \mathrm{E}_{q(\mathbf{Z})}\left[\mathbf{m}_{n,j}\mathbf{m}_{n,j}^\top\right] + \mathrm{E}_{q(\mathbf{K}_\kappa)}\left[\mathbf{K}_\kappa\right] \\
\mathbf{R}_{\mathbf{X}|\bar{\mathbf{M}}} &= \mathbf{R}_{\mathbf{XX}} - \mathbf{R}_{\mathbf{XM}}\mathbf{R}_{\bar{\mathbf{M}}\bar{\mathbf{M}}}^{-1}\mathbf{R}_{\mathbf{XM}}^\top.
\end{aligned}$$

Hence we conclude that

$$q(\Xi|\rho_q) = \mathcal{MN}\left(\Xi; \hat{\Xi}_\mu, \frac{1}{\rho_q}\cdot\mathbf{I}, \hat{\mathbf{K}}^{-1}\right)$$

where $\hat{\Xi}_\mu = \mathbf{R}_{\mathbf{XM}}\mathbf{R}_{\bar{\mathbf{M}}\bar{\mathbf{M}}}^{-1}$ and $\hat{\mathbf{K}} = \mathbf{R}_{\bar{\mathbf{M}}\bar{\mathbf{M}}}$. In the above expressions we can see that, to find $q(\Xi|\rho_q)$ we need $\mathrm{E}_{q(\mathbf{K}_\kappa)}\left[\mathbf{K}_\kappa\right]$ and the expectation over the sufficient statistics formed by $\{\mathbf{z}_{n,j}\}$ with respect to $q(\mathbf{Z})$. We will present the former term later in this subsection and return to the latter in the next subsection.

Next we determine $q(\rho_q)$ using the equality $\log q(\rho_q) = \log q(\rho_q, \Xi) - \log q(\Xi)$, we have

$$\begin{aligned}
\log q(\rho_q) &= (\alpha_q - 1)\log\rho_q - \beta_q\rho_q + \frac{JNK}{2}\log\rho_q - \frac{\rho_q}{2}\mathrm{tr}(\mathbf{R}_{\mathbf{X}|\bar{\mathbf{M}}}) \\
&\quad + \text{const.}
\end{aligned}$$

From which we have

$$q(\rho_q) = \mathcal{G}\left(\rho_q; \hat{\alpha}_q, \hat{\beta}_q\right)$$

where $\hat{\alpha}_q = \alpha_q + JNK/2$ and $\hat{\beta}_q = \beta_q + 1/2\cdot\mathrm{tr}(\mathbf{R}_{\mathbf{X}|\bar{\mathbf{M}}})$.

The updating procedure for $q(\mathbf{K}_\kappa)$ can be found by inspecting

$$\begin{aligned}
\log q(\mathbf{K}_\kappa) &= \log p(\mathbf{K}_\kappa) + \mathrm{E}_{q(\rho_q,\Xi)}\left[\log p(\Xi|1/\rho_q\cdot\mathbf{I}, \mathbf{K}_\kappa^{-1})\right] + \text{const.} \\
&= \log p(\mathbf{K}_\kappa) + \frac{K}{2}\log|\mathbf{K}_\kappa| - \frac{1}{2}\mathrm{tr}\left[\mathbf{K}_\kappa \mathrm{E}_{q(\rho_q,\Xi)}\left[\rho_q\Xi^\mathsf{T}\Xi\right]\right] + \text{const.} \\
&= \left(\frac{\nu_\kappa - 1 + K}{2}\right)\log|\mathbf{K}_\kappa| - \frac{1}{2}\mathrm{tr}\left[\mathbf{K}_\kappa\left(\mathbf{W}_\kappa^{-1} + \mathrm{E}_{q(\rho_q,\Xi)}\left[\rho_q\Xi^\mathsf{T}\Xi\right]\right)\right] + \text{const.}
\end{aligned}$$

In the expression we need $E_{q(\rho_q,\Xi)}\left[\rho_q \Xi^{\mathsf{T}}\Xi\right]$, which can be shown to be [116]

$$E_{q(\rho_q,\Xi)}\left[\rho_q \Xi^{\mathsf{T}}\Xi\right] = K \cdot \hat{\mathbf{K}}^{-1} + \frac{\hat{\alpha}_q}{\hat{\beta}_q} \cdot \hat{\boldsymbol{\Xi}}_{\mu}^{\top}\hat{\boldsymbol{\Xi}}_{\mu}.$$

Thus we conclude that

$$q(\mathbf{K}_{\kappa}) = \mathcal{W}(\mathbf{K}_{\kappa}; \hat{\nu}_{\kappa}, \hat{\mathbf{W}}_{\kappa})$$

where $\hat{\nu}_{\kappa} = \nu_{\kappa} + K$ and $\hat{\mathbf{W}}_{\kappa} = \left[\mathbf{W}_{\kappa}^{-1} + E_{q(\rho_q,\Xi)}\left[\rho_q \Xi^{\mathsf{T}}\Xi\right]\right]^{-1}$. With $q(\mathbf{K}_{\kappa})$ found, it is straight forward that $E_{q(\mathbf{K}_{\kappa})}\left[\mathbf{K}_{\kappa}\right] = \hat{\nu}_{\kappa}\hat{\mathbf{W}}_{\kappa}$.

The updating procedure for $q(\rho_r)$ can be found through

$$\begin{aligned}
\log q(\rho_r) &= \log p(\rho_r) + E_{q(\mathbf{Z})}\left[\log p(\mathbf{Y}|\mathbf{Z}, \rho_r)\right] + \text{const.} \\
&= (\alpha_r - 1)\log\rho_r - \beta_r\rho_r + \frac{JNM}{2} \cdot \log\rho_r - \frac{\rho_r}{2}\sum_{j=1}^{J}\sum_{n=1}^{N} \operatorname{tr}\left[\mathbf{y}_{n,j}\mathbf{y}_{n,j}^{\top}\right. \\
&\quad -\mathbf{y}_{n,j}E_{q(\mathbf{Z})}\left[\mathbf{z}_{n,j}^{\top}\right]\boldsymbol{\Lambda}^{\top}\mathbf{H}^{\top} - \mathbf{H}\boldsymbol{\Lambda}E_{q(\mathbf{Z})}\left[\mathbf{z}_{n,j}\right]\mathbf{y}_{n,j}^{\top} \\
&\quad \left.+\mathbf{H}\boldsymbol{\Lambda} \cdot E_{q(\mathbf{Z})}\left[\mathbf{z}_{n,j}\mathbf{z}_{n,j}^{\top}\right]\boldsymbol{\Lambda}^{\top}\mathbf{H}^{\top}\right] + \text{const.}
\end{aligned}$$

which leads to the updating equation given as

$$q(\rho_r) = \mathcal{G}(\rho_r; \hat{\alpha}_r, \hat{\beta}_r)$$

where $\hat{\alpha}_r = \alpha_r + JNM/2$, $\hat{\beta}_r = \beta_r + 1/2 \cdot \operatorname{tr}\left[\mathbf{R}_{\mathbf{YY}} - \mathbf{R}_{\mathbf{YZ}}\boldsymbol{\Lambda}^{\top}\mathbf{H}^{\top} - \mathbf{H}\boldsymbol{\Lambda}\mathbf{R}_{\mathbf{ZY}} + \mathbf{H}\boldsymbol{\Lambda}\mathbf{R}_{\mathbf{ZZ}}\boldsymbol{\Lambda}^{\top}\mathbf{H}^{\top}\right]$, $\mathbf{R}_{\mathbf{YY}} = \sum_{j=1}^{J}\sum_{n=1}^{N}\mathbf{y}_{n,j}\mathbf{y}_{n,j}^{\top}$, and $\mathbf{R}_{\mathbf{YZ}} = \sum_{j=1}^{J}\sum_{n=1}^{N}\mathbf{y}_{n,j} \cdot E_{q(\mathbf{Z})}\left[\mathbf{z}_{n,j}^{\top}\right]$

Finally the updating equation for $q(\rho_{s_i})$'s can be found by inspecting

$$\begin{aligned}
\log q(\rho_{s_i}) &= \log p(\rho_{s_i}) + \sum_{j=1}^{J} E_{q(\mathbf{Z})}\left[\log p(\mathbf{z}_{0,j}|\boldsymbol{\pi}_0, \boldsymbol{\Sigma}_0)\right] + \text{const.} \\
&= (\alpha_s - 1)\log\rho_{s_i} - \beta_s\rho_{s_i} + \frac{J}{2}\log\rho_{s_i} - \frac{1}{2}[\mathbf{V}_0]_{i,i} \cdot \rho_{s_i} + \text{const.}
\end{aligned}$$

where $\mathbf{V}_0 = \sum_{j=1}^{J} E_{q(\mathbf{Z})}\left[\left(\mathbf{z}_{0,j} - E_{q(\mathbf{Z})}[\mathbf{z}_{0,j}]\right)\left(\mathbf{z}_{0,j} - E_{q(\mathbf{Z})}[\mathbf{z}_{0,j}]\right)^{\top}\right] + \left(E_{q(\mathbf{Z})}[\mathbf{z}_{0,j}] - \boldsymbol{\pi}_0\right)\left(E_{q(\mathbf{Z})}[\mathbf{z}_{0,j}] - \boldsymbol{\pi}_0\right)^{\top}$. Hence we have

$$q(\rho_{s_i}) = \mathcal{G}(\rho_{s_i}; \hat{\alpha}_{s_i}, \hat{\beta}_{s_i})$$

for $i = 1, 2, \cdots, Kp$, $\hat{\alpha}_{s_i} = \alpha_s + J/2$ and $\hat{\beta}_{s_i} = \beta_s + 1/2 \cdot [\mathbf{V}_0]_{i,i}$ and $[\mathbf{M}]_{i,i}$ denotes the i-th diagonal component of the square matrix $\mathbf{M}$.

Independent ARD prior In the independent ARD prior, we assume that the approximate posterior $q(\theta)$ assumes the factorized form of

$$q(\theta) = \prod_{i=1}^{K}\prod_{j=1}^{Kp} q(\gamma_{a_{i,j}}) \prod_{i=1}^{K}\prod_{j=1}^{\ell} q(\gamma_{b_{i,j}}) q(\rho_q, \boldsymbol{\xi}) q(\rho_r) \prod_{i=1}^{Kp} q(\rho_{s_i}).$$

The updating procedure of $q(\rho_r)$ and $q(\rho_{s_i})$'s are the same as those for the matrix normal prior. We will focus on the updating procedures of the rest factors.

We begin by finding the updating equation for $q(\rho_q, \boldsymbol{\xi})$, of which the logarithm takes the form of

$$\begin{aligned} \log q(\rho_q, \boldsymbol{\xi}) &= \log p(\rho_q) + \mathbb{E}_{q(\boldsymbol{\gamma}_A)q(\boldsymbol{\gamma}_B)}\left[\log p(\boldsymbol{\xi}|\rho_q, \boldsymbol{\Lambda}_\xi)\right] \\ &\quad + \mathbb{E}_{q(\mathbf{Z})}\left[\log p(\mathbf{Z}|\rho_q, \boldsymbol{\xi})\right] + \text{const.} \end{aligned}$$

Similar to the derivations of $q(\rho_q, \Xi)$ for the matrix normal prior, we write out $\log q(\boldsymbol{\xi}|\rho_q)$ as

$$\begin{aligned} \log q(\boldsymbol{\xi}|\rho_q) &= \mathbb{E}_{q(\boldsymbol{\gamma}_A)q(\boldsymbol{\gamma}_B)}\left[\log p(\boldsymbol{\xi}|\rho_q, \boldsymbol{\Lambda}_\xi)\right] + \mathbb{E}_{q(\mathbf{Z})}\left[\log p(\mathbf{Z}|\rho_q, \boldsymbol{\xi})\right] + \text{const.} \\ &= -\frac{\rho_q}{2}\boldsymbol{\xi}^\top \mathbb{E}_{q(\boldsymbol{\gamma}_A)q(\boldsymbol{\gamma}_B)}\left[\boldsymbol{\Lambda}_\xi\right]\boldsymbol{\xi} \\ &\quad + \rho_q \boldsymbol{\xi}^\top \text{vec}(\mathbf{R}_{\mathbf{XM}}) - \frac{\rho_q}{2}\boldsymbol{\xi}^\top(\mathbf{R}_{\mathbf{MM}} \otimes \mathbf{I})\boldsymbol{\xi} + \text{const.} \\ &= -\frac{\rho_q}{2}\boldsymbol{\xi}^\top(\mathbb{E}_{q(\boldsymbol{\gamma}_A)q(\boldsymbol{\gamma}_B)}\left[\boldsymbol{\Lambda}_\xi\right] + \mathbf{R}_{\mathbf{MM}} \otimes \mathbf{I})\boldsymbol{\xi} + \rho_q \boldsymbol{\xi}^\top \text{vec}(\mathbf{R}_{\mathbf{XM}}) + \text{const.} \end{aligned}$$

From which we conclude that

$$q(\boldsymbol{\xi}|\rho_q) = \mathcal{N}(\boldsymbol{\xi}; \hat{\boldsymbol{\xi}}_\mu, \frac{1}{\rho_q} \cdot \hat{\boldsymbol{\Lambda}}_\xi^{-1})$$

where $\hat{\boldsymbol{\Lambda}}_\xi = \mathbb{E}_{q(\boldsymbol{\gamma}_A)q(\boldsymbol{\gamma}_B)}\left[\boldsymbol{\Lambda}_\xi\right] + \mathbf{R}_{\mathbf{MM}} \otimes \mathbf{I}$ and $\hat{\boldsymbol{\xi}}_\mu = \hat{\boldsymbol{\Lambda}}_\xi^{-1}\text{vec}(\mathbf{R}_{\mathbf{XM}})$.

With that, we have $\log q(\rho_q) = \log q(\rho_q, \boldsymbol{\xi}) - \log q(\boldsymbol{\xi}|\rho_q)$ given as

$$\begin{aligned} \log q(\rho_q) &= (\alpha_q - 1)\log \rho_q - \beta_q \rho_q + \frac{JNK}{2}\log \rho_q - \frac{\rho_q}{2}\text{tr}(\mathbf{R}_{\mathbf{XX}}) \\ &\quad + \frac{\rho_q}{2}\hat{\boldsymbol{\xi}}_\mu^\top \hat{\boldsymbol{\Lambda}}_\xi \hat{\boldsymbol{\xi}}_\mu + \text{const.} \end{aligned}$$

From which we conclude that

$$q(\rho_q) = \mathcal{G}(\rho_q; \hat{\alpha}_q, \hat{\beta}_q)$$

where $\hat{\alpha}_q = \alpha_q + JNK/2$ and $\hat{\beta}_q = \beta_q + 1/2 \cdot \left[\mathrm{tr}(\mathbf{R}_{\mathbf{XX}}) - \hat{\boldsymbol{\xi}}_\mu^\top \hat{\boldsymbol{\Lambda}}_\xi \hat{\boldsymbol{\xi}}_\mu\right]$.

As for the updating procedure for the $q(\gamma_{a_{i,j}})$'s, we have $\log q(\gamma_{a_{i,j}})$ expressed as

$$\begin{aligned}\log q(\gamma_{a_{i,j}}) &= \log p(\gamma_{a_{i,j}}) + E_{q(\rho_q,\xi)}\left[\log p(\boldsymbol{\xi}|\rho_q, \boldsymbol{\Lambda}_\xi)\right] + \text{const.}\\ &= (\alpha_a - 1)\log\gamma_{a_{i,j}} - \beta_a \gamma_{a_{i,j}} + \frac{1}{2}\log\gamma_{a_{i,j}}\\ &\quad -\gamma_{a_{i,j}} \cdot \frac{1}{2}\left[E_{q(\rho_q,\xi)}\left[\rho_q \boldsymbol{\xi}\boldsymbol{\xi}^\top\right]\right]_{(j-1)K+i,(j-1)K+i} + \text{const.}\end{aligned}$$

where $E_{q(\rho_q,\xi)}\left[\rho_q \boldsymbol{\xi}\boldsymbol{\xi}^\top\right] = \hat{\boldsymbol{\Lambda}}_\xi^{-1} + \hat{\alpha}_q/\hat{\beta}_q \cdot \hat{\boldsymbol{\xi}}_\mu \hat{\boldsymbol{\xi}}_\mu^\top$. Hence we have

$$q(\gamma_{a_{i,j}}) = \mathcal{G}(\gamma_{a_{i,j}}; \hat{\alpha}_{a_{i,j}}, \hat{\beta}_{a_{i,j}})$$

for $i = 1, 2, \cdots, K$, $j = 1, 2, \cdots, Kp$ where $\hat{\alpha}_{a_{i,j}} = \alpha_a + 1/2$ and $\hat{\beta}_{a_{i,j}} = \beta_a + 1/2 \cdot \left[E_{q(\rho_q,\xi)}\left[\rho_q \boldsymbol{\xi}\boldsymbol{\xi}^\top\right]\right]_{(j-1)K+i,(j-1)K+i}$. Similarly, we have

$$q(\gamma_{b_{i,j}}) = \mathcal{G}(\gamma_{b_{i,j}}; \hat{\alpha}_{b_{i,j}}, \hat{\beta}_{b_{i,j}})$$

for $i = 1, 2, \cdots, K$, $j = 1, 2, \cdots, \ell$ where $\hat{\alpha}_{b_{i,j}} = \alpha_b + 1/2$ and $\hat{\beta}_{b_{i,j}} = \beta_b + 1/2 \cdot \left[E_{q(\rho_q,\xi)}\left[\rho_q \boldsymbol{\xi}\boldsymbol{\xi}^\top\right]\right]_{K^2p+(j-1)K+i,K^2p+(j-1)K+i}$.

Self-Connected ARD prior In the self-connected ARD prior, we consider $q(\theta)$ with the factorized form of

$$q(\theta) = \prod_{i=1}^{K}\prod_{j=1}^{K} q(\gamma_{a_{i,j}}) \prod_{i=1}^{\ell} q(\gamma_{b,i}) q(\rho_q, \boldsymbol{\xi}) q(\rho_r) \prod_{i=1}^{Kp} q(\rho_{s_i}).$$

The updating procedure for all factors are the same as those of the independent ARD prior expect for $q(\gamma_{a_{i,j}})$ and $q(\gamma_{b_i})$.

We can again find the updating procedure for $q(\gamma_{a_{i,j}})$ by inspecting $\log q(\gamma_{a_{i,j}})$

$$\begin{aligned}\log q(\gamma_{a_{i,j}}) &= \log p(\gamma_{a_{i,j}}) + \mathrm{E}_{q(\rho_q,\boldsymbol{\xi})}\left[\log p(\boldsymbol{\xi}|\rho_q,\boldsymbol{\Lambda}_{\xi})\right] + \text{const.}\\ &= (\alpha_a - 1)\log\gamma_{a_{i,j}} - \beta_a\gamma_{a_{i,j}} + \frac{p}{2}\log\gamma_{a_{i,j}}\\ &\quad -\gamma_{a_{i,j}}\cdot\frac{1}{2}\sum_{m=1}^{p}\left[\mathrm{E}_{q(\rho_q,\boldsymbol{\xi})}\left[\rho_q\boldsymbol{\xi}\boldsymbol{\xi}^\top\right]\right]_{(m-1)K^2+(j-1)K+i,(m-1)K^2+(j-1)K+i}\\ &\quad +\text{const.}\end{aligned}$$

Hence we have

$$q(\gamma_{a_{i,j}}) = \mathcal{G}(\gamma_{a_{i,j}}; \hat{\alpha}'_{a_{i,j}}, \hat{\beta}'_{a_{i,j}})$$

for $i = 1, 2, \cdots, K$, $j = 1, 2, \cdots, K$ where $\hat{\alpha}'_{a_{i,j}} = \alpha_a + p/2$ and $\hat{\beta}'_{a_{i,j}} = \beta_a + 1/2 \cdot \sum_{m=1}^{p}\left[\mathrm{E}_{q(\rho_q,\boldsymbol{\xi})}\left[\rho_q\boldsymbol{\xi}\boldsymbol{\xi}^\top\right]\right]_{(m-1)K^2+(j-1)K+i,(m-1)K^2+(j-1)K+i}$. Similarly,

$$q(\gamma_{b_i}) = \mathcal{G}(\gamma_{b_i}; \hat{\alpha}'_{b_i}, \hat{\beta}'_{b_i})$$

for $i = 1, 2, \cdots, \ell$, where $\hat{\alpha}'_{b_i} = \alpha_b + K/2$ and $\hat{\beta}'_{b_i} = \beta_b + 1/2 \cdot \sum_{m=1}^{K}\left[\ \mathrm{E}_{q(\rho_q,\boldsymbol{\xi})}\left[\ \rho_q\boldsymbol{\xi}\boldsymbol{\xi}^\top\right]\right]_{K^2p+K(i-1)+m,K^2p+K(i-1)+m}$.

Update of $q(\mathbf{Z})$

As can be seen from the previous section, in order to update $q(\theta_i)$, instead of $q(\mathbf{Z})$ itself, we are mainly interested in the expectation of the sufficient statistics $\mathrm{E}_{q(\mathbf{Z})}\left[\mathbf{z}_{n,j}\right]$, $\mathrm{E}_{q(\mathbf{Z})}\left[\mathbf{z}_{n,j}\mathbf{z}_{n,j}^\top\right]$, and $\mathrm{E}_{q(\mathbf{Z})}\left[\mathbf{z}_{n,j}\mathbf{z}_{n-1,j}^\top\right]$. In the following we will begin by inspecting $\log q(\mathbf{Z})$. Then we follow the procedures as proposed in [114] to augment the measurement $\mathbf{y}_{n,j}$ to derive Kalman filtering and smoothing procedure to find the moments.

The logarithm of the updating step for $q(\mathbf{Z})$ can be expressed as

$$
\begin{aligned}
\log q(\mathbf{Z}) &= \mathrm{E}_{q(\rho_q,\mathbf{A},\mathbf{B})q(\rho_r)q(\boldsymbol{\rho}_s)}\big[\log p(\mathbf{Y},\mathbf{Z}|\mathbf{A},\mathbf{B},\rho_q,\rho_r,\boldsymbol{\rho}_s)\big] + \text{const.} \\
&= -\frac{1}{2}\sum_{j=1}^{J}\Bigg\{\sum_{n=1}^{N}\mathrm{E}\big[\rho_r\big](\mathbf{y}_{n,j}-\mathbf{C}\mathbf{z}_{n,j})^\top(\mathbf{y}_{n,j}-\mathbf{C}\mathbf{z}_{n,j}) \\
&\quad +\mathrm{E}\big[(\mathbf{z}_{n,j}-\mathbf{A}_s\mathbf{z}_{n-1,j}-\mathbf{B}_s\mathbf{u}_{n,j})^\top\boldsymbol{\Lambda}_{\mathbf{Q},s}(\mathbf{z}_{n,j}-\mathbf{A}_s\mathbf{z}_{n-1,j}-\mathbf{B}_s\mathbf{u}_{n,j})\big] \\
&\quad +(\mathbf{z}_{0,j}-\boldsymbol{\pi}_0)^\top\mathrm{E}\big[\boldsymbol{\Sigma}_0^{-1}\big](\mathbf{z}_{0,j}-\boldsymbol{\pi}_0)\Big\} + \text{const.} \\
&= -\frac{1}{2}\sum_{j=1}^{J}\Bigg\{\sum_{n=1}^{N}\mathrm{E}\big[\rho_r\big](\mathbf{y}_{n,j}-\mathbf{C}\mathbf{z}_{n,j})^\top(\mathbf{y}_{n,j}-\mathbf{C}\mathbf{z}_{n,j}) \\
&\quad +(\mathbf{z}_{n,j}-\mathrm{E}\big[\mathbf{A}_s\big]\mathbf{z}_{n-1,j}-\mathrm{E}\big[\mathbf{B}_s\big]\mathbf{u}_{n,j})^\top\mathrm{E}\big[\boldsymbol{\Lambda}_{\mathbf{Q},s}\big](\mathbf{z}_{n,j}-\mathrm{E}\big[\mathbf{A}_s\big]\mathbf{z}_{n-1,j}-\mathrm{E}\big[\mathbf{B}_s\big]\mathbf{u}_{n,j}) \\
&\quad +\begin{bmatrix}\mathbf{z}_{n-1,j}\\ \mathbf{u}_{n,j}\end{bmatrix}^\top\begin{bmatrix}\boldsymbol{\Sigma}_A & \boldsymbol{\Sigma}_{AB}\\ \boldsymbol{\Sigma}_{AB}^\top & \boldsymbol{\Sigma}_B\end{bmatrix}\begin{bmatrix}\mathbf{z}_{n-1,j}\\ \mathbf{u}_{n,j}\end{bmatrix} \\
&\quad +(\mathbf{z}_{0,j}-\boldsymbol{\pi}_0)^\top\mathrm{E}\big[\boldsymbol{\Sigma}_0^{-1}\big](\mathbf{z}_{0,j}-\boldsymbol{\pi}_0)\Big\} + \text{const.}
\end{aligned}
$$

where $\boldsymbol{\Sigma}_A = \mathrm{E}\big[\mathbf{A}^\top\boldsymbol{\Lambda}_\mathbf{Q}\mathbf{A}\big]-\mathrm{E}\big[\mathbf{A}^\top\big]\mathrm{E}\big[\boldsymbol{\Lambda}_\mathbf{Q}\big]\mathrm{E}\big[\mathbf{A}\big]$, $\boldsymbol{\Sigma}_{AB} = \mathrm{E}\big[\mathbf{A}^\top\boldsymbol{\Lambda}_\mathbf{Q}\mathbf{B}\big]-\mathrm{E}\big[\mathbf{A}^\top\big]\mathrm{E}\big[\boldsymbol{\Lambda}_\mathbf{Q}\big]\mathrm{E}\big[\mathbf{B}\big]$, and $\boldsymbol{\Sigma}_B = \mathrm{E}\big[\mathbf{B}^\top\boldsymbol{\Lambda}_\mathbf{Q}\mathbf{B}\big]-\mathrm{E}\big[\mathbf{B}^\top\big]\mathrm{E}\big[\boldsymbol{\Lambda}_\mathbf{Q}\big]\mathrm{E}\big[\mathbf{B}\big]$. With the second equality, we have separated the variance over the fluctuations of the parameters from the mean in the quadratic forms. In the above expressions $\mathrm{E}[\mathbf{A}]$ and $\mathrm{E}[\mathbf{B}]$ can be obtained from the mean of $q(\Xi|\rho_q)$ for the matrix normal prior or the mean of $q(\xi|\rho_q)$ for the ARD priors. $\mathrm{E}[\mathbf{A}_s]$ and $\mathrm{E}[\mathbf{B}_s]$ are formed by substituting $\mathrm{E}[\mathbf{A}]$ and $\mathrm{E}[\mathbf{B}]$ for $\mathbf{A}$ and $\mathbf{B}$ in $\mathbf{A}_s$ and $\mathbf{B}_s$, respectively. In addition, $\mathrm{E}\big[\boldsymbol{\Lambda}_\mathbf{Q}\big] = \hat{\alpha}_q/\hat{\beta}_q\cdot\mathbf{I}$ and $\mathrm{E}\big[\boldsymbol{\Lambda}_{\mathbf{Q},s}\big] = \text{block\_diag}(\mathrm{E}\big[\boldsymbol{\Lambda}_\mathbf{Q}\big],\mathbf{0})$. For all three priors, we can evaluate $\mathrm{E}_{q(\rho_q,\Xi)}\Big[(\Xi-\mathrm{E}\big[\Xi\big])^\top\boldsymbol{\Lambda}_\mathbf{Q}(\Xi-\mathrm{E}\big[\Xi\big])\Big]$ then index proper submatrices to get $\boldsymbol{\Sigma}_A$, $\boldsymbol{\Sigma}_{AB}$, and $\boldsymbol{\Sigma}_B$. With the matrix normal prior, $\mathrm{E}_{q(\rho_q,\Xi)}\big[(\Xi-\mathrm{E}\big[\Xi\big])^\top\boldsymbol{\Lambda}_\mathbf{Q}(\Xi-\mathrm{E}\big[\Xi\big])\big] = \mathrm{E}_{q(\rho_q,\Xi)}\big[\rho_q\cdot(\Xi-\mathrm{E}\big[\Xi\big])^\top(\Xi-\mathrm{E}\big[\Xi\big])\big] = K\cdot\hat{\mathbf{K}}^{-1}$. With the ARD priors, we have

$$
\begin{aligned}
&\text{vec}\Big(\mathrm{E}_{q(\rho_q,\xi)}\big[(\Xi-\mathrm{E}\big[\Xi\big])^\top\boldsymbol{\Lambda}_\mathbf{Q}(\Xi-\mathrm{E}\big[\Xi\big])\big]\Big)^\top \\
= \;&\text{vec}\Big(\mathrm{E}_{q(\rho_q,\xi)}\big[\rho_q(\Xi-\mathrm{E}\big[\Xi\big])^\top(\Xi-\mathrm{E}\big[\Xi\big])\big]\Big)^\top \\
= \;&\text{vec}(\mathbf{I})^\top\mathrm{E}_{q(\rho_q,\xi)}\Big[\rho_q\cdot(\Xi-\hat{\Xi}_\mu)\otimes(\Xi-\hat{\Xi}_\mu)\Big]
\end{aligned}
$$

where $E_{q(\rho_q,\xi)}\left[\rho_q \cdot (\Xi - \hat{\Xi}_\mu) \otimes (\Xi - \hat{\Xi}_\mu)\right]$ can be obtained by rearranging the components of $E_{q(\rho_q,\xi)}[\rho_q \cdot (\xi - \hat{\xi}_\mu)(\xi - \hat{\xi}_\mu)^\top] = \hat{\boldsymbol{\Lambda}}_\xi^{-1}$. Hence $E_{q(\rho_q,\Xi)}\left[(\Xi - E\left[\Xi\right])^\top \boldsymbol{\Lambda}_{\mathbf{Q}}(\Xi - E\left[\Xi\right])\right]$ can be obtained by rearranging $\hat{\boldsymbol{\Lambda}}_\xi^{-1}$ and aggregating the components.

We can further re-express $\log q(\mathbf{Z})$ by grouping the terms that involve $\mathbf{z}_{n,j}$ based on time index n

$$
\begin{aligned}
\log q(\mathbf{Z}) \;=\; & -\frac{1}{2}\sum_{j=1}^{J}\left\{\sum_{n=1}^{N-1}\left[E\left[\rho_r\right](\mathbf{y}_{n,j} - \mathbf{C}\mathbf{z}_{n,j})^\top(\mathbf{y}_{n,j} - \mathbf{C}\mathbf{z}_{n,j})\right.\right.\\
& +\mathbf{z}_{n,j}^\top \boldsymbol{\Sigma}_A \mathbf{z}_{n,j} + 2\mathbf{u}_{n+1,j}^\top \boldsymbol{\Sigma}_{AB}^\top \mathbf{z}_{n,j}\\
& \left.+(\mathbf{z}_{n,j} - E\left[\mathbf{A}_s\right]\mathbf{z}_{n-1,j} - E\left[\mathbf{B}_s\right]\mathbf{u}_{n,j})^\top E\left[\boldsymbol{\Lambda}_{\mathbf{Q},s}\right](\mathbf{z}_{n,j} - E\left[\mathbf{A}_s\right]\mathbf{z}_{n-1,j} - E\left[\mathbf{B}_s\right]\mathbf{u}_{n,j})\right]\\
& +E\left[\rho_r\right](\mathbf{y}_{N,j} - \mathbf{C}\mathbf{z}_{N,j})^\top(\mathbf{y}_{N,j} - \mathbf{C}\mathbf{z}_{N,j})\\
& +(\mathbf{z}_{N,j} - E\left[\mathbf{A}_s\right]\mathbf{z}_{N-1,j} - E\left[\mathbf{B}_s\right]\mathbf{u}_{N,j})^\top E\left[\boldsymbol{\Lambda}_{\mathbf{Q},s}\right]\\
& \quad \cdot(\mathbf{z}_{N,j} - E\left[\mathbf{A}_s\right]\mathbf{z}_{N-1,j} - E\left[\mathbf{B}_s\right]\mathbf{u}_{N,j})\\
& \left.+(\mathbf{z}_{0,j} - \boldsymbol{\pi}_0)^\top E\left[\boldsymbol{\Sigma}_0^{-1}\right](\mathbf{z}_{0,j} - \boldsymbol{\pi}_0) + \mathbf{z}_{0,j}^\top \boldsymbol{\Sigma}_A \mathbf{z}_{0,j} + 2\mathbf{u}_{1,j}^\top \boldsymbol{\Sigma}_{AB}^\top \mathbf{z}_{0,j}\right\} + \text{const.}
\end{aligned}
$$

We define the augmented observation variable as $\tilde{\mathbf{y}}_{n,j} = [\mathbf{y}_{n,j}^\top, -\mathbf{u}_{n+1,j}^\top \mathbf{L}_B, \mathbf{0}_{kp\times 1}^\top]^\top$, for $n = 1, 2, \cdots, N-1$, and $j = 1, 2, \cdots, J$. We also augment the observation matrix and observation covariance matrix as: $\tilde{\mathbf{C}} = \left[\mathbf{C}^\top, \boldsymbol{\Sigma}_{AB}(\mathbf{L}_B^{-1})^\top, \mathbf{L}_{A|B}\right]^\top$ and $\tilde{\boldsymbol{\Lambda}}_{\mathbf{R}}$ = block_diag($E\left[\rho_r\right]\mathbf{I}_k$, $\mathbf{I}_\ell$, $\mathbf{I}_{np}$) where $\boldsymbol{\Sigma}_B = \mathbf{L}_B \mathbf{L}_B^\top$, $\boldsymbol{\Sigma}_A - \boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_B^{-1}\boldsymbol{\Sigma}_{AB}^\top = \mathbf{L}_{A|B}\mathbf{L}_{A|B}^\top$. Finally, we let $\tilde{\boldsymbol{\Sigma}}_0 = (E\left[\boldsymbol{\Sigma}_0^{-1}\right] + \boldsymbol{\Sigma}_A)^{-1}$ and $\tilde{\boldsymbol{\pi}}_{\mathbf{0}} = \tilde{\boldsymbol{\Sigma}}_0\Big(E\left[\boldsymbol{\Sigma}_0^{-1}\right]\boldsymbol{\pi}_0 - \boldsymbol{\Sigma}_{AB}\mathbf{u}_{1,j}\Big)$. We can rewrite $\log q(\mathbf{Z})$ as

$$
\begin{aligned}
& \log q(\mathbf{Z})\\
= \; & -\frac{1}{2}\sum_{j=1}^{J}\left\{\sum_{n=1}^{N-1}\left[(\tilde{\mathbf{y}}_{n,j} - \tilde{\mathbf{C}}\mathbf{z}_{n,j})^\top \tilde{\boldsymbol{\Lambda}}_{\mathbf{R}}(\tilde{\mathbf{y}}_{n,j} - \tilde{\mathbf{C}}\mathbf{z}_{n,j})\right.\right.\\
& \left.+(\mathbf{z}_{n,j} - E\left[\mathbf{A}_s\right]\mathbf{z}_{n-1,j} - E\left[\mathbf{B}_s\right]\mathbf{u}_{n,j})^\top E\left[\boldsymbol{\Lambda}_{\mathbf{Q},s}\right](\mathbf{z}_{n,j} - E\left[\mathbf{A}_s\right]\mathbf{z}_{n-1,j} - E\left[\mathbf{B}_s\right]\mathbf{u}_{n,j})\right]\\
& +E\left[\rho_r\right](\mathbf{y}_{N,j} - \mathbf{C}\mathbf{z}_{N,j})^\top(\mathbf{y}_{N,j} - \mathbf{C}\mathbf{z}_{N,j})\\
& +(\mathbf{z}_{N,j} - E\left[\mathbf{A}_s\right]\mathbf{z}_{N-1,j} - E\left[\mathbf{B}_s\right]\mathbf{u}_{N,j})^\top E\left[\boldsymbol{\Lambda}_{\mathbf{Q},s}\right](\mathbf{z}_{N,j} - E\left[\mathbf{A}_s\right]\mathbf{z}_{N-1,j} - E\left[\mathbf{B}_s\right]\mathbf{u}_{N,j})\\
& \left.+(\mathbf{z}_{0,j} - \tilde{\boldsymbol{\pi}}_0)^\top \tilde{\boldsymbol{\Sigma}}_0^{-1}(\mathbf{z}_{0,j} - \tilde{\boldsymbol{\pi}}_0)\right\} + \text{const.}
\end{aligned}
$$

As we noted before, what is of main interest is the expectation of the sufficient statistics. With the above expressions, we can follow the Kalman filter procedure (as listed in Algorithm 1) and the standard Kalman smoother procedure (with $\mathsf{E}\left[\mathbf{A}_s\right]$ and $\mathsf{E}\left[\mathbf{B}_s\right]$ substituted for $\mathbf{A}_s$ and $\mathbf{B}_s$, respectively, c.f. Appendix A) to find the expectation of the sufficient statistics [126]. Note that similar to the EM algorithm for TMS-EEG data, in filtering and smoothing the covariance matrices $\mathbf{V}_{n|n'}$ and cross covariance matrices $\mathbf{V}_{n,n-1|n'}$, we only need to compute the filtering and smoothing procedure over a single epoch, as the expected values of the model coefficients remain the same throughout the epochs. Similar to the EM algorithm, the sufficient statistics are given by

$$\begin{aligned}
\mathsf{E}_{q(\mathbf{Z})}\left[\mathbf{z}_{n,j}\right] &= \mathbf{z}_{n|N,j} \quad n = 0, 1, \cdots, N \\
\mathsf{E}_{q(\mathbf{Z})}\left[\mathbf{z}_{n,j}\mathbf{z}_{n,j}^\top\right] &= \mathbf{V}_{n|N} + \mathbf{z}_{n|N,j}\mathbf{z}_{n|N,j}^\top \quad n = 0, 1, \cdots, N \\
\mathsf{E}_{q(\mathbf{Z})}\left[\mathbf{z}_{n,j}\mathbf{z}_{n-1,j}^\top\right] &= \mathbf{V}_{n,n-1|N} + \mathbf{z}_{n|N,j}\mathbf{z}_{n-1|N,j}^\top \quad n = 1, 2, \cdots, N
\end{aligned}$$

for $j = 1, 2, \cdots, J$.

Algorithm 1 Kalman Filter

```
procedure FORWARD
    $\mathbf{z}_{0|0,j} = \tilde{\boldsymbol{\pi}}_0$
    $\mathbf{V}_{0|0} = \tilde{\boldsymbol{\Sigma}}_0$
    for $n = 1$ to $N - 1$ do
        $\mathbf{z}_{n|n-1,j} = \mathsf{E}\left[\mathbf{A}_s\right]\mathbf{z}_{n-1|n-1,j} + \mathsf{E}\left[\mathbf{B}_s\right]\mathbf{u}_{n,j}$
        $\mathbf{V}_{n|n-1} = \mathsf{E}\left[\mathbf{A}_s\right]\mathbf{V}_{n-1|n-1}\mathsf{E}\left[\mathbf{A}_s^\top\right] + \text{block\_diag}(\mathsf{E}\left[\boldsymbol{\Lambda}_{\mathbf{Q}}\right]^{-1}, \mathbf{0})$
        $\mathbf{K} = \mathbf{V}_{n|n-1}\tilde{\mathbf{C}}^\top(\tilde{\mathbf{C}}\mathbf{V}_{n|n-1}\tilde{\mathbf{C}}^\top + \tilde{\boldsymbol{\Lambda}}_{\mathbf{R}}^{-1})^{-1}$
        $\mathbf{z}_{n|n,j} = \mathbf{z}_{n|n-1,j} + \mathbf{K}(\tilde{\mathbf{y}}_{n,j} - \tilde{\mathbf{C}}\mathbf{z}_{n|n-1,j})$
        $\mathbf{V}_{n|n} = (\mathbf{I} - \mathbf{K}\tilde{\mathbf{C}})\mathbf{V}_{n|n-1}$
    end for
    $\mathbf{z}_{N|N-1,j} = \mathsf{E}\left[\mathbf{A}_s\right]\mathbf{z}_{N-1|N-1,j} + \mathsf{E}\left[\mathbf{B}_s\right]\mathbf{u}_{N,j}$
    $\mathbf{V}_{N|N-1} = \mathsf{E}\left[\mathbf{A}_s\right]\mathbf{V}_{N-1|N-1}\mathsf{E}\left[\mathbf{A}_s^\top\right] + \text{block\_diag}(\mathsf{E}\left[\boldsymbol{\Lambda}_{\mathbf{Q}}\right]^{-1}, \mathbf{0})$
    $\mathbf{K} = \mathbf{V}_{N|N-1}\mathbf{C}^\top(\mathbf{C}\mathbf{V}_{N|N-1}\mathbf{C}^\top + \mathsf{E}\left[\rho_r\right]^{-1} \cdot \mathbf{I})^{-1}$
    $\mathbf{z}_{N|N,j} = \mathbf{z}_{N|N-1,j} + \mathbf{K}(\tilde{\mathbf{y}}_{N,j} - \mathbf{C}\mathbf{z}_{N|N-1,j})$
    $\mathbf{V}_{N|N} = (\mathbf{I} - \mathbf{K}\mathbf{C})\mathbf{V}_{N|N-1}$
end procedure
```

5.4 Evidence Lower Bound

In this section, we discuss the evaluation of the evidence lower bound (ELOB), which is used to check convergence of the variational inference procedure. The ELOB, as we have seen before, is given by

$$\begin{aligned}\mathcal{L} &= \mathrm{E}_q\big[\log p(\mathbf{Y},\mathbf{Z},\theta)\big] + \mathrm{H}(q)\\ &= \mathrm{E}_{q(\theta)q(\mathbf{Z})}\big[\log p(\mathbf{Y},\mathbf{Z}|\theta)\big] - \mathrm{E}_{q(\mathbf{Z})}\big[\log q(\mathbf{Z})\big] - \mathrm{KL}(q(\theta)\|p(\theta)).\end{aligned}$$

The KL divergence in the above expression differs among the priors and the evaluation of this term is standard, which we report in Appendix C.2. As for the first two terms, [110] showed how they can be evaluated with autoregressive model of order 1. Here we extend their work to order $p > 1$. We start by inspecting the two terms:

$$\begin{aligned}&\mathrm{E}_{q(\theta)q(\mathbf{Z})}\big[\log p(\mathbf{Y},\mathbf{Z}|\theta)\big] - \mathrm{E}_{q(\mathbf{Z})}\big[\log q(\mathbf{Z})\big]\\ =\ &\mathrm{E}_{q(\theta)q(\mathbf{Z})}\big[\log p(\mathbf{Y},\mathbf{Z}|\theta)\big] - \mathrm{E}_{q(\mathbf{Z})}\big[\mathrm{E}_{q(\theta)}\big[\log p(\mathbf{Y},\mathbf{Z}|\theta)\big] - \log Z'\big]\\ =\ &\log Z'\end{aligned}$$

where $\log Z' = \log \int \exp\big(\mathrm{E}_{q(\theta)}\big[\log p(\mathbf{Y},\mathbf{Z}|\theta)\big]\big)\,d\mathbf{Z}$. The evaluation of $\log Z'$ follows closely the procedures for evaluating the log likelihood function in the EM algorithm [10] and was addressed in the accompanying code of [114]. Our model is differing from [114] in our consideration of $p > 1$ and cortical signal orientations $\boldsymbol{\Lambda}$, We report the derivation and evaluation procedures for $\log Z'$ in Appendix C.1.

Hyperparameter Learning

The hyperparameters in the priors can be optimized by maximizing the ELOB. We proceed by taking derivative of the ELOB with respect to the hyperparameter and find the updates by solving the equation with the derivative equated to 0. We are mainly interested in optimizing the hyperparameters of two classes of distributions: the gamma distribution and the Wishart distribution. We will use the optimizing steps of the hyperparameters of $p(\rho_s)$ and $p(\mathbf{K}_\kappa)$ as an example.

We begin by finding the updating procedure for the hyperparameters of $p(\rho_s)$.

Setting the derivatives of ELOB with respect to α_s and β_s to 0, we have the equations:

$$\begin{aligned}\psi(\alpha_s) &= \log\beta_s + \frac{1}{Kp}\sum_{i=1}^{Kp} E_{q(\rho_s)}\big[\log\rho_{s,i}\big] \\ \frac{1}{\beta_s} &= \frac{1}{\alpha_s\cdot Kp}\sum_{i=1}^{Kp} E_{q(\rho_s)}\big[\rho_{s,i}\big].\end{aligned}$$

From which we can simplify by setting $c = 1/Kp\cdot\sum_{i=1}^{Kp} E_{q(\rho_s)}\big[\log\rho_{s,i}\big]$, $d = 1/Kp\cdot\sum_{i=1}^{Kp} E_{q(\rho_s)}\big[\rho_{s,i}\big]$ and $\beta_s = \alpha_s/d$ to arrive at

$$\psi(\alpha_s) - \log\alpha_s + \log d - c = 0.$$

Hence the solution α_s to the above equation is the update to α_s. Following [110], we optimize α_s by with iterating with the Newton step

$$\alpha_s^{\text{new}} \leftarrow \alpha_s \exp\left(-\frac{\psi(\alpha_s)-\log\alpha_s+\log d - c}{\alpha_s\psi'(\alpha_s)-1}\right)$$

where $\psi(x)$ is the trigamma function. Upon convergence, we update α_s by setting it to α_s^{new} and update β_s by setting it to α_s^{new}/d.

The hyperparameters of $p(\mathbf{K}_\kappa)$ can be found similarly by setting the derivatives with respect to ν_k and $\mathbf{W}_k$ to zero:

$$\begin{aligned}\psi_{Kp+\ell}(\frac{\nu_k}{2}) &= -\log|\mathbf{W}_k| - (Kp+\ell)\log 2 + E_{q(\mathbf{K}_\kappa)}\big[\log|\mathbf{K}_\kappa|\big] \\ \nu_k\mathbf{W}_k^{-1} &= \mathbf{W}_k^{-1}E_{q(\mathbf{K}_\kappa)}\big[\mathbf{K}_\kappa\big]\mathbf{W}_k^{-1}.\end{aligned}$$

Let $c = -(Kp+\ell)\log 2 + E_{q(\mathbf{K}_\kappa)}\big[\log|\mathbf{K}_\kappa|\big]$, $\mathbf{D} = E_{q(\mathbf{K}_\kappa)}\big[\mathbf{K}_\kappa\big]$, and $\mathbf{W}_k = 1/\nu_k\cdot\mathbf{D}$, we have

$$\psi_{Kp+\ell}(\frac{\nu_k}{2}) - (Kp+\ell)\log\nu_k + \log|\mathbf{D}| - c = 0.$$

We again iterate over the follow Newton step to find the update for ν_k [127]:

$$\begin{aligned}\nu_k^{\text{new}} \leftarrow\;& (\nu_k - (Kp+\ell)+1) \\ &\cdot\exp\left[\frac{\psi_{Kp+\ell}(\nu_k/2) - (Kp+\ell)\log\nu_k + \log|\mathbf{D}| - c}{(\nu_k-(Kp+\ell)+1)/2\cdot\psi'_{Kp+\ell}(\nu_k/2) - (Kp+\ell)(\nu_k-(Kp+\ell)+1)/\nu_k}\right] \\ &+(Kp+\ell)-1\end{aligned}$$

where $\psi_{Kp+\ell}(x)$ is the multivariate trigamma function. On convergence the ν_k is set to ν_k^{new} and $\mathbf{W}_k$ is set to $\mathbf{D}/\nu_k^{\text{new}}$.

Finally, we update π_0 and $\boldsymbol{\Lambda}$ by

$$\pi_0 = \sum_{j=1}^{J} \mathbf{z}_{0,j|N}$$

and

$$\begin{aligned}\boldsymbol{\Lambda} &= \arg\max_{\Lambda} -\frac{1}{2}\sum_{j=1}^{J}\sum_{n=1}^{N-1}\left[\mathbf{z}_{n,j|N}^{\top}\tilde{\mathbf{C}}^{\top}\tilde{\mathbf{C}}\mathbf{z}_{n,j|N} - 2\mathbf{z}_{n,j|N}^{\top}\tilde{\mathbf{C}}^{\top}\tilde{\mathbf{y}}_{n,j}\right] \\ &\quad +\mathbf{z}_{N,j|N}^{\top}\mathbf{C}^{\top}\mathbf{C}\mathbf{z}_{N,j|N} - 2\mathbf{z}_{N,j|N}^{\top}\mathbf{C}^{\top}\mathbf{y}_{N,j} \\ &\quad \text{subject to } ||\lambda_i|| = 1 \quad \text{for } i = 1, 2, \cdots, K\end{aligned}$$

where $\mathbf{C} = [\mathbf{H}\boldsymbol{\Lambda},\ \mathbf{0}]$, $\boldsymbol{\Lambda} = \text{block\_diag}(\lambda_1, \lambda_2, \cdots, \lambda_K)$. We note that the updating steps for π_0 and $\boldsymbol{\Lambda}$ resembles those of M-step in the EM algorithm for the maximum likelihood estimate. Similar to EM algorithm, an iteration of Kalman filtering and smoothing of $q(\mathbf{Z})$ and updating π_0 and $\boldsymbol{\Lambda}$ monotonically increases a lower bound to $\log Z'$.

5.5 Simulation Results

In the first simulation we simulate data from a true model with six ROIs and MVAR model order p of ten. We consider a sparse network that contains one-third of the possible connections. The coefficients in $\mathbf{A}$ are turned on/off groupwise as in the self-connected prior, six out of the 30 non-diagonal connections are present with the rest being off. All diagonal connections are present. Data of four epoch sizes: four, sixteen, 64, and 256 epochs are generated. For each epoch size we simulate ten different realizations of data. All ten realizations follow the same network matrix $\mathbf{A}$ and state-space model parameters. The model is set up so that the average evoked response is 10 dB larger than the spontaneous signal - the observed signal components due to the non-deterministic components in the cortical signal. In other words, the spontaneous signal is $\mathbf{C}\mathbf{z}_{n,j}$ minus the average evoked response. Moreover the spontaneous signal is 3 dB larger than the observed noise $\mathbf{v}_{n,j}$.

In Figures 5.2 and 5.3 we depict Hinton diagrams of the **A** matrices in the true model and of the posterior mean for different priors and different number of epochs in a single realization. The Hinton diagram represents each coefficient in the **A** matrix as a square. The color of the squares denotes the sign of the coefficients: white (+) and black (-) and the area of the squares denotes the magnitude of the coefficients. The similarity in the patterns between the posterior mean and the true coefficient shows the performance of the estimated posterior distribution. We observe that the self-connected ARD prior, having more information about the true model, outperforms the other two priors. Even with relatively few number of epochs, the posterior mean of the self-connected ARD prior detects the sparsity pattern in the true model. The independent ARD prior is inferior to the self-connected ARD prior but also identifies the pattern with intermediate size of data (64 epochs). The posterior mean of the matrix normal prior is always dense, which is what we would expect from the prior assumption. The difference in the posterior means of the priors are more significant when data is of smaller size. As the number of epochs increase, the difference becomes less significant. This observation agrees with our understanding of Bayesian inference in that the prior beliefs affect the posterior more when we have less data and the posterior mean of a conjugate prior becomes the maximum likelihood estimate when we have infinite amount of data.

In Figure 5.4, the posterior mean-squared error (MSE) in **A** is depicted. The posterior MSE is defined as

$$\text{MSE} = \sum_{i=1}^{K} \sum_{j=1}^{Kp} E[(a_{i,j} - \bar{a}_{i,j})^2]$$

where the expectation is taken over the posterior marginal distribution $q(\mathbf{A})$ and $a_{i,j}$ denote the (i, j)-th component of **A**, which is a random variable drawn from $q(\mathbf{A})$ and $\bar{a}_{i,j}$ is the (i, j)-th component of the true autoregressive coefficient $\bar{\mathbf{A}}$. In Figure 5.4 (A), we show the MSE over the coefficients that corresponds to the 'Off' connections in **A**. The MSE can be analytical evaluated as individual coefficient in **A** is Gaussian given the state precision ρ_q and is marginally t-distributed. In the figure, each point is the average posterior MSEs, averaged over the posterior MSEs from each of the ten data realizations. The error bars indicate the standard deviations of the posterior MSEs. The posterior MSEs show similar trends as we observe from the

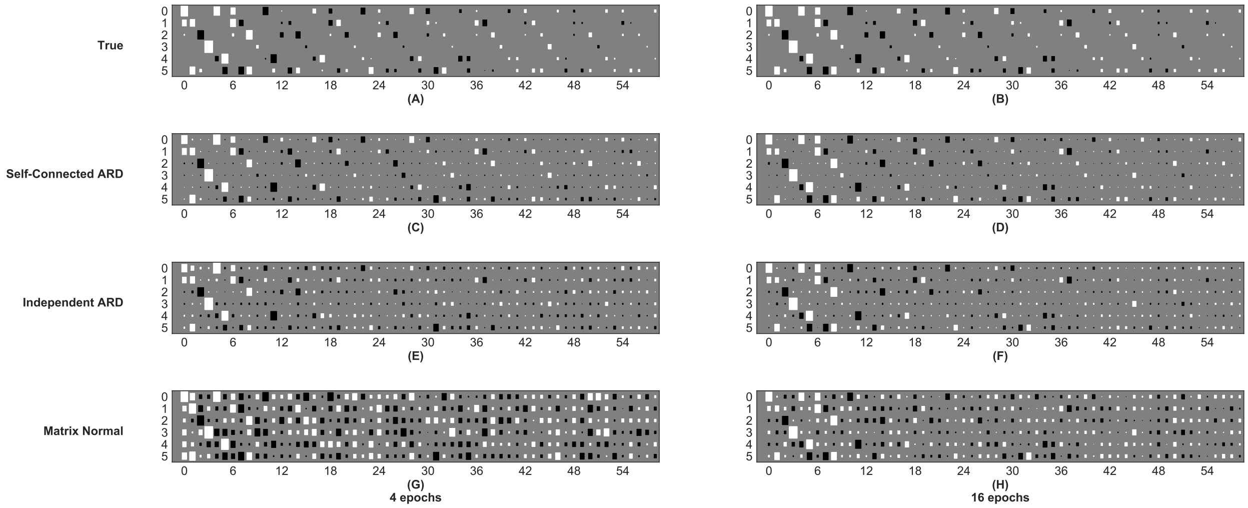


Figure 5.2: **Hinton diagram of the sparse network.** In each panel we show the Hinton diagrams of **A** matrices of the true model and the posterior mean of the approximate posterior with different priors and different number of epochs. The true model is a sparse network in that only 6 out of 30 non-diagonal connections in the true model exists. The diagonal connections are always present. Each rows from top to bottom correspond to: (A) - (B) True model. (E) - (H) Self-connected ARD prior. (I) - (L) Independent ARD Prior. (M) - (P) Matrix normal prior. Columns from left to right correspond to: 4 and 16 epochs.

Hinton diagrams of a single realization in Figures 5.2 and 5.3. The self-connected ARD outperforms the rest two priors and the MSE and the difference among the three priors decreases as we have more data. In Figure 5.4 (B) we show the MSE over the 'On' connections and in Figure 5.4 (C) the normalized mean squared error (NMSE) over the 'On' connections is depicted. The NMSE is the ratio of the sum of mean squared errors of all 'On' coefficients to the sum of squared 'On' coefficient values and is given as

$$\mathrm{NMSE} = \frac{\sum_{\{i,j|\bar{a}_{i,j}\neq 0\}} E[(a_{i,j} - \bar{a}_{i,j})^2]}{\sum_{\{i,j|\bar{a}_{i,j}\neq 0\}} \bar{a}_{i,j}^2}.$$

The NMSE indicates how well the posteriors characterize the true coefficients with larger values. The NMSE shows similar trends as the MSE of the 'Off' connections (in Figure 5.4 (A)) but the difference between the priors are less significant. The

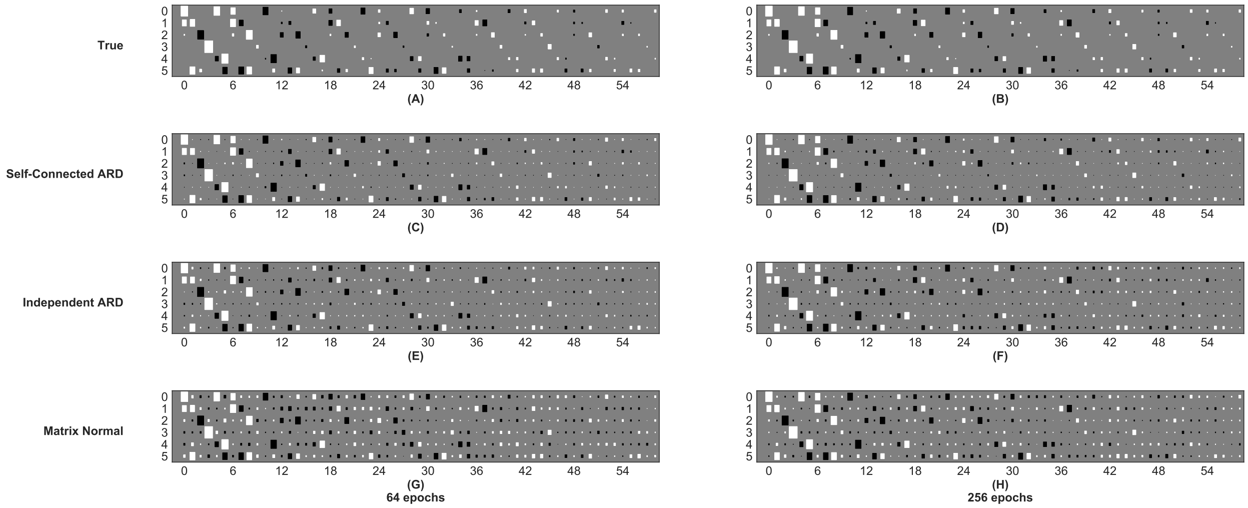
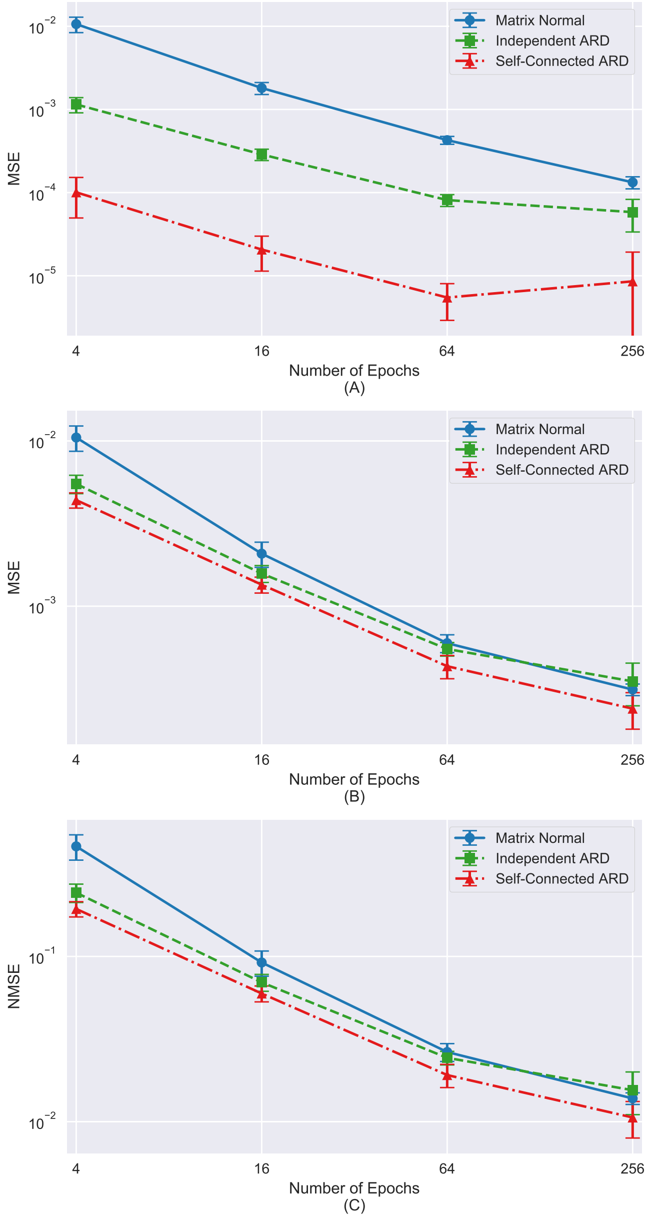


Figure 5.3: **Hinton diagram of the sparse network.** In each panel we show the Hinton diagrams of **A** matrices of the true model and the posterior mean of the approximate posterior with different priors and different number of epochs. The true model is a sparse network in that only 6 out of 30 non-diagonal connections in the true model exists. The diagonal connections are always present. Each rows from top to bottom correspond to: (A) - (D) True model. (E) - (H) Self-connected ARD prior. (I) - (L) Independent ARD Prior. (M) - (P) Matrix normal prior. Columns from left to right correspond to: 64 nd 256 epochs.

posterior variance of the 'Off' coefficients of the ARD priors are shrunken to zero. On the other hand the posterior of the 'On' coefficients remain non-zero valued. Thus the difference in the NMSE among the priors are less significant over the 'On' coefficients. The self-connected ARD prior outperforms the other two priors as we have coded more information about the structure of the true model in the prior.

In Figure 5.5, we illustrate how the posterior distribution characterizes the evoked response. The model evoked response are generated with the posterior mean of **A** and **B**, which are the maximum a posteriori estimates of **A** and **B** from the approximate posteriors. Each point is the average mean squared error over the 10 realizations and the error bars indicate the standard deviations. The trend in the MSE of the evoked response resembles the trend in the NMSE of the 'On' coefficients but the differences among the priors are less significant.

In the second simulation we simulate data from a true model of a dense network



(A)

(B)

(C)

Figure 5.4: **Posterior mean squared error in A, with the true A being a sparse network.** Each data point is the average posterior mean squared error in **A** where the average is done across the posterior distributions obtained from each of the 10 realizations. The error bars show the standard deviation in the mean squared errors. (A) Mean squared error over zero-valued true coefficients ('Off' connections). (B) Mean squared error over non-zero-valued true coefficients ('On' connections). (C) Normalized mean squared error over non-zero-valued true coefficients ('On' connections).

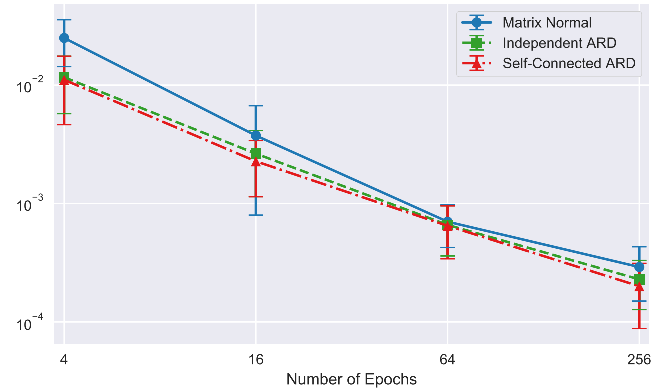


Figure 5.5: **Mean squared error in the evoked response of the sparse network.**

over six ROIs. The coefficients in **A** are turned on/off groupwise. The six diagonal connections are all present. 24 out of the 30 non-diagonal connections are present with the rest being off. Thus the network is more dense compared to the first network. We consider data of four epoch sizes: four, sixteen, 64, and 256 epochs are generated. For each epoch size we simulate 10 different realizations of data from the same model **A** and state-space model parameters.

In Figures 5.6 and 5.7, we show the Hinton diagrams of the **A** matrices of the true model and the posterior mean of the approximate posterior for each prior. Compared to Figures 5.2 and 5.3, the difference across the posteriors are less significant in this case. In Figure 5.8 we depict the posterior MSE in **A**. Similar to what we see from the Hinton diagrams in Figures 5.6 and 5.7, the difference among the priors are less significant, compared to the sparse true model in the first example. We still see some performance gain from the ARD priors when we have less data. This is intuitive as the ARD priors would prune away the posterior mean and covariance of the zero-valued true coefficients. The matrix normal prior outperform the ARD priors in estimating the 'On' coefficients when the data is of larger size, which means the ARD prior can overshrink the estimates. On the other hand the matrix normal prior doesn't encode structure preference and outperforms in estimating the actual value of the 'On' coefficients when the data is larger enough.

In Figure 5.9, we show the mean squared error in the evoked response, the model evoked response is generated by the posterior mean of **A** and **B**. The difference among the priors is insignificant for this case, compared to what we see from the sparse network in Figure 5.5.

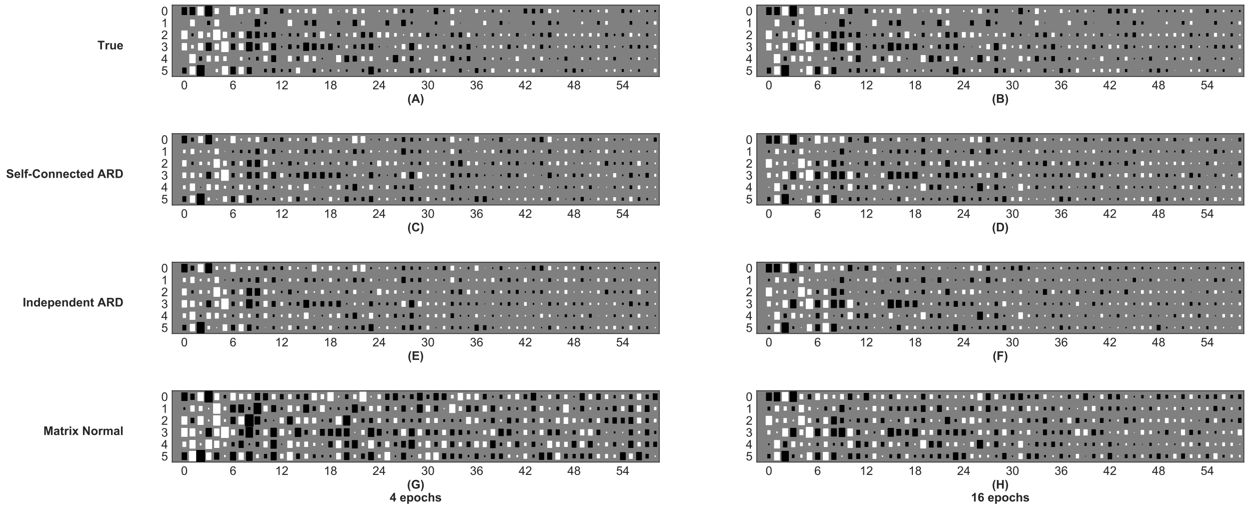


Figure 5.6: **Hinton diagram of the dense network.** Hinton diagram of **A** matrices of the true model and the posterior mean of the approximate posterior with different priors and different number of epochs. The true model is a dense network in that 24 out of 30 non-diagonal connections in the true model exists. The diagonal connections are always present. Each rows from top to bottom correspond to: (A) - (D) True model. (E) - (H) Self-connected ARD prior. (I) - (L) Independent ARD Prior. (M) - (P) Matrix normal prior. Columns from left to right correspond to: 4 and 16 epochs.

In the third simulated case, we revisit the sparse model as in the first case. We generate ten different sparse networks, each of them has six out of 30 non-diagonal groups of coefficients being switched on, but the six groups vary across the ten networks. In addition, all six diagonal group of connections are on. Figure 5.10 depicts the average posterior MSE over the ten networks. The performance difference among the priors over the 'Off' coefficients trend similarly as we observed from Figure 5.4. That is, the self-connected ARD prior outperform the rest two priors as it encodes the information of sparsity and group structure. The independent ARD prior comes in second since it encodes sparsity. Moreover the similarity in trends among the panels in Figures 5.4 and 5.10 shows the stability of our algorithm. Despite the difference in network structure, the algorithm concludes with agreeing result among networks with shared structure properties. On the other hand, the difference in NMSE in the 'On' coefficients in Figure 5.10 among the priors is less significant,

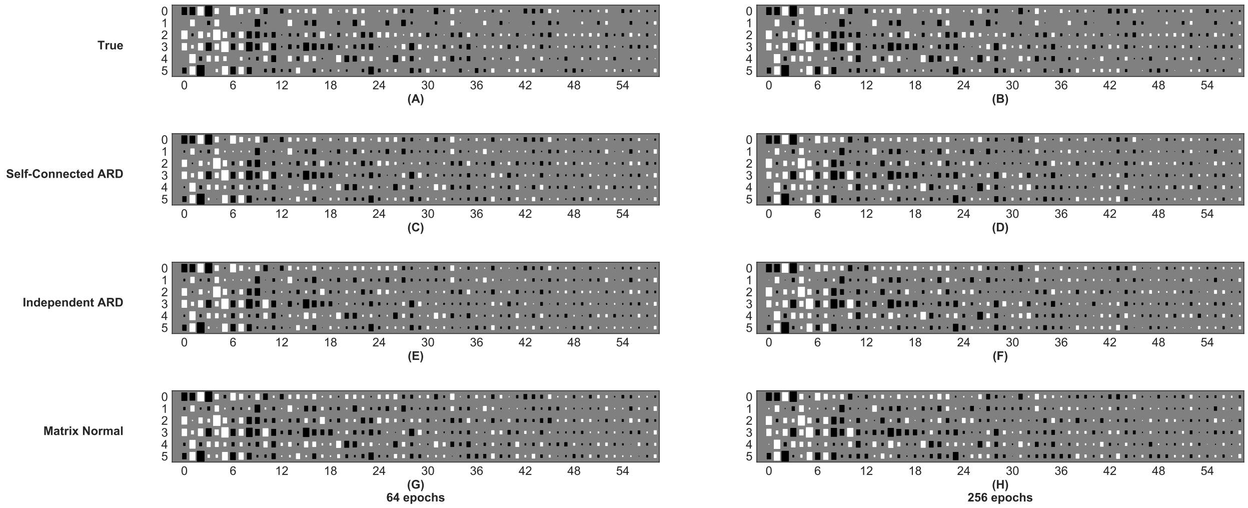


Figure 5.7: **Hinton diagram of the dense network.** Hinton diagram of **A** matrices of the true model and the posterior mean of the approximate posterior with different priors and different number of epochs. The true model is a dense network in that 24 out of 30 non-diagonal connections in the true model exists. The diagonal connections are always present. Each rows from top to bottom correspond to: (A) - (D) True model. (E) - (H) Self-connected ARD prior. (I) - (L) Independent ARD Prior. (M) - (P) Matrix normal prior. Columns from left to right correspond to: 64 and 256 epochs.

compared to what we saw in Figure 5.4, but we still observe the superiority of the self-connected ARD prior.

In the fourth simulated case, we consider the patterns in which the coefficients don't switch on/off groupwise. Out of the 300 non-diagonal coefficients in the **A** matrix, 60 of them are switched on. The 60 diagonal coefficients are all 'On'. The number of 'On' coefficients remain unchanged from the previous simulation. The sets of 'On' coefficients vary across the ten simulated networks. In Figure 5.11 we show the average posterior MSE among the ten networks. One notable difference from Figure 5.10 is that the independent ARD prior outperforms the other two priors, particularly over the 'Off' coefficients. This again shows that the prior helps identifying the pattern if we code the proper structural belief in the prior.

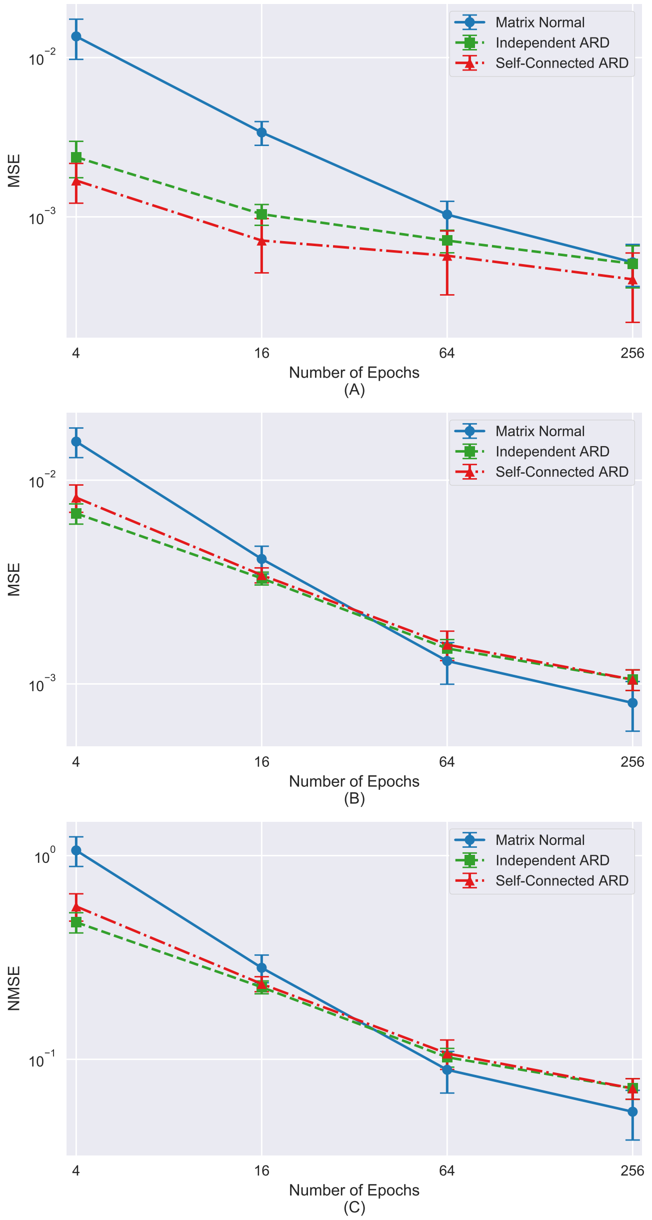


Figure 5.8: **Posterior mean squared error in A, with the true A being a dense network.** Each data point is the average posterior mean squared error in **A** where the average is done across the posterior distributions obtained from each of the 10 realizations. The error bars show the standard deviation in the mean squared errors. (A) Mean squared error over zero-valued true coefficients ('Off' connections). (B) Mean squared error over non-zero-valued true coefficients ('On' connections). (C) Normalized mean squared error over non-zero-valued true coefficients ('On' connections).

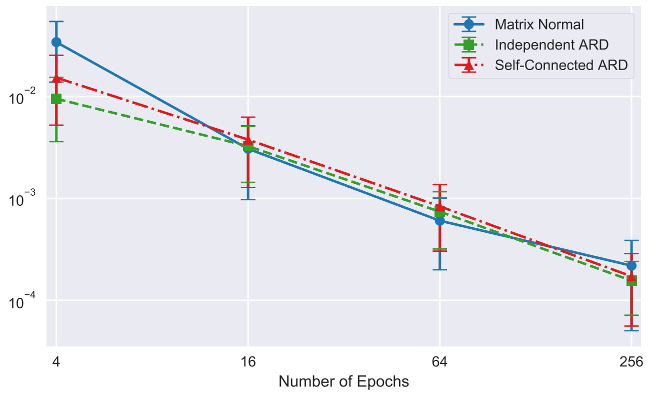


Figure 5.9: **Mean squared error in the evoked response of the dense network.**

5.6 Conclusion

In this chapter we presented the mean field variational inference procedure for the Bayesian linear state-space model of the TMS-EEG data. We considered cortical signals of fixed but unknown orientations and autoregressive model of order $p > 1$. We considered two classes of priors with different structural preferences: the matrix normal prior and the ARD prior. We presented variational inference procedures for learning approximate posteriors of the parameters for both classes of priors. We compared these priors by studying several simulated cases and we showed that the priors, when specified properly with structural information, help identify the actual network structure.

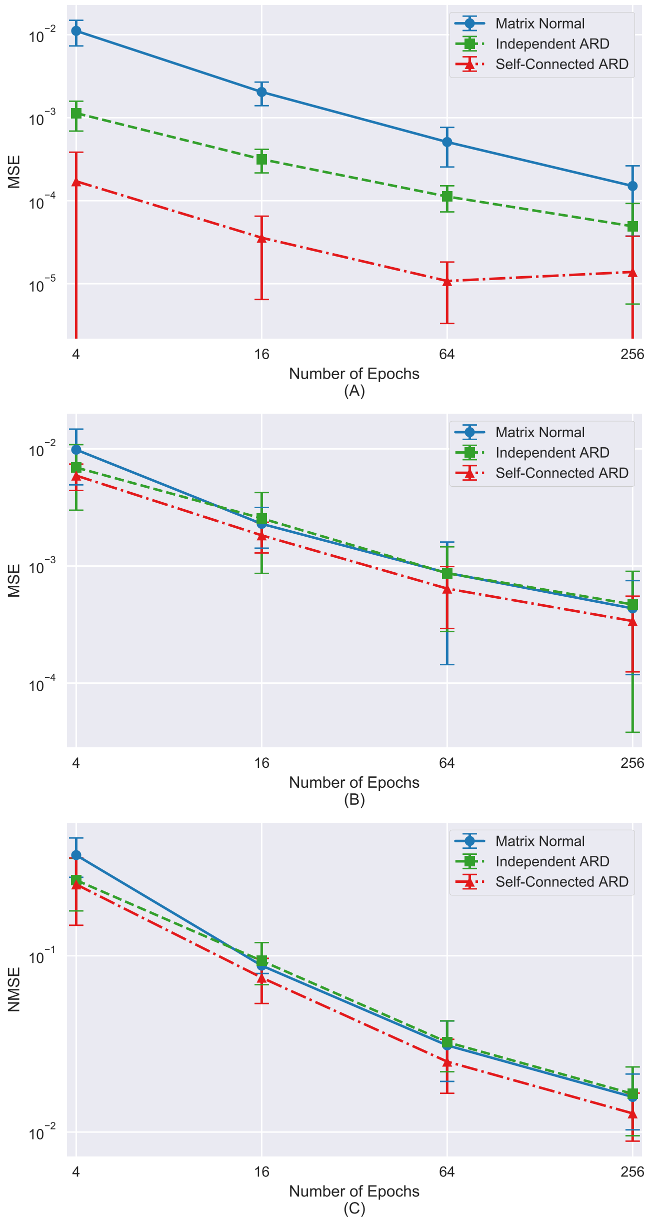


Figure 5.10: **Posterior mean squared error in A, with A being of ten different sparse networks with group structure.** Each data point is the average posterior mean squared error in **A** where the average is done across the posterior distributions obtained from each of the ten different sparse networks. The coefficients are switched on group-wise. The error bars show the standard deviation in the mean squared errors. (A) Mean squared error over zero-valued true coefficients ('Off' connections). (B) Mean squared error over non-zero-valued true coefficients ('On' connections). (C) Normalized mean squared error over non-zero-valued true coefficients ('On' connections).

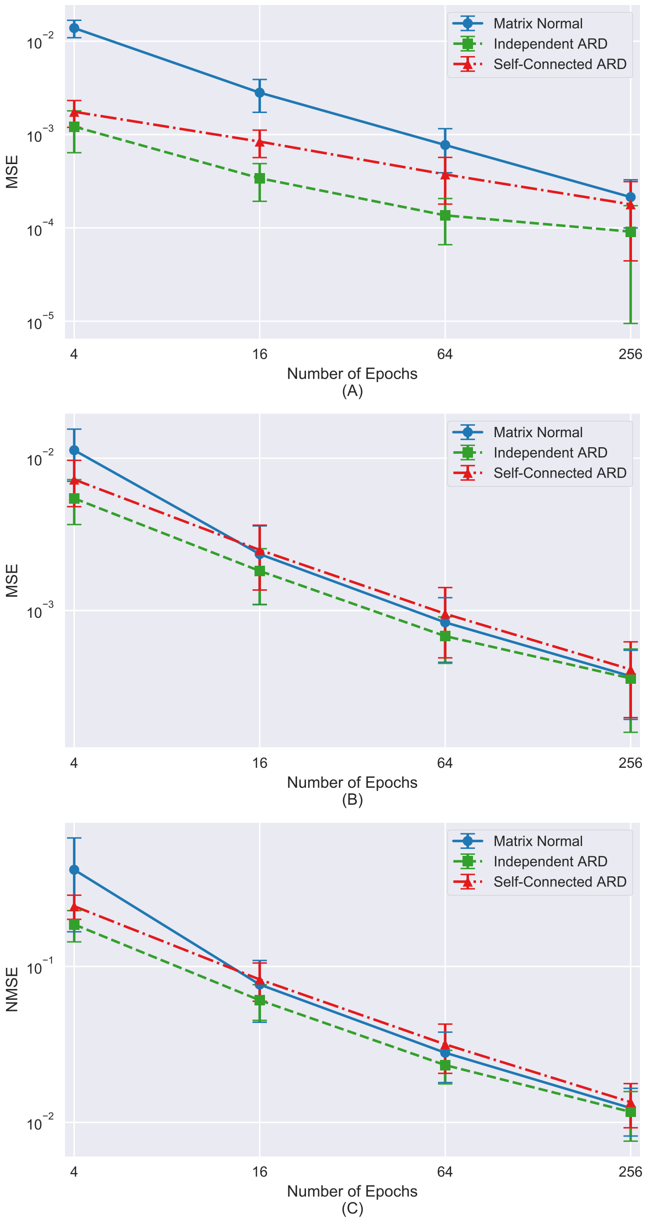


Figure 5.11: **Posterior mean squared error in A, with A being of ten different sparse networks with no group structure.** Each data point is the average posterior mean squared error in **A** where the average is done across the posterior distributions obtained from each of the ten different sparse networks. The coefficients are switched on independently. The error bars show the standard deviation in the mean squared errors. (A) Mean squared error over zero-valued true coefficients ('Off' connections). (B) Mean squared error over non-zero-valued true coefficients ('On' connections). (C) Normalized mean squared error over non-zero-valued true coefficients ('On' connections).

6 FUTURE WORK AND SUMMARY

6.1 Future Work

In the previous chapters we showed how the MVARX model can be used to model the cortical level and source level measurements. We demonstrated how structural assumption of the models can be utilized to show difference in model complexity between models of data in wakefulness and sleep. We also explored the potential of learning the structure of brain network without explicitly imposing constraints with the self-connected group lasso regularization and Bayesian inference procedures.

While we have investigated analysis of evoked response triggered by TMS during wakefulness and sleep, the functional interactions of the brain during various forms of stimulation remain to be explored. An interesting future work is to analyze evoked responses triggered by brain stimulation modalities such as sensory motor stimulation, optogenetic stimulation, and deep brain stimulation.

Instead of explicitly adopting structural hypotheses to learn different models as we did in integrated and segregated models, a future work would be to utilize the self-connected group lasso regularization in the linear state space model to learn the structural pattern of brain network in a data-driven way. The Bayesian linear state space model we presented in Chapter 5 can also be used to identify the network structure. In particular, one can follow similar procedures as in DCM to perform model selection by comparing the evidence lower bounds of priors with different structural preferences [5].

Another interesting future work is to analyze multiple subjects under the same condition altogether. Recall that the Bayesian priors discussed in Chapter 5 are hierarchical priors. Specifically, we can learn common posteriors of the higher level parameters for all subjects and learn independent posteriors of the lower level coefficients for each subject. For instance, with the independent ARD prior, we can learn group level means $\hat{\alpha}_{a_{i,j}}/\hat{\beta}_{a_{i,j}}$ of the posteriors of the gamma distribution that the precision coefficients draw from and learn the individual posterior mean of the precision $\hat{\gamma}_{a_{i,j}}$ for each subject. The group level means of posterior distributions can be utilized to compare different groups of subjects. The difference between group level means between, for instance, groups of subjects in wakefulness and sleep, can help us learn new insights about the brain.

Finally in Chapters 4 and 5 we assumed that the source orientation $\boldsymbol{\lambda}$ is of fixed direction throughout the measurement session. The assumption can be relaxed by assuming that the orientation varies across trials or within a trial for every several milliseconds. We can also consider assessing the assumption by extending the Bayesian framework. For example we can assume that the source orientations of each source in different trials are independent and identically distributed and follow the von Mises-Fisher distribution [128]. With the prior, the posteriors of the source orientations will be the vector Bingham-von Mises-Fisher distribution and the posterior means cannot be analytically evaluated. However one can apply the Gibbs sampling procedures proposed in [128] to estimate the posterior means. The full Bayesian linear state space model will be useful in validating our modeling assumptions and understanding of the data.

6.2 Summary

Chapter 2 presented the application of the MVARX model in modeling intracerebral EEG triggered by current stimulation. We presented an outlier detection process to exclude outlying epochs from the data analysis. The model order p was selected with cross-validation and model checking was done with a residual whiteness test. The performance of the model in characterizing the data was evaluated with evoked response and one-step prediction of the test datasets. Moreover, we estimated models under two different structural hypotheses - full (integrated) and diagonal (segregated) - and demonstrated that the models are capable of characterizing the difference between level of integration in wakefulness and sleep. Finally an application of the MVARX model in computing the integrated information of the brain in wakefulness and sleep was presented.

Chapter 3 extended the results of Chapter 2 to consider learning MVARX model from intracerebral EEG with self-connected group lasso regularization. We compared sparse models learned in a data-driven way to full models introduced in Chapter 2. We demonstrated that sparse models generally outperform the full models in one-step prediction and evoked-response characterization. We also showed that the group lasso regularization identifies sparser brain connectivity patterns during sleep, compared to those during wakefulness. Moreover the regularization encourages denser interconnections between nodes in a common cortical area than those between different areas.

Chapter 4 built upon the results of Chapter 2 and [10] to model the scalp evoked responses to exogenous stimulation with a linear state space model in which the state equation characterizes the cortical activity excited by the exogenous stimulation as an MVARX process. We demonstrated the application of the model by modeling TMS-EEG measurements. We selected the ROIs based on cortical signal power and estimated the model parameters with an EM algorithm. An application of the model was shown to learn models according to two structural hypotheses - integrated and segregated - and we showed that the feedback path supported by the integrated model is necessary to characterize the evoked response in wakefulness.

Chapter 5 considered imposing priors over the parameters in the linear state space model discussed in Chapter 4 and presented variational Bayesian inference procedures for learning approximated posteriors of the parameters and cortical activities. We considered two classes of priors with differing structural preferences and demonstrated the learning procedures for both classes of priors. We also showed how the evidence lower bound can be evaluated. The lower bound can be utilized to determine convergence of the algorithm and to perform model selection. Finally, the priors were compared with simulation studies to show their performance in learning the parameters.

A EM ALGORITHM FOR TMS/EEG

Eq. (4.1) in the paper may be rewritten as the state equation:

$$\mathbf{z}_{n,j} = \mathbf{A}_s \mathbf{z}_{n-1,j} + \mathbf{B}_s \mathbf{u}_{n,j} + \tilde{\mathbf{w}}_{n,j} \tag{A.1}$$

where the Kp by one state vector $\mathbf{z}_{n,j}$ is the concatenation of present cortical signals and the previous $p-1$ cortical signals, $\mathbf{z}_{n,j} = [\mathbf{x}_{n,j}^\top, \mathbf{x}_{n-1,j}^\top, \cdots, \mathbf{x}_{n-p+1,j}^\top]^\top$, the ℓ by one input vector $\mathbf{u}_{n,j}$ is the concatenation of ℓ consecutive samples of $u_{n,j}$ in time, $\mathbf{u}_{n,j} = [u_{n,j}, u_{n-1,j}, \cdots, u_{n-\ell+1,j}]^\top$, and the Kp by one state noise $\tilde{\mathbf{w}}_{n,j} = [\mathbf{w}_{n,j}^\top, \mathbf{0}]^\top$. The Kp by Kp state transition matrix $\mathbf{A}_s$ is given as

$$\mathbf{A}_s = \begin{bmatrix} \mathbf{A} & \\ \mathbf{I} & \mathbf{0} \end{bmatrix}$$

and the Kp by ℓ matrix $\mathbf{B}_s$ is given as

$$\mathbf{B}_s = \begin{bmatrix} \mathbf{B} \\ \mathbf{0} \end{bmatrix}.$$

Similarly, Eq. (4.2) can be reexpressed as

$$\begin{aligned} \mathbf{y}_{n,j} &= \mathbf{H}\Lambda\mathbf{x}_{n,j} + \mathbf{v}_{n,j} \\ &= \mathbf{C}\mathbf{z}_{n,j} + \mathbf{v}_{n,j} \end{aligned}$$

where the M by 3K matrix $\mathbf{H} = [\mathbf{H}_1, \mathbf{H}_2, \cdots, \mathbf{H}_K]$, the 3K by K matrix $\boldsymbol{\Lambda}$ is given as

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \lambda_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \lambda_K \end{bmatrix},$$

and the M by Kp matrix $\mathbf{C} = [\mathbf{H}\Lambda, \mathbf{0}]$.

The maximum likelihood estimate (MLE) is found by solving

$$\max_\theta \log p(\mathbf{Y}|\mathbf{U}, \theta) = \max_\theta \int \log p(\mathbf{Y}, \mathbf{X}|\mathbf{U}, \theta) d\mathbf{X}.$$

The EM algorithm is an iterative method for finding the MLEs in the presence of latent variables $\mathbf{X}$. Each iteration of the EM algorithm involves two steps. Given $\theta^{(k)}$, the estimate at iteration k, we first complete the E step by computing

$$\mathcal{Q}(\theta, \theta^{(k)}) = \mathbb{E}_{\mathbf{X}|\mathbf{Y},\mathbf{U},\theta^{(k)}}[\log p(\mathbf{Y}, \mathbf{X}|\mathbf{U}, \theta)].$$

Next, we find $\theta^{(k+1)}$ in the M step as $\theta^{(k+1)} = \arg\max_\theta \mathcal{Q}(\theta, \theta^{(k)})$. The details of the E- and M-steps are as follows.

E-step In the E-step we evaluate $\mathcal{Q}(\theta, \theta^{(k)}) = \mathbb{E}_{\mathbf{X}|\mathbf{Y},\mathbf{U},\theta^{(k)}}[\log p(\mathbf{Y}, \mathbf{X}|\mathbf{U}, \theta)]$.

The complete data log likelihood function of the parameter θ, $\log p(\mathbf{Y}, \mathbf{X}|\mathbf{U}, \theta)$, is given as

$$\begin{aligned} &\log p(\mathbf{Y}, \mathbf{X}|\mathbf{U}, \theta) \\ = &\sum_{j=1}^{J} \left[\log p(\mathbf{z}_{0,j}|\pi_0, \sigma_0^2 \cdot \mathbf{I})) + \sum_{n=1}^{N} \log \left(p(\mathbf{y}_{n,j}|\mathbf{z}_{n,j}, \Lambda, \mathbf{R}) p(\mathbf{z}_{n,j}|\mathbf{z}_{n-1,j}, \mathbf{u}_{n,j}, \mathbf{A}_s, \mathbf{B}_s, \mathbf{Q}_s) \right) \right] \end{aligned}$$

where

$$\begin{aligned} p(\mathbf{z}_{0,j}|\pi_0, \sigma_0^2 \cdot \mathbf{I}) &= \mathcal{N}(\mathbf{z}_{0,j}; \pi_0, \sigma_0^2 \cdot \mathbf{I}) \\ p(\mathbf{y}_{n,j}|\mathbf{z}_{n,j}, \Lambda, \mathbf{R}) &= \mathcal{N}(\mathbf{y}_{n,j}; \mathbf{C}\mathbf{z}_{n,j}, \mathbf{R}) \\ p(\mathbf{z}_{n,j}|\mathbf{z}_{n-1,j}, \mathbf{u}_{n,j}, \mathbf{A}_s, \mathbf{B}_s, \mathbf{Q}_s) &= \mathcal{N}(\mathbf{z}_{n,j}; \mathbf{A}_s\mathbf{z}_{n-1,j} + \mathbf{B}_s\mathbf{u}_{n,j}, \mathbf{Q}_s). \end{aligned}$$

Here $\mathcal{N}(\mathbf{x}; \pi, \mathbf{V})$ denotes the probability density function of a Gaussian random vector $\mathbf{x}$ with mean π and covariance matrix $\mathbf{V}$. Hence the E-step involves evaluating the posterior expectations of the following sufficient statistics:

$$\begin{aligned} \mathbf{z}_{n|N,j} &= \mathbb{E}_{\mathbf{X}|\mathbf{Y},\mathbf{U},\theta^{(k)}}[\mathbf{z}_{n,j}] && \text{(A.2)} \\ \mathbf{P}_{n|N,j} &= \mathbb{E}_{\mathbf{X}|\mathbf{Y},\mathbf{U},\theta^{(k)}}[\mathbf{z}_{n,j}\mathbf{z}_{n,j}^\top] && \text{(A.3)} \\ \mathbf{P}_{n,n-1|N,j} &= \mathbb{E}_{\mathbf{X}|\mathbf{Y},\mathbf{U},\theta^{(k)}}[\mathbf{z}_{n,j}\mathbf{z}_{n-1,j}^\top], && \text{(A.4)} \end{aligned}$$

for $n = 1, 2, \cdots, N$, $j = 1, 2, \cdots, J$. These sufficient statistics can be computed with the fixed-interval smoother (also known as the Rauch-Tung-Striebel smoother) [129, 126]. The fixed interval smoother starts with Kalman filtering (forward pass), which is as follows.

- Initialize with $\mathbf{z}_{0|0,j} = \pi_0^{(k)}$ and $\mathbf{V}_{0|0} = \mathbf{V}_0^{(k)}$ where $\pi_0^{(k)}$ and $\sigma_0^{2,(k)} \cdot \mathbf{I}$ are respectively the estimates of initial mean and covariance matrix of each trial after the k-th iteration.

- Prediction

$$\begin{aligned}\mathbf{z}_{n|n-1,j} &= \mathbf{A}_s^{(k)}\mathbf{z}_{n-1|n-1,j} + \mathbf{B}_s^{(k)}\mathbf{u}_{n,j},\\ \mathbf{V}_{n|n-1} &= \mathbf{A}_s^{(k)}\mathbf{V}_{n-1|n-1}(\mathbf{A}_s^{(k)})^\top + \mathbf{Q}_s^{(k)},\end{aligned}$$

  where $\mathbf{A}_s^{(k)}$, $\mathbf{B}_s^{(k)}$, and $\mathbf{Q}_s^{(k)}$ are respectively the estimates of state transition matrix, exogenous input matrix, and state noise covariance matrix after the k-th iteration.

- Updating

$$\begin{aligned}\mathbf{K}_n &= \mathbf{V}_{n|n-1}(\mathbf{C}^{(k)})^\top(\mathbf{R}^{(k)} + \mathbf{C}^{(k)}\mathbf{V}_{n|n-1}(\mathbf{C}^{(k)})^\top)^{-1},\\ \mathbf{z}_{n|n,j} &= \mathbf{z}_{n|n-1,j} + \mathbf{K}_n(\mathbf{y}_{n,j} - \mathbf{C}^{(k)}\mathbf{z}_{n|n-1,j}),\\ \mathbf{V}_{n|n} &= (\mathbf{I} - \mathbf{K}_n\mathbf{C}^{(k)})\mathbf{V}_{n|n-1},\end{aligned}$$

  where $\mathbf{C}^{(k)} = [\mathbf{H}\Lambda^{(k)}, \mathbf{0}]$; the matrices $\Lambda^{(k)}$ and $\mathbf{R}^{(k)}$ are respectively the estimates of Λ and $\mathbf{R}$ after the k-th iteration.

Kalman filtering proceeds sequentially from $n = 0$ to N for all j. Note that the covariance matrices $\mathbf{V}_{n|n-1}$, $\mathbf{V}_{n|n}$ and the Kalman gain $\mathbf{K}_n$ are independent of the trial index j. Thus, the computational cost is reduced by implementing updates for these matrices in only one trial. The updates for the states $\mathbf{z}_{n|n-1,j}$ and $\mathbf{z}_{n|n,j}$ must be performed for all trials.

The log likelihood function at $\theta^{(k)}$ is expressed as [97]:

$$\begin{aligned}L(\theta^{(k)}) &= -\frac{1}{2}\sum_{j=1}^{J}\sum_{n=1}^{N}\log|\mathbf{V}_{n|n-1}|\\ &\quad +(\mathbf{y}_{n,j} - \mathbf{C}^{(k)}\mathbf{z}_{n|n-1,j})^\top(\mathbf{V}_{n|n-1})^{-1}(\mathbf{y}_{n,j} - \mathbf{C}^{(k)}\mathbf{z}_{n|n-1,j}).\end{aligned}$$

The EM algorithm is terminated at iteration k if $[L(\theta^{(k)}) - L(\theta^{(k-1)}]/[L(\theta^{(k)}) - L(\theta^{(0)})]$ is sufficiently small.

The fixed-interval smoother uses the results of the Kalman filter as initial value to obtain the sufficient statics in (A.2) - (A.4) by computing the following recursions backward from $n = N - 1$ all the way to $n = 0$.

- Backward recursion for $n = N - 1, N - 2, \cdots, 0$:

$$\begin{aligned}
\mathbf{J}_n &= \mathbf{V}_{n|n}(\mathbf{A}_s^{(k)})^\top(\mathbf{V}_{n+1|n})^{-1}, \\
\mathbf{z}_{n|N,j} &= \mathbf{z}_{n|n,j} + \mathbf{J}_n(\mathbf{z}_{n+1|N,j} - \mathbf{A}_s^{(k)}\mathbf{z}_{n|n,j} - \mathbf{B}_s^{(k)}\mathbf{u}_{n+1,j}), \\
\mathbf{V}_{n|N} &= \mathbf{V}_{n|n} + \mathbf{J}_n(\mathbf{V}_{n+1|N} - \mathbf{V}_{n+1|n})(\mathbf{J}_n)^\top, \\
\mathbf{V}_{n+1,n|N} &= \mathbf{V}_{n+1|N}(\mathbf{J}_n)^\top.
\end{aligned}$$

The smoothing steps on the covariance matrices $\mathbf{V}_{n|N}$, $\mathbf{V}_{n+1,n|N}$, and the matrix $\mathbf{J}_n$ are independent of trial index and thus only need to be computed for one trial.

The results of the fixed-interval smoother are used to form the conditional expectation of the sufficient statistics Eqs. (A.2)-(A.4). $\mathbf{z}_{n|N,j}$ is already computed, $\mathbf{P}_{n|N,j} = \mathbf{V}_{n|N} + \mathbf{z}_{n|N,j}(\mathbf{z}_{n|N,j})^\top$, and $\mathbf{P}_{n,n-1|N,j} = \mathbf{V}_{n,n-1|N} + \mathbf{z}_{n|N,j}(\mathbf{z}_{n-1|N,j})^\top$.

Square-root Kalman filtering and smoothing minimize numerical computation error propagation [130] and increase the stability of E-step.

M-step In the following we provide the M-step updates of the estimates for $\mathbf{A}, \mathbf{B}, \Lambda, \mathbf{R}, \mathbf{Q}, \pi_0$, and $\mathbf{V}_0$.

- Update for Λ:
  Define the matrices **F** and **G** as

$$\mathbf{F} = \sum_{j=1}^{J}\sum_{n=1}^{N}\mathbf{x}_{n|N,j}(\mathbf{x}_{n|N,j})^\top$$

  and

$$\mathbf{G} = \sum_{i=1}^{J}\sum_{n=1}^{N}\mathbf{y}_{n,j}(\mathbf{x}_{n|N,j})^\top$$

  where $\mathbf{x}_{n|N,j}$ is formed by the first K elements of $\mathbf{z}_{n|N,j}$.

The terms in $\mathcal{Q}(\theta, \theta^{(k)})$ that involve Λ are given as:

$$\begin{aligned} q(\Lambda) \triangleq & -\frac{1}{2}\mathrm{tr}\Big\{\mathbf{R}^{-1}\Big[\sum_{j=1}^{J}\sum_{n=1}^{N}\mathbf{y}_{n,j}(\mathbf{y}_{n,j})^\top - \mathbf{H}\Lambda\sum_{j=1}^{J}\sum_{n=1}^{N}\mathbf{x}_{n|N,j}(\mathbf{y}_{n,j})^\top \\ & -\sum_{i=1}^{J}\sum_{n=1}^{N}\mathbf{y}_{n,j}(\mathbf{x}_{n|N,j})^\top\Lambda^\top\mathbf{H}^\top + \mathbf{H}\Lambda\sum_{j=1}^{J}\sum_{n=1}^{N}\mathbf{x}_{n|N,j}(\mathbf{x}_{n|N,j})^\top\Lambda^\top\mathbf{H}^\top\Big]\Big\} \end{aligned}$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix. In the following we will substitute $\mathbf{R}^{(k)}$ for $\mathbf{R}$ in $q(\Lambda)$. To find $\Lambda^{(k+1)}$, we need to solve the following optimization problem

$$\begin{aligned} \min_{\Lambda} \quad & -q(\Lambda) \\ \text{subject to} \quad & \|\lambda_i\|_2^2 = 1, \quad i = 1, 2, \cdots, K. \end{aligned} \tag{A.5}$$

We solve this problem with the interior-point method [131]. We set the initial guess Λ_0 by finding an unconstrained analytical solution that minimizes $-q(\Lambda)$. Equating the derivative of $q(\Lambda)$ with respect to each of λ_i to zero and reformulating gives the system of linear equation as

$$\begin{aligned} & \begin{bmatrix} \mathbf{H}_1^\top(\mathbf{R}^{(k)})^{-1}\mathbf{H}_1 f_{1,1} & \cdots & \mathbf{H}_1^\top(\mathbf{R}^{(k)})^{-1}\mathbf{H}_K f_{K,1} \\ \mathbf{H}_2^\top(\mathbf{R}^{(k)})^{-1}\mathbf{H}_1 f_{1,2} & \cdots & \mathbf{H}_2^\top(\mathbf{R}^{(k)})^{-1}\mathbf{H}_K f_{K,2} \\ \vdots & \ddots & \vdots \\ \mathbf{H}_K^\top(\mathbf{R}^{(k)})^{-1}\mathbf{H}_1 f_{1,K} & \cdots & \mathbf{H}_K^\top(\mathbf{R}^{(k)})^{-1}\mathbf{H}_K f_{K,K} \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_K \end{bmatrix} \\ = & \begin{bmatrix} \mathbf{H}_1^\top(\mathbf{R}^{(k)})\mathbf{g}^1 \\ \mathbf{H}_2^\top(\mathbf{R}^{(k)})\mathbf{g}^2 \\ \vdots \\ \mathbf{H}_K^\top(\mathbf{R}^{(k)})\mathbf{g}^K \end{bmatrix} \end{aligned} \tag{A.6}$$

where $f_{i,j}$ denotes the (i, j)-th element of $\mathbf{F}$ and $\mathbf{g}^i$ denotes the i-th column of $\mathbf{G}$. Each λ_i obtained by solving Eq. (A.6) is normalized to obtain Λ_0. The result of solving the constrained optimization problem (A.5) is $\Lambda^{(k+1)}$.

- Update for $\mathbf{R}$:

$$
\begin{aligned}
\mathbf{R}^{(k+1)} &= \sigma_R^2 \cdot \mathbf{I} \\
\sigma_R^2 &= \frac{1}{JNM} \cdot \mathrm{tr}\Bigg[\sum_{j=1}^{J} \sum_{n=1}^{N} \mathbf{y}_{n,j}(\mathbf{y}_{n,j})^\top - \mathbf{H}\Lambda^{(k+1)}\mathbf{G}^\top - \mathbf{G}(\Lambda^{(k+1)})^\top \mathbf{H}^\top \\
&\quad + \mathbf{H}\Lambda^{(k+1)}\mathbf{F}(\Lambda^{(k+1)})^\top \mathbf{H}^\top \Bigg].
\end{aligned}
$$

Note that we assume isotropic observation noise in our study. The update is easily modified for $\mathbf{R}$ an arbitrary covariance matrix.

- Update for $\mathbf{A}$ and $\mathbf{B}$:

Define the $Kp + \ell$ by $Kp + \ell$ matrix $\mathbf{K}$ and the K by Kp matrix $\mathbf{L}$ as

$$
\mathbf{K} = \begin{bmatrix} \sum_{j=1}^{J} \sum_{n=1}^{N} \mathbf{P}_{n|N,j} & \sum_{j=1}^{J} \sum_{n=1}^{N} \mathbf{z}_{n|N,j}\mathbf{u}_{n,j}{}^\top \\ \sum_{j=1}^{J} \sum_{n=1}^{N} \mathbf{u}_{n,j}\mathbf{z}_{n|N,j}^\top & \sum_{j=1}^{J} \sum_{n=1}^{N} \mathbf{u}_{n,j}\mathbf{u}_{n,j}^\top \end{bmatrix}
$$

$$
\mathbf{L} = [\mathbf{I}_K, \mathbf{0}] \left[\sum_{j=1}^{J} \sum_{n=1}^{N} \mathbf{P}_{n,n-1|N,j} \right].
$$

The derivative of $\mathcal{Q}(\theta, \theta^{(k)})$ with respect to $[\mathbf{A},\ \mathbf{B}]$ is,

$$
\frac{\partial \mathcal{Q}(\theta, \theta^{(k)})}{\partial [\mathbf{A},\ \mathbf{B}]} = \mathbf{Q}^{-1}\left(\mathbf{L} + [\mathbf{A}, \mathbf{B}]\mathbf{K}\right).
$$

Equating this derivative to zero gives

$$
[\mathbf{A}^{(k+1)},\ \mathbf{B}^{(k+1)}] = \mathbf{L}\mathbf{K}^{-1}.
$$

- Update for $\mathbf{Q}$:

$$
\mathbf{Q}^{(k+1)} = \frac{1}{JN}\left(\mathbf{F} - \mathbf{L}\mathbf{K}^{-1}\mathbf{L}^\top\right)
$$

- Update for π_0:

$$\pi_0^{(k+1)} = \frac{1}{J}\sum_{j=1}^{J} \mathbf{z}_{0|N,j}$$

- Update of σ_0^2:

$$\sigma_0^{2,(k+1)} = \frac{1}{JKp}\sum_{j=1}^{J} \operatorname{tr}\Big[\mathbf{V}_{0|N,j} + (\mathbf{z}_{0|N,j} - \pi_0^{(k+1)})(\mathbf{z}_{0|N,j} - \pi_0^{(k+1)})^\top\Big].$$

Note that this form of update ensures numerical stability by computing the estimate as a sum of positive semidefinite matrices.

B REGION SELECTION PROCEDURE

We select ROIs for inclusion in the model by first using a minimum norm method to reconstruct source activity at 3004 dipoles tesselating the cortical surface. Next the estimated source activity is used to estimate the combined normalized power in each cortical patch across both wakefulness and sleep data sets for each subject. The ROIs in the model are selected based on the combined normalized power across the entire patch set.

Consider the M by P leadfield matrix $\mathbf{L}$ where $P = 3004$ is the number of dipoles. Let the dipole activity originating from voxel β be represented by the signal $r^{\beta}_{n,j}$, and the concatenation of the dipole signals be the P by 1 vector $\mathbf{r}_{n,j} = [r^1_{n,j}, r^2_{n,j}, \cdots, r^P_{n,j}]^\top$. The measured data can be expressed as

$$\mathbf{y}_{n,j} = \mathbf{L}\mathbf{r}_{n,j} + \mathbf{n}_{n,j}$$

where $\mathbf{n}_{n,j}$ is noise. The minimum norm solution for $\mathbf{r}_{n,j}$ satisfies

$$\min_{\mathbf{r}_{n,j}} \sum_{j=1}^{J} \sum_{n=1}^{N} ||\mathbf{y}_{n,j} - \mathbf{L}\mathbf{r}_{n,j}||_2^2 + \eta M ||\mathbf{r}_{n,j}||_2^2.$$

We choose η using the generalized cross-validation (GCV) method proposed in [132]. The GCV objective is given as

$$\hat{\eta} = \arg\min_{\eta} \frac{\frac{1}{M}\sum_{j=1}^{J}\sum_{n=1}^{N} ||(\mathbf{I} - \mathcal{A}(\eta))\mathbf{y}_{n,j}||_2^2}{[\frac{1}{M}\mathrm{tr}(\mathbf{I} - \mathcal{A}(\eta))]^2}$$

where $\mathcal{A}(\eta) = \mathbf{L}(\mathbf{L}^\top\mathbf{L} + \eta M \cdot \mathbf{I})^{-1}\mathbf{L}^\top$. Once $\hat{\eta}$ is selected, the signals from all dipoles are reconstructed as [133]

$$\mathbf{r}_{n,j} = \mathbf{L}^\top(\mathbf{L}\mathbf{L}^\top + (\hat{\eta}M)\mathbf{I})^{-1}\mathbf{y}_{n,j}.$$

Define $\mathbf{L}^\ell$ as the M by G^ℓ matrix containing the G^ℓ columns of $\mathbf{L}$ corresponding to the G^ℓ dipoles within patch ℓ and let $\mathbf{L}^\ell = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$ be the singular value decomposition. Define the M by 3 patch basis [95] $\mathbf{D}^\ell = \tilde{\mathbf{U}}^\ell\tilde{\boldsymbol{\Sigma}}^\ell$ where the M by 3 matrix $\tilde{\mathbf{U}}^\ell$ contains the three dominant left singular vectors and $\tilde{\boldsymbol{\Sigma}}^\ell$ is a 3 by 3 diagonal matrix of the three largest singular values. Assuming $\mathbf{L}^\ell$ is well approximated by a rank 3 matrix [95], then the Gram matrix $\mathbf{L}^\ell(\mathbf{L}^\ell)^\top$ can be approximated as $\mathbf{D}^\ell(\mathbf{D}^\ell)^\top$. This result will be used below.

Next we use the reconstructed dipole activity to estimate the cortical signal power associated with the ℓ-th patch. The signals from dipoles within patch ℓ are $\mathbf{r}^\ell_{n,j} = (\mathbf{L}^\ell)^\top(\mathbf{L}\mathbf{L}^\top + (\hat{\eta}M)\mathbf{I})^{-1}\mathbf{y}_{n,j}$. We model the cortical signal from each patch as a scalar $s^\ell_{n,j}$ with the orientation of the signal with respect to the columns of $\mathbf{D}^\ell$ given by λ^ℓ. That is, the measured signal components originating from patch ℓ are modeled as $\mathbf{y}^\ell_{n,j} \approx \mathbf{D}^\ell\lambda^\ell s^\ell_{n,j}$. Ignoring noise, we have $\mathbf{y}^\ell_{n,j} = \mathbf{L}^\ell\mathbf{r}^\ell_{n,j}$ and thus

$$\begin{aligned} s^\ell_{n,j} &= (\lambda^\ell)^\top(\tilde{\mathbf{\Sigma}}^\ell)^{-1}(\tilde{\mathbf{U}}^\ell)^\top\mathbf{y}^\ell_{n,j} \\ &= (\lambda^\ell)^\top(\tilde{\mathbf{\Sigma}}^\ell)^{-1}(\tilde{\mathbf{U}}^\ell)^\top\mathbf{L}^\ell(\mathbf{L}^\ell)^\top(\mathbf{L}\mathbf{L}^\top + (\hat{\eta}M)\mathbf{I})^{-1}\mathbf{y}_{n,j} \\ &= (\lambda^\ell)^\top(\mathbf{D}^\ell)^\top(\mathbf{L}\mathbf{L}^\top + (\hat{\eta}M)\mathbf{I})^{-1}\mathbf{y}_{n,j} \end{aligned}$$

where we used $\mathbf{L}^\ell(\mathbf{L}^\ell)^\top \approx \mathbf{D}^\ell(\mathbf{D}^\ell)^\top = \tilde{\mathbf{U}}^\ell(\tilde{\mathbf{\Sigma}}^\ell)^2(\tilde{\mathbf{U}}^\ell)^\top$ and $(\tilde{\mathbf{\Sigma}}^\ell)^{-1}(\tilde{\mathbf{U}}^\ell)^\top\tilde{\mathbf{U}}^\ell(\tilde{\mathbf{\Sigma}}^\ell)^2(\tilde{\mathbf{U}}^\ell)^\top = \tilde{\mathbf{\Sigma}}^\ell(\tilde{\mathbf{U}}^\ell)^\top = (\mathbf{D}^\ell)^\top$. We choose the orientation λ^ℓ to maximize the power associated with $s^\ell_{n,j}$, that is

$$\lambda^\ell = \arg\max_{\lambda^\ell} (\lambda^\ell)^\top(\mathbf{D}^\ell)^\top(\mathbf{L}\mathbf{L}^\top + (\hat{\eta}M)\mathbf{I})^{-1}\mathbf{R}_\mathbf{y}(\mathbf{L}\mathbf{L}^\top + (\hat{\eta}M)\mathbf{I})^{-1}\mathbf{D}^\ell\lambda^\ell$$

subject to $(\lambda^\ell)^\top\lambda^\ell = 1$ as in [133]. Here the spatial covariance matrix $\mathbf{R}_\mathbf{y} = 1/JN \cdot \sum_{j=1}^{J}\sum_{n=1}^{N}\mathbf{y}_{n,j}(\mathbf{y}_{n,j})^\top - \mu_\mathbf{y}\mu_\mathbf{y}^\top$ and $\mu_\mathbf{y} = 1/JN \cdot \sum_{j=1}^{J}\sum_{n=1}^{N}\mathbf{y}_{n,j}$. The solution is to choose λ^ℓ as the eigenvector corresponding to the maximum eigenvalue of $(\mathbf{D}^\ell)^\top(\mathbf{L}\mathbf{L}^\top + (\hat{\eta}M)\mathbf{I})^{-1}\mathbf{R}_\mathbf{y}(\mathbf{L}\mathbf{L}^\top + (\hat{\eta}M)\mathbf{I})^{-1}\mathbf{D}^\ell$. The power $\sigma(\ell)$ associated with $s^\ell_{n,j}$ is equal to the maximum eigenvalue.

The power associated with the ℓth patch of the data collected from a subject during wakefulness and sleep is denoted $\sigma_w(\ell)$ and $\sigma_s(\ell)$, respectively. We form a weighted average of wakefulness and sleep patch powers to obtain a normalized index

$$\overline{\sigma}(\ell) = \frac{\sigma_w(\ell)}{\mathrm{tr}(\mathbf{R}_{\mathbf{y},w})} + \frac{\sigma_s(\ell)}{\mathrm{tr}(\mathbf{R}_{\mathbf{y},s})}$$

where the normalizing weights $\mathrm{tr}(\mathbf{R}_{\mathbf{y},w})$ and $\mathrm{tr}(\mathbf{R}_{\mathbf{y},s})$ are overall signal power of the data in wakefulness and sleep, respectively.

The first ROI is selected as the one that maximizes $\overline{\sigma}(\ell)$. The second ROI maximizes $\overline{\sigma}(\ell)$ after all ROIs that overlap with the first one are excluded. The third patch maximizes $\overline{\sigma}(\ell)$ after all patches that overlap with the first two are excluded. This procedure continues until $\lfloor M/3 \rfloor$ patches are selected.

C.1 Evaluation of $\log Z'$

We can reexpress the $\log Z'$

$$
\begin{aligned}
& \log Z' \\
= \; & \sum_{j=1}^{J} \log \int d\mathbf{z}_{0,j} d\mathbf{x}_{1:N,j} \exp\left(\mathrm{E}_{q(\theta)}\left[\log p(\mathbf{Y}, \mathbf{Z}|\theta) \right] \right) \\
= \; & \sum_{j=1}^{J} \log \int d\mathbf{z}_{0,j} d\mathbf{x}_{1:N,j} \exp\Bigg\{ -\frac{Kp}{2} \log 2\pi + \frac{1}{2} \mathrm{E}_{q(\rho_s)}\left[\log |\Sigma_0^{-1}| \right] \\
& -\frac{1}{2} (\mathbf{z}_{0,j} - \boldsymbol{\pi}_0)^\top \mathrm{E}_{q(\rho_s)}\left[\Sigma_0^{-1} \right] (\mathbf{z}_{0,j} - \boldsymbol{\pi}_0) \\
& + \sum_{n=1}^{N} \Bigg[-\frac{K}{2} \log 2\pi + \frac{K}{2} \mathrm{E}_{q(\rho_q)}\left[\log \rho_q \right] \\
& -\frac{1}{2} \mathrm{E}_{q(\rho_q, \Xi)}\left[\rho_q \cdot (\mathbf{x}_{n,j} - \mathbf{A}\mathbf{z}_{n-1,j} - \mathbf{B}\mathbf{u}_{n,j})^\top (\mathbf{x}_{n,j} - \mathbf{A}\mathbf{z}_{n-1,j} - \mathbf{B}\mathbf{u}_{n,j}) \right] \\
& -\frac{M}{2} \log 2\pi + \frac{M}{2} \mathrm{E}_{q(\rho_r)}\left[\log \rho_r \right] \\
& -\frac{\mathrm{E}_{q(\rho_r)}\left[\rho_r \right]}{2} (\mathbf{y}_{n,j} - \mathbf{C}\mathbf{z}_{n,j})^\top (\mathbf{y}_{n,j} - \mathbf{C}\mathbf{z}_{n,j})] \Bigg] \Bigg\} \\
= \; & \sum_{j=1}^{J} \log \int d\mathbf{z}_{0,j} d\mathbf{x}_{1:N,j} \exp\Bigg\{ -\frac{Kp}{2} \log 2\pi + \frac{1}{2} \log |\tilde{\boldsymbol{\Sigma}}_0^{-1}| - \frac{1}{2} (\mathbf{z}_{0,j} - \tilde{\boldsymbol{\pi}}_0)^\top \tilde{\boldsymbol{\Sigma}}_0^{-1} (\mathbf{z}_{0,j} - \tilde{\boldsymbol{\pi}}_0) \\
& + \frac{1}{2} \mathrm{E}_{q(\rho_s)}\left[\log |\Sigma_0^{-1}| \right] - \frac{1}{2} \log |\tilde{\boldsymbol{\Sigma}}_0^{-1}| + \frac{1}{2} \tilde{\boldsymbol{\pi}}_0^\top \tilde{\boldsymbol{\Sigma}}_0^{-1} \tilde{\boldsymbol{\pi}}_0 - \frac{1}{2} \boldsymbol{\pi}_0^\top \mathrm{E}_{q(\rho_s)}\left[\Sigma_0^{-1} \right] \boldsymbol{\pi}_0 - \frac{1}{2} \mathbf{u}_{1,j}^\top \Sigma_B \mathbf{u}_{1,j} \\
& + \sum_{n=1}^{N-1} \Bigg[-\frac{K}{2} \log 2\pi + \frac{K}{2} \log \mathrm{E}_{q(\rho_q)}\left[\rho_q \right] - \frac{\mathrm{E}_{q(\rho_q)}\left[\rho_q \right]}{2} (\mathbf{x}_{n,j} - \mathrm{E}[\mathbf{A}]\mathbf{z}_{n-1,j} - \mathrm{E}[\mathbf{B}]\mathbf{u}_{n,j})^\top \\
& \quad \cdot (\mathbf{x}_{n,j} - \mathrm{E}[\mathbf{A}]\mathbf{z}_{n-1,j} - \mathrm{E}[\mathbf{B}]\mathbf{u}_{n,j}) + \frac{K}{2} \mathrm{E}_{q(\rho_q)}\left[\log \rho_q \right] - \frac{K}{2} \log \mathrm{E}_{q(\rho_q)}\left[\rho_q \right] \\
& -\frac{M + Kp + \ell}{2} \log 2\pi + \frac{1}{2} \log |\tilde{\boldsymbol{\Lambda}}_{\mathbf{R}}| - \frac{1}{2} (\tilde{\mathbf{y}}_{n,j} - \tilde{\mathbf{C}}\mathbf{z}_{n,j})^\top \tilde{\boldsymbol{\Lambda}}_{\mathbf{R}}^{-1} \\
& \quad \cdot (\tilde{\mathbf{y}}_{n,j} - \tilde{\mathbf{C}}\mathbf{z}_{n,j}) + \frac{Kp + \ell}{2} \log 2\pi + \frac{M}{2} \mathrm{E}\left[\log \rho_r \right] - \frac{1}{2} \log |\tilde{\boldsymbol{\Lambda}}_{\mathbf{R}}| \Bigg] \\
& -\frac{K}{2} \log 2\pi + \frac{K}{2} \log \mathrm{E}_{q(\rho_q)}\left[\rho_q \right] - \frac{\mathrm{E}_{q(\rho_q)}\left[\rho_q \right]}{2} (\mathbf{x}_{N,j} - \mathrm{E}[\mathbf{A}]\mathbf{z}_{N-1,j} - \mathrm{E}[\mathbf{B}]\mathbf{u}_{N,j})^\top
\end{aligned}
$$

$$
\begin{aligned}
&\cdot(\mathbf{x}_{N,j} - E[\mathbf{A}]\mathbf{z}_{N-1,j} - E[\mathbf{B}]\mathbf{u}_{N,j}) + \frac{K}{2}E_{q(\rho_q)}\big[\log\rho_q\big] - \frac{K}{2}\log E_{q(\rho_q)}\big[\rho_q\big] \\
&-\frac{M}{2}\log 2\pi + \frac{M}{2}\log E[\rho_r] - \frac{E[\rho_r]}{2}(\mathbf{y}_{N,j} - \mathbf{C}\mathbf{z}_{N,j})^\top(\mathbf{y}_{N,j} - \mathbf{C}\mathbf{z}_{N,j}) \\
&\left.+\frac{M}{2}E\big[\log\rho_r\big] - \frac{M}{2}\log E[\rho_r]\right\} \\
= &\sum_{j=1}^{J}\left\{\frac{1}{2}E_{q(\rho_s)}\big[\log|\Sigma_0^{-1}|\big] - \frac{1}{2}\log|\tilde{\mathbf{\Sigma}}_0^{-1}| + \frac{1}{2}\tilde{\boldsymbol{\pi}}_0^\top\tilde{\mathbf{\Sigma}}_0^{-1}\tilde{\boldsymbol{\pi}}_0 - \frac{1}{2}\boldsymbol{\pi}_0^\top E_{q(\rho_s)}\big[\Sigma_0^{-1}\big]\boldsymbol{\pi}_0\right. \\
&-\frac{1}{2}\mathbf{u}_{1,j}^\top\Sigma_B\mathbf{u}_{1,j} + N\cdot\left(\frac{K}{2}E_{q(\rho_q)}\big[\log\rho_q\big] - \frac{K}{2}\log E_{q(\rho_q)}\big[\rho_q\big]\right. \\
&\quad\left.+\frac{M}{2}E\big[\log\rho_r\big] - \frac{M}{2}\log E[\rho_r]\right) + (N-1)\cdot\frac{Kp+\ell}{2}\log 2\pi \\
&+\log\int d\mathbf{z}_{0,j}\,d\mathbf{x}_{1:N,j}\cdot\left[\mathcal{N}(\mathbf{z}_{0,j};\tilde{\boldsymbol{\pi}}_0,\tilde{\mathbf{\Sigma}}_0)\prod_{n=1}^{N-1}\big[\mathcal{N}(\mathbf{x}_{n,j};E[\mathbf{A}]\mathbf{z}_{n-1,j} - E[\mathbf{B}]\mathbf{u}_{n,j}, E\big[\rho_q\big]^{-1}\cdot\mathbf{I})\right. \\
&\quad\cdot\mathcal{N}(\tilde{\mathbf{y}}_{n,j},\tilde{\mathbf{C}}\mathbf{z}_{n,j},\tilde{\mathbf{\Lambda}}_{\mathbf{R}}^{-1})\big] \\
&\left.\left.\cdot\mathcal{N}(\mathbf{x}_{N,j};E[\mathbf{A}]\mathbf{z}_{N-1,j} - E[\mathbf{B}]\mathbf{u}_{N,j}, E\big[\rho_q\big]^{-1}\cdot\mathbf{I})\mathcal{N}(\mathbf{y}_{N,j},\mathbf{C}\mathbf{z}_{N,j},E\big[\rho_r\big]^{-1}\cdot\mathbf{I})\right]\right\}.
\end{aligned}
$$

Thus we can let the last term as $\log f(\tilde{\mathbf{y}}_{1,j},\tilde{\mathbf{y}}_{2,j},\cdots,\tilde{\mathbf{y}}_{N-1,j},\mathbf{y}_{N,j})$, which we can further expand as $\log f(\tilde{\mathbf{y}}_{1,j},\tilde{\mathbf{y}}_{2,j},\cdots,\tilde{\mathbf{y}}_{N-1,j},\mathbf{y}_{N,j}) = \log f(\tilde{\mathbf{y}}_{1,j}) + \sum_{n=2}^{N-1}\log f(\tilde{\mathbf{y}}_{n,j}|\tilde{\mathbf{y}}_{1:n-1,j}) + \log f(\mathbf{y}_{N,j}|\tilde{\mathbf{y}}_{1:N-1,j})$. Hence we conclude that the $\ln Z'$ can be evaluated with the Kalman filter procedure similar to how we evaluate log likelihood function for the frequentist EM algorithm:

$$
\begin{aligned}
\log Z' = &\sum_{j=1}^{J}\left\{\frac{1}{2}E_{q(\rho_s)}\big[\log|\Sigma_0^{-1}|\big] - \frac{1}{2}\log|\tilde{\mathbf{\Sigma}}_0^{-1}| + \frac{1}{2}\tilde{\boldsymbol{\pi}}_0^\top\tilde{\mathbf{\Sigma}}_0^{-1}\tilde{\boldsymbol{\pi}}_0 - \frac{1}{2}\boldsymbol{\pi}_0^\top E_{q(\rho_s)}\big[\Sigma_0^{-1}\big]\boldsymbol{\pi}_0\right. \\
&-\frac{1}{2}\mathbf{u}_{1,j}^\top\Sigma_B\mathbf{u}_{1,j} + N\cdot\left(\frac{K}{2}E_{q(\rho_q)}\big[\log\rho_q\big] - \frac{K}{2}\log E_{q(\rho_q)}\big[\rho_q\big]\right. \\
&\quad\left.+\frac{M}{2}E\big[\log\rho_r\big] - \frac{M}{2}\log E[\rho_r]\right) + (N-1)\cdot\frac{Kp+\ell}{2}\log 2\pi \\
&-(N-1)\cdot\frac{M+Kp+\ell}{2}\log 2\pi - \frac{1}{2}\sum_{n=1}^{N-1}\big[\log|\mathbf{E}_{n|n-1}| + \mathbf{e}_{n|n-1,j}^\top\mathbf{E}_{n|n-1}^{-1}\mathbf{e}_{n|n-1,j}\big] \\
&\left.-\frac{M}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{E}_{N|N-1}| + \mathbf{e}_{N|N-1,j}^\top\mathbf{E}_{N|N-1}^{-1}\mathbf{e}_{N|N-1,j}\right\}
\end{aligned}
$$

where $\mathbf{e}_{n|n-1,j} = \tilde{\mathbf{y}}_{n,j} - \tilde{\mathbf{C}}\mathbf{z}_{n|n-1,j}$, $\mathbf{E}_{n|n-1} = \tilde{\mathbf{C}}\mathbf{V}_{n|n-1}\tilde{\mathbf{C}}^{\top} + \tilde{\mathbf{\Lambda}}_{\mathbf{R}}^{-1}$ for $n = 1, 2, \cdots, N-1$, $\mathbf{e}_{N|N-1,j} = \mathbf{y}_{N,j} - \mathbf{C}\mathbf{z}_{N|N-1,j}$ and $\mathbf{E}_{N|N-1}\mathbf{C}\mathbf{V}_{N|N-1}\mathbf{C}^{\top} + E[\rho_r]^{-1} \cdot \mathbf{I}$.

C.2 Evaluation KL$(q(\theta)||p(\theta))$ in the Evidence Lower Bound

Matrix Normal Prior The KL divergence between the approximate posterior $q(\theta)$ and the prior $p(\theta)$ can be written out as

$$\begin{aligned}
\mathrm{KL}(q(\theta)||p(\theta)) &= -H(q(\theta)) - E\big[\log p(\theta)\big] \\
&= -H(q(\rho_q)) - E_{q(\rho_q)}\big[H(q(\Xi|\rho_q)\big] - H(q(\mathbf{K}_\kappa)) \\
&\quad\; -H(q(\rho_r)) - H(q(\rho_{\mathbf{s}})) \\
&\quad -E_{q(\rho_q)}\big[\log p(\rho_q)\big] - E_{q(\rho_q,\Xi)q(\mathbf{K}_\kappa)}\big[\log p(\Xi|\rho_q, \mathbf{K}_\kappa)\big] \\
&\quad\; -E_{q(\mathbf{K}_\kappa)}\big[\log p(\mathbf{K}_\kappa)\big] - E_{q(\rho_r)}\big[\log p(\rho_r)\big] - E_{q(\rho_{\mathbf{s}})}\big[\log p(\rho_{\mathbf{s}})\big]
\end{aligned}$$

where

$$\begin{aligned}
H(q(\rho_q)) &= \log\Gamma(\hat{\alpha}_q) - \log\hat{\beta}_q + (1-\hat{\alpha}_q)\psi(\hat{\alpha}_q) + \hat{\alpha}_q \\
E_{q(\rho_q)}\big[H(q(\Xi|\rho_q)\big] &= \frac{K(Kp+\ell)}{2}\log 2\pi e - \frac{K(Kp+\ell)}{2}\big(\psi(\hat{\alpha}_q) - \log\hat{\beta}_q\big) - \frac{K}{2}\log|\hat{\mathbf{K}}| \\
H(q(\mathbf{K}_\kappa)) &= \log\Gamma_{Kp+\ell}\big(\frac{\hat{\nu}_\kappa}{2}\big) + \frac{(Kp+\ell)(Kp+\ell+1)}{2}\log 2 + \frac{Kp+\ell+1}{2}\log|\hat{\mathbf{W}}_\kappa| \\
&\quad -\frac{\hat{\nu}_k - (Kp+\ell) - 1}{2}\psi_{Kp+\ell}\big(\frac{\hat{\nu}_k}{2}\big) + \frac{\hat{\nu}_k(Kp+\ell)}{2} \\
H(q(\rho_r)) &= \log\Gamma(\hat{\alpha}_r) - \log\hat{\beta}_r + (1-\hat{\alpha}_r)\psi(\hat{\alpha}_r) + \hat{\alpha}_r \\
H(q(\rho_{\mathbf{s}})) &= \sum_{i=1}^{Kp}\big[\log\Gamma(\hat{\alpha}_{s,i}) - \log\hat{\beta}_{s,i} + (1-\hat{\alpha}_{s,i})\psi(\hat{\alpha}_{s,i}) + \hat{\alpha}_{s,i}\big] \\
E_{q(\rho_q)}\big[\log p(\rho_q)\big] &= -\log\Gamma(\alpha_q) + \alpha_q\log\beta_q + (\alpha_q - 1)(\psi(\hat{\alpha}_q) - \log\hat{\beta}_q) - \frac{\beta_q\hat{\alpha}_q}{\hat{\beta}_q}
\end{aligned}$$

$$\begin{aligned}
E\big[\log p(\boldsymbol{\Xi}|\rho_q,\mathbf{K}_\kappa)\big] &= -\frac{K(Kp+\ell)}{2}\log 2\pi + \frac{K(Kp+\ell)}{2}(\psi(\hat{\alpha}_q)-\log\hat{\beta}_q) \\
&\quad +\frac{K}{2}\big(\psi_{Kp+\ell}(\frac{\hat{\nu}_\kappa}{2})+(Kp+\ell)\log 2+\log|\hat{\mathbf{W}}_\kappa|\big) \\
&\quad -\frac{\hat{\nu}_k}{2}\mathrm{tr}\Big[\big(K\cdot\hat{\mathbf{K}}^{-1}+\frac{\hat{\alpha}_q}{\hat{\beta}_q}\hat{\boldsymbol{\Xi}}_\mu^\top\hat{\boldsymbol{\Xi}}_\mu\big)\hat{\mathbf{W}}_\kappa\Big] \\
E_{q(\mathbf{K}_\kappa)}\big[\log p(\mathbf{K}_\kappa)\big] &= -\log\Gamma_{Kp+\ell}(\frac{\nu_\kappa}{2})-\frac{\nu_k}{2}\log|\mathbf{W}_k|-\frac{\nu_k(Kp+\ell)}{2}\log 2 \\
&\quad +\frac{\nu_k-(Kp+\ell)-1}{2}\big(\psi_{Kp+\ell}(\frac{\hat{\nu}_k}{2})+(Kp+\ell)\log 2+\log|\hat{\mathbf{W}}_k|\big) \\
&\quad -\frac{\hat{\nu}_k}{2}\mathrm{tr}(\mathbf{W}_k^{-1}\hat{\mathbf{W}}_k) \\
E_{q(\rho_r)}\big[\log p(\rho_r)\big] &= -\log\Gamma(\alpha_r)+\alpha_r\log\beta_r+(\alpha_r-1)(\psi(\hat{\alpha}_r)-\log\hat{\beta}_r)-\frac{\beta_r\hat{\alpha}_r}{\hat{\beta}_r} \\
E_{q(\rho_s)}\big[\log p(\rho_{\mathbf{s}})\big] &= Kp(-\log\Gamma(\alpha_s)+\alpha_s\log\beta_s) \\
&\quad +\sum_{i=1}^{Kp}(\alpha_s-1)(\psi(\hat{\alpha}_{s,i})-\log\hat{\beta}_{s,i})-\frac{\beta_s\hat{\alpha}_{s,i}}{\hat{\beta}_{s,i}},
\end{aligned}$$

$\psi(x) = d\log\Gamma(x)/dx$ is the digamma function and $\psi_p(x) = d\log\Gamma_p(x)/dx$ is the multivariate digamma function.

Independent ARD prior The KL divergence from the approximate posterior $q(\theta)$ to the prior $p(\theta)$ can be shown to be

$$\begin{aligned}
\mathrm{KL}(q(\theta)||p(\theta)) &= -H(q(\theta))-E_{q(\theta)}\big[\log p(\theta)\big] \\
&= -H(q(\rho_q))-E_{q(\rho_q)}\big[H(q(\xi|\rho_q))\big]-H(q(\gamma_{\mathbf{A}})) \\
&\quad -H(q(\gamma_{\mathbf{B}}))-H(q(\rho_r))-H(q(\rho_{\mathbf{s}})) \\
&\quad -E_{q(\rho_q)}\big[\log p(\rho_q)\big]-E_{q(\rho_q,\xi)q(\gamma_{\mathbf{A}})q(\gamma_{\mathbf{B}})}\big[\log p(\xi|\rho_q,\gamma_{\mathbf{A}},\gamma_{\mathbf{B}})\big] \\
&\quad -E_{q(\gamma_{\mathbf{A}})}\big[\log p(\gamma_{\mathbf{A}})\big]-E_{q(\gamma_{\mathbf{B}})}\big[\log p(\gamma_{\mathbf{B}})\big] \\
&\quad -E_{q(\rho_r)}\big[\log p(\rho_r)\big]-E_{q(\rho_s)}\big[\log p(\rho_{\mathbf{s}})\big]
\end{aligned}$$

where $H(q(\rho_q))$, $H(q(\rho_r))$, $H(q(\rho_\mathbf{s}))$, $E_{q(\rho_q)}\big[\log p(\rho_q)\big]$, $E_{q(\rho_r)}\big[\log p(\rho_r)\big]$, and $E_{q(\rho_\mathbf{s})}\big[\log p(\rho_\mathbf{s})\big]$ are the same as those in the matrix normal prior, and

$$\begin{aligned}
E_{q(\rho_q)}\big[H(q(\xi|\rho_q))\big] &= \frac{K(Kp+\ell)}{2}\log 2\pi e - \frac{K(Kp+\ell)}{2}\big(\psi(\hat{\alpha}_q)-\log\hat{\beta}_q\big) - \frac{1}{2}\log|\hat{\boldsymbol{\Lambda}}_\xi| \\
H(q(\gamma_\mathbf{A})) &= \sum_{i=1}^{K}\sum_{j=1}^{Kp}\log\Gamma(\hat{\alpha}_{a,i,j}) - \log\hat{\beta}_{a,i,j} + (1-\hat{\alpha}_{a,i,j})\psi(\hat{\alpha}_{a,i,j}) + \hat{\alpha}_{a,i,j} \\
H(q(\gamma_\mathbf{B})) &= \sum_{i=1}^{K}\sum_{j=1}^{\ell}\log\Gamma(\hat{\alpha}_{b,i,j}) - \log\hat{\beta}_{b,i,j} + (1-\hat{\alpha}_{b,i,j})\psi(\hat{\alpha}_{b,i,j}) + \hat{\alpha}_{b,i,j} \\
E\big[\log p(\xi|\rho_q,\gamma_\mathbf{A},\gamma_\mathbf{B})\big] &= \frac{-K(KP+\ell)}{2}\log 2\pi + \frac{K(Kp+\ell)}{2}\big(\psi(\hat{\alpha}_q)-\log\hat{\beta}_q\big) \\
&\quad + \sum_{i=1}^{K}\sum_{j=1}^{Kp}\big(\psi(\hat{\alpha}_{a,i,j})-\log\hat{\beta}_{a,i,j}\big) + \sum_{i=1}^{K}\sum_{j=1}^{\ell}\big(\psi(\hat{\alpha}_{b,i,j})-\log\hat{\beta}_{b,i,j}\big) \\
&\quad -\frac{1}{2}\mathrm{tr}\bigg(E_{q(\gamma_\mathbf{A})q(\gamma_\mathbf{B})}\big[\Lambda_\xi\big]\big(\hat{\boldsymbol{\Lambda}}_\xi^{-1} + \frac{\hat{\alpha}_q}{\hat{\beta}_q}\hat{\boldsymbol{\xi}}_\mu\hat{\boldsymbol{\xi}}_\mu^\top\big)\bigg) \\
E_{q(\gamma_\mathbf{A})}\big[\log p(\gamma_\mathbf{A})\big] &= K^2p\big(-\log\Gamma(\alpha_a) + \alpha_a\log\beta_a\big) \\
&\quad + \sum_{i=1}^{K}\sum_{j=1}^{Kp}(\alpha_a-1)(\psi(\hat{\alpha}_{a,i,j})-\log\hat{\beta}_{a,i,j}) - \frac{\beta_a\hat{\alpha}_{a,i,j}}{\hat{\beta}_{a,i,j}} \\
E_{q(\gamma_\mathbf{B})}\big[\log p(\gamma_\mathbf{B})\big] &= K\ell\big(-\log\Gamma(\alpha_b) + \alpha_b\log\beta_b\big) \\
&\quad + \sum_{i=1}^{K}\sum_{j=1}^{\ell}(\alpha_b-1)(\psi(\hat{\alpha}_{b,i,j})-\log\hat{\beta}_{b,i,j}) - \frac{\beta_b\hat{\alpha}_{b,i,j}}{\hat{\beta}_{b,i,j}}.
\end{aligned}$$

Self-Connected ARD The KL divergence from the approximate posterior $q(\theta)$ to the prior $p(\theta)$ can be shown to be

$$\begin{aligned}
KL(q(\theta)||p(\theta)) &= -H(q(\theta)) - E_{q(\theta)}\big[\log p(\theta)\big] \\
&= -H(q(\rho_q)) - E_{q(\rho_q)}\big[H(q(\xi|\rho_q))\big] - H(q(\gamma_\mathbf{a})) \\
&\qquad -H(q(\gamma_\mathbf{b})) - H(q(\rho_r)) - H(q(\rho_\mathbf{s})) \\
&\quad -E_{q(\rho_q)}\big[\log p(\rho_q)\big] - E_{q(\rho_q,\xi)q(\gamma_a)q(\gamma_b)}\big[\log p(\xi|\rho_q,\gamma_\mathbf{a},\gamma_\mathbf{b})\big] \\
&\qquad -E_{q(\gamma_a)}\big[\log p(\gamma_\mathbf{a})\big] - E_{q(\gamma_b)}\big[\log p(\gamma_\mathbf{b})\big] \\
&\qquad -E_{q(\rho_r)}\big[\log p(\rho_r)\big] - E_{q(\rho_s)}\big[\log p(\rho_\mathbf{s})\big]
\end{aligned}$$

where the only differing term from the independent ARD are $H(q(\gamma_{\mathbf{a}}))$, $H(q(\gamma_{\mathbf{b}}))$, $E_{q(\rho_q,\xi)q(\gamma_a)q(\gamma_b)}\big[\log p(\xi|\rho_q,\gamma_{\mathbf{a}},\gamma_{\mathbf{b}})\big]$, $E_{q(\gamma_a)}\big[\log p(\gamma_{\mathbf{a}})\big]$ and $E_{q(\gamma_b)}\big[\log p(\gamma_{\mathbf{b}})\big]$ which are given as

$$\begin{aligned}
H(q(\gamma_{\mathbf{a}})) &= \sum_{i=1}^{K}\sum_{j=1}^{K} \log\Gamma(\hat{\alpha}_{a',i,j}) - \log\hat{\beta}_{a',i,j} + (1-\hat{\alpha}_{a',i,j})\psi(\hat{\alpha}_{a',i,j}) + \hat{\alpha}_{a',i,j} \\
H(q(\rho_{\mathbf{b}})) &= \sum_{i=1}^{\ell} \log\Gamma(\hat{\alpha}_{b',i}) - \log\hat{\beta}_{b',i} + (1-\hat{\alpha}_{b',i})\psi(\hat{\alpha}_{b',i}) + \hat{\alpha}_{b',i} \\
E\big[\log p(\xi|\rho_q,\gamma_{\mathbf{a}},\gamma_{\mathbf{b}})\big] &= \frac{-K(KP+\ell)}{2}\log 2\pi + \frac{K(Kp+\ell)}{2}\big(\psi(\hat{\alpha}_q) - \log\hat{\beta}_q\big) \\
&\quad + \sum_{i=1}^{K}\sum_{j=1}^{K}\big(\psi(\hat{\alpha}_{a',i,j}) - \log\hat{\beta}_{a',i,j}\big) + \sum_{i=1}^{\ell}\big(\psi(\hat{\alpha}_{b',i}) - \log\hat{\beta}_{b',i}\big) \\
&\quad - \frac{1}{2}\mathrm{tr}\Big(E_{q(\gamma_a)q(\gamma_b)}\big[\Lambda_\xi\big]\big(\hat{\mathbf{\Lambda}}_\xi^{-1} + \frac{\hat{\alpha}_q}{\hat{\beta}_q}\hat{\boldsymbol{\xi}}_\mu\hat{\boldsymbol{\xi}}_\mu^\top\big)\Big) \\
E_{q(\gamma_a)}\big[\log p(\gamma_{\mathbf{a}})\big] &= K^2\big(-\log\Gamma(\alpha_a) + \alpha_a\log\beta_a\big) \\
&\quad + \sum_{i=1}^{K}\sum_{j=1}^{K}(\alpha_a - 1)(\psi(\hat{\alpha}_{a',i,j}) - \log\hat{\beta}_{a',i,j}) - \frac{\beta_a\hat{\alpha}_{a',i,j}}{\hat{\beta}_{a',i,j}}
\end{aligned}$$

$$\begin{aligned}
E_{q(\gamma_b)}\big[\log p(\gamma_{\mathbf{b}})\big] &= \ell\big(-\log\Gamma(\alpha_b) + \alpha_b\log\beta_b\big) \\
&\quad + \sum_{i=1}^{\ell}(\alpha_b - 1)(\psi(\hat{\alpha}_{b',i}) - \log\hat{\beta}_{b',i}) - \frac{\beta_b\hat{\alpha}_{b',i}}{\hat{\beta}_{b',i}}.
\end{aligned}$$

[1] R. E. Hoffman, N. N. Boutros, S. Hu, R. M. Berman, J. H. Krystal, and D. S. Charney, “Transcranial magnetic stimulation and auditory hallucinations in schizophrenia,” *The Lancet*, vol. 355, no. 9209, pp. 1073–1075, 2000.

[2] “Brain stimulation therapies.” [Online]. Available: https://www.nimh.nih.gov/health/topics/brain-stimulation-therapies/brain-stimulation-therapies.shtml

[3] M. Cossu, M. Schiariti, S. Francione, D. Fuschillo, F. Gozzo, L. Nobili, F. Cardinale, L. Castana, and G. L. Russo, “Stereoelectroencephalography in the presurgical evaluation of focal epilepsy in infancy and early childhood,” *Journal of Neurosurgery: Pediatrics*, vol. 9, no. 3, pp. 290–300, 2012.

[4] M. Massimini, F. Ferrarelli, R. Huber, S. K. Esser, H. Singh, and G. Tononi, “Breakdown of cortical effective connectivity during sleep,” *Science*, vol. 309, no. 5744, pp. 2228–2232, 2005.

[5] S. J. Kiebel, M. I. Garrido, R. J. Moran, and K. J. Friston, “Dynamic causal modelling for eeg and meg,” *Cognitive neurodynamics*, vol. 2, no. 2, p. 121, 2008.

[6] M. Boly, R. Moran, M. Murphy, P. Boveroux, M.-A. Bruno, Q. Noirhomme, D. Ledoux, V. Bonhomme, J.-F. Brichant, G. Tononi *et al.*, “Connectivity changes underlying spectral eeg changes during propofol-induced loss of consciousness,” *Journal of Neuroscience*, vol. 32, no. 20, pp. 7082–7090, 2012.

[7] M. I. Garrido, J. M. Kilner, S. J. Kiebel, K. E. Stephan, and K. J. Friston, “Dynamic causal modelling of evoked potentials: a reproducibility study,” *Neuroimage*, vol. 36, no. 3, pp. 571–580, 2007.

[8] M. Ding, S. L. Bressler, W. Yang, and H. Liang, “Short-window spectral analysis of cortical event-related potentials by adaptive multivariate autoregressive modeling: data preprocessing, model validation, and variability assessment.” *Biological Cybernetics*, vol. 83, pp. 35–45, 2000.

[9] L. Astolfi, F. Cincotti, D. Mattia, F. D. V. Fallani, A. Tocci, A. Colosimo, S. Salinari, M. G. Marciani, W. Hesse, H. Witte *et al.*, “Tracking the time-varying

cortical connectivity patterns by adaptive multivariate estimators," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 3, pp. 902–913, 2008.

[10] B. L. P. Cheung, B. A. Riedner, G. Tononi, and B. D. Van Veen, "Estimation of cortical connectivity from eeg using state-space models," *IEEE Transactions on Biomedical Engineering*, vol. 57, pp. 2122–2134, 2010.

[11] M. Ding, Y. Chen, and S. L. Bressler, *Granger Causality: Basic Theory and Application to Neuroscience*. Wiley-VCH Verlag GmbH & Co. KGaA, 2006, ch. 17, pp. 437–460. [Online]. Available: http://dx.doi.org/10.1002/9783527609970.ch17

[12] A. Brovelli, M. Ding, A. Ledberg, Y. Chen, R. Nakamura, and S. L. Bressler, "Beta oscillations in a large-scale sensorimotor cortical network: directional influences revealed by Granger causality," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 101, pp. 9849–9854, 2004.

[13] C. Bernasconi and P. König, "On the directionality of cortical interactions studied by structural analysis of electrophysiological recordings," *Biological Cybernetics*, vol. 81, pp. 199–210, 1999. [Online]. Available: http://dx.doi.org/10.1007/s004220050556

[14] M. Winterhalder, B. Schelter, W. Hesse, K. Schwab, L. Leistritz, D. Klan, R. Bauer, J. Timmer, and H. Witte, "Comparison of linear signal processing techniques to infer directed interactions in multivariate neural systems," *Signal Process.*, vol. 85, pp. 2137–2160, 2005. [Online]. Available: http://dx.doi.org/10.1016/j.sigpro.2005.07.011

[15] A. Korzeniewska, C. M. Crainiceanu, R. Kuś, P. J. Franaszczuk, and N. E. Crone, "Dynamics of event-related causality in brain electrical activity," *Hum Brain Mapp*, vol. 29, no. 10, pp. 1170–1192, 2008.

[16] F. Babiloni, F. Cincotti, C. Babiloni, F. Carducci, D. Mattia, L. Astolfi, A. Basilisco, P. M. Rossini, L. Ding, Y. Ni, J. Cheng, K. Christine, J. Sweeney, and B. He, "Estimation of the cortical functional connectivity with the multimodal integration of high-resolution EEG and fMRI data by directed transfer function," *Neuroimage*, vol. 24, pp. 118–131, 2005.

[17] S. Malekpour, Z. Li, B. Cheung, E. Castillo, L. Papanicolaou, A. Kramer, J. Fletcher, and B. Van Veen, "Interhemispheric effective and functional cortical connectivity signatures of spina bifida are consistent with callosal anomaly," *Brain Connectivity*, vol. 2, no. 3, pp. 142–154, 2012.

[18] M. J. Kamiński and K. J. Blinowska, "A new method of the description of the information flow in the brain structures," *Biol Cybern*, vol. 65, no. 3, pp. 203–210, 1991.

[19] L. A. Baccalá and K. Sameshima, "Partial directed coherence: a new concept in neural structure determination," *Biol Cybern*, vol. 84, no. 6, pp. 463–474, 2001.

[20] J. F. Geweke, "Measures of Conditional Linear Dependence and Feedback Between Time Series," *Journal of the American Statistical Association*, vol. 79, pp. 907–915, 1984. [Online]. Available: http://www.jstor.org/stable/2288723

[21] A. B. Barrett and A. K. Seth, "Practical measures of integrated information for time-series data," *PLoS Comput Biol*, vol. 7, p. e1001052, 2011.

[22] E. Möller, B. Schack, M. Arnold, and H. Witte, "Instantaneous multivariate EEG coherence analysis by means of adaptive high-dimensional autoregressive models," *J. Neurosci. Methods*, vol. 105, no. 2, pp. 143–158, 2001.

[23] L. Astolfi, F. Cincotti, D. Mattia, F. De Vico Fallani, A. Tocci, A. Colosimo, S. Salinari, M. G. Marciani, W. Hesse, H. Witte, M. Ursino, M. Zavaglia, and F. Babiloni, "Tracking the time-varying cortical connectivity patterns by adaptive multivariate estimators," *IEEE Trans Biomed Eng*, vol. 55, no. 3, pp. 902–913, 2008.

[24] C. Munari, D. Hoffmann, S. Francione, P. Kahane, L. Tassi, G. Lo Russo, and A. L. Benabid, "Stereo-electroencephalography methodology: advantages and limits," *Acta Neurol. Scand., Suppl.c*, vol. 152, pp. 56–67, 1994.

[25] H. Lütkepohl, *New Introduction to Multiple Time Series Analysis*. Springer, 2006.

[26] J. B. Ranck, “Which elements are excited in electrical stimulation of mammalian central nervous system: A review,” *Brain Research*, vol. 98, pp. 417 – 440, 1975.

[27] G. Tononi, “An information integration theory of consciousness.” *BMC neuroscience*, vol. 5, p. 42, 2004. [Online]. Available: http://dx.doi.org/10.1186/1471-2202-5-42

[28] S. Dehaene, J. P. Changeux, L. Naccache, J. Sackur, and C. Sergent, “Conscious, preconscious, and subliminal processing: a testable taxonomy,” *Trends Cogn. Sci. (Regul. Ed.)*, vol. 10, no. 5, pp. 204–211, 2006.

[29] S. Laureys, “The neural correlate of (un)awareness: lessons from the vegetative state,” *Trends Cogn. Sci. (Regul. Ed.)*, vol. 9, no. 12, pp. 556–559, 2005.

[30] A. K. Seth, Z. Dienes, A. Cleeremans, M. Overgaard, and L. Pessoa, “Measuring consciousness: relating behavioural and neurophysiological approaches,” *Trends Cogn. Sci. (Regul. Ed.)*, vol. 12, no. 8, pp. 314–321, 2008.

[31] M. Cossu, F. Cardinale, L. Castana, A. Citterio, S. Francione, L. Tassi, A. L. Benabid, and G. Lo Russo, “Stereoelectroencephalography in the presurgical evaluation of focal epilepsy: a retrospective analysis of 215 procedures,” *Neurosurgery*, vol. 57, pp. 706–718, 2005.

[32] A. Valentín, M. Anderson, G. Alarcón, J. J. Garcia Seoane, R. Selway, C. D. Binnie, and C. E. Polkey, “Responses to single pulse electrical stimulation identify epileptogenesis in the human brain in vivo,” *Brain*, vol. 125, pp. 1709–1718, 2002.

[33] A. Valentín, G. Alarcón, M. Honavar, J. J. Garcia Seoane, R. P. Selway, C. E. Polkey, and C. D. Binnie, “Single pulse electrical stimulation for identification of structural abnormalities and prediction of seizure outcome after epilepsy surgery: a prospective study,” *Lancet Neurol*, vol. 4, pp. 718–726, 2005.

[34] L. Nobili, M. Ferrara, F. Moroni, L. De Gennaro, G. L. Russo, C. Campus, F. Cardinale, and F. De Carli, “Dissociated wake-like and sleep-like electrocortical activity during sleep,” *Neuroimage*, vol. 58, no. 2, pp. 612–619, 2011.

[35] L. Nobili, L. De Gennaro, P. Proserpio, F. Moroni, S. Sarasso, A. Pigorini, F. De Carli, and M. Ferrara, "Local aspects of sleep: observations from intracerebral recordings in humans," *Prog. Brain Res.*, vol. 199, pp. 219–232, 2012.

[36] A. Rechtschaffen and A. Kales, Eds., *A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects*, ser. NIH Publication No. 204. Washington DC: US Government Printing Office, National Institute of Health Publication, 1968.

[37] P. Bloomfield, *Fourier Analysis of Time Series: An Introduction (Wiley Series in Probability and Statistics)*, 2nd ed. Wiley-Interscience, 2 2000.

[38] L. Barnett and A. K. Seth, "Behaviour of Granger causality under filtering: theoretical invariance and practical application," *J. Neurosci. Methods*, vol. 201, no. 2, pp. 404–419, Oct 2011.

[39] K. I. Penny, "Appropriate critical values when testing for a single multivariate outlier by using the mahalanobis distance," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 45, no. 1, pp. pp. 73–81, 1996.

[40] R. Matsumoto, D. R. Nair, E. LaPresto, I. Najm, W. Bingaman, H. Shibasaki, and H. O. Lüders, "Functional connectivity in the human language system: a cortico-cortical evoked potential study," *Brain*, vol. 127, pp. 2316–2330, 2004.

[41] A. D. R. McQuarrie and C.-L. Tsai, *Regression and Time Series Model Selection*. World Scientific Pub Co Inc, 1998.

[42] B. L. P. Cheung, R. D. Nowak, H. C. Lee, W. van Drongelen, and B. D. Van Veen, "Cross validation for selection of cortical interaction models from scalp eeg or meg," *IEEE Trans. Biomed. Engineering*, vol. 59, pp. 504–514, 2012.

[43] Y. Hong, "Consistent testing for serial correlation of unknown form," *Econometrica*, vol. 64, pp. 837–64, 1996.

[44] P. Duchesne and R. Roy, "On consistent testing for serial correlation of unknown form in vector time series models," *J. Multivar. Anal.*, vol. 89, pp. 148–180, 2004.

[45] G. Tononi, “Consciousness as integrated information: a provisional manifesto.” *The Biological Bulletin*, vol. 215, pp. 216–242, 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19098144

[46] ——, “Information integration: its relevance to brain function and consciousness,” *Archives Italiennes de Biologie*, vol. 148, pp. 299–322, 2010.

[47] D. Balduzzi and G. Tononi, “Integrated information in discrete dynamical systems: Motivation and theoretical framework,” *PLoS Comput Biol*, vol. 4, no. 6, p. e1000091, 2008.

[48] T. M. Cover and J. A. Thomas, *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.

[49] R. Matsumoto, D. R. Nair, E. LaPresto, W. Bingaman, H. Shibasaki, and H. O. Lüders, “Functional connectivity in human cortical motor system: a cortico-cortical evoked potential study,” *Brain*, vol. 130, pp. 181–197, 2007.

[50] R. Matsumoto, D. R. Nair, A. Ikeda, T. Fumuro, E. LaPresto, N. Mikuni, W. Bingaman, S. Miyamoto, H. Fukuyama, R. Takahashi, I. Najm, H. Shibasaki, and H. O. Lüders, “Parieto-frontal network in humans studied by cortico-cortical evoked potential,” *Human Brain Mapping*, pp. 1–17, 2011. [Online]. Available: http://dx.doi.org/10.1002/hbm.21407

[51] M. Massimini, F. Ferrarelli, R. Huber, S. K. Esser, H. Singh, and G. Tononi, “Breakdown of cortical effective connectivity during sleep,” *Science*, vol. 309, no. 5744, pp. 2228–2232, Sep 2005.

[52] P. A. Valdes-Sosa, A. Roebroeck, J. Daunizeau, and K. Friston, “Effective connectivity: influence, causality and biophysical modeling,” *Neuroimage*, vol. 58, no. 2, pp. 339–361, 2011.

[53] B. L. P. Cheung, B. A. Riedner, G. Tononi, and B. D. Van Veen, “Estimation of cortical connectivity from eeg using state-space models,” *IEEE Transactions on Biomedical engineering*, vol. 57, no. 9, pp. 2122–2134, 2010.

[54] A. Brovelli, M. Ding, A. Ledberg, Y. Chen, R. Nakamura, and S. L. Bressler, “Beta oscillations in a large-scale sensorimotor cortical network: directional

influences revealed by granger causality," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 26, pp. 9849–9854, 2004.

[55] H. Lütkepohl, *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.

[56] E. Bullmore and O. Sporns, "The economy of brain network organization," *Nature Reviews Neuroscience*, vol. 13, no. 5, pp. 336–349, 2012.

[57] A. Bolstad, B. D. Van Veen, and R. Nowak, "Causal network inference via group sparse regularization," *Signal Processing, IEEE Transactions on*, vol. 59, no. 6, pp. 2628–2641, 2011.

[58] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[59] C. Munari, D. Hoffmann, S. Fracione, P. Kahane, L. Tassi, G. L. Russo, and A. Benabid, "Stereo-electroencephalography methodology: advantages and limits," *Acta Neurologica Scandinavica*, vol. 89, no. S152, pp. 56–67, 1994.

[60] M. Ding, S. L. Bressler, W. Yang, and H. Liang, "Short-window spectral analysis of cortical event-related potentials by adaptive multivariate autoregressive modeling: data preprocessing, model validation, and variability assessment," *Biological cybernetics*, vol. 83, no. 1, pp. 35–45, 2000.

[61] J. Huang, P. Breheny, and S. Ma, "A selective review of group selection in high-dimensional models," *Statistical Science*, vol. 27, no. 4, pp. 481–499, 2012.

[62] P. Duchesne and R. Roy, "On consistent testing for serial correlation of unknown form in vector time series models," *Journal of Multivariate Analysis*, vol. 89, no. 1, pp. 148–180, 2004.

[63] J.-Y. Chang, A. Pigorini, M. Massimini, G. Tononi, L. Nobili, and B. D. Van Veen, "Multivariate autoregressive models with exogenous inputs for intracerebral responses to direct electrical stimulation of the human brain," *Frontiers in human neuroscience*, vol. 6, 2012.

[64] R. Matsumoto, D. R. Nair, A. Ikeda, T. Fumuro, E. LaPresto, N. Mikuni, W. Bingaman, S. Miyamoto, H. Fukuyama, R. Takahashi *et al.*, "Parieto-frontal network in humans studied by cortico-cortical evoked potential," *Human brain mapping*, vol. 33, no. 12, pp. 2856–2872, 2012.

[65] R. Mazumder, J. H. Friedman, and T. Hastie, "Sparsenet: Coordinate descent with nonconvex penalties," *Journal of the American Statistical Association*, vol. 106, no. 495, 2011.

[66] Y. Chen, S. L. Bressler, and M. Ding, "Frequency decomposition of conditional granger causality and application to multivariate neural field potential data," *Journal of neuroscience methods*, vol. 150, no. 2, pp. 228–237, 2006.

[67] J. Virtanen, J. Ruohonen, R. Näätänen, and R. Ilmoniemi, "Instrumentation for the measurement of electric brain responses to transcranial magnetic stimulation," *Medical & biological engineering & computing*, vol. 37, no. 3, pp. 322–326, 1999.

[68] K. Iramina, T. Maeno, Y. Nonaka, and S. Ueno, "Measurement of evoked electroencephalography induced by transcranial magnetic stimulation," *Journal of applied physics*, vol. 93, no. 10, pp. 6718–6720, 2003.

[69] G. Thut, J. R. Ives, F. Kampmann, M. A. Pastor, and A. Pascual-Leone, "A new device and protocol for combining tms and online recordings of eeg and evoked potentials," *Journal of neuroscience methods*, vol. 141, no. 2, pp. 207–217, 2005.

[70] F. Ferrarelli, M. Massimini, S. Sarasso, A. Casali, B. A. Riedner, G. Angelini, G. Tononi, and R. A. Pearce, "Breakdown in cortical effective connectivity during midazolam-induced loss of consciousness," *Proceedings of the National Academy of Sciences*, vol. 107, no. 6, pp. 2681–2686, 2010.

[71] M. Rosanova, O. Gosseries, S. Casarotto, M. Boly, A. G. Casali, M.-A. Bruno, M. Mariotti, P. Boveroux, G. Tononi, S. Laureys *et al.*, "Recovery of cortical effective connectivity and recovery of consciousness in vegetative patients," *Brain*, p. awr340, 2012.

[72] A. G. Casali, O. Gosseries, M. Rosanova, M. Boly, S. Sarasso, K. R. Casali, S. Casarotto, M.-A. Bruno, S. Laureys, G. Tononi *et al.*, "A theoretically based index of consciousness independent of sensory processing and behavior," *Science translational medicine*, vol. 5, no. 198, pp. 198ra105–198ra105, 2013.

[73] S. Sarasso, M. Boly, M. Napolitani, O. Gosseries, V. Charland-Verville, S. Casarotto, M. Rosanova, A. G. Casali, J.-F. Brichant, P. Boveroux *et al.*, "Consciousness and complexity during unresponsiveness induced by propofol, xenon, and ketamine," *Current Biology*, vol. 25, no. 23, pp. 3099–3105, 2015.

[74] F. Cona, M. Zavaglia, M. Massimini, M. Rosanova, and M. Ursino, "A neural mass model of interconnected regions simulates rhythm propagation observed via tms-eeg," *NeuroImage*, vol. 57, no. 3, pp. 1045–1058, 2011.

[75] S. K. Esser, S. Hill, and G. Tononi, "Breakdown of effective connectivity during slow wave sleep: investigating the mechanism underlying a cortical gate using large-scale modeling," *Journal of Neurophysiology*, vol. 102, no. 4, pp. 2096–2111, 2009.

[76] A. Pigorini, S. Sarasso, P. Proserpio, C. Szymanski, G. Arnulfo, S. Casarotto, M. Fecchio, M. Rosanova, M. Mariotti, G. L. Russo *et al.*, "Bistability breaks-off deterministic responses to intracortical stimulation during non-rem sleep," *Neuroimage*, vol. 112, pp. 105–113, 2015.

[77] A. G. Casali, S. Casarotto, M. Rosanova, M. Mariotti, and M. Massimini, "General indices to characterize the electrical response of the cerebral cortex to tms," *Neuroimage*, vol. 49, no. 2, pp. 1459–1468, 2010.

[78] S. Komssi, P. Savolainen, J. Heiskala, and S. Kähkönen, "Excitation threshold of the motor cortex estimated with transcranial magnetic stimulation electroencephalography," *Neuroreport*, vol. 18, no. 1, pp. 13–16, 2007.

[79] M. Massimini, F. Ferrarelli, S. K. Esser, B. A. Riedner, R. Huber, M. Murphy, M. J. Peterson, and G. Tononi, "Triggering sleep slow waves by transcranial magnetic stimulation," *Proceedings of the National Academy of Sciences*, vol. 104, no. 20, pp. 8496–8501, 2007.

[80] F. Ferrarelli, M. Massimini, M. J. Peterson, B. A. Riedner, M. Lazar, M. J. Murphy, R. Huber, M. Rosanova, A. L. Alexander, N. Kalin *et al.*, "Reduced evoked gamma oscillations in the frontal cortex in schizophrenia patients: a tms/eeg study," *American Journal of Psychiatry*, 2008.

[81] E. M. ter Braack, C. C. de Vos, and M. J. van Putten, "Masking the auditory evoked potential in tms–eeg: a comparison of various methods," *Brain topography*, vol. 28, no. 3, pp. 520–528, 2015.

[82] O. Gosseries, S. Sarasso, S. Casarotto, M. Boly, C. Schnakers, M. Napolitani, M.-A. Bruno, D. Ledoux, J.-F. Tshibanda, M. Massimini *et al.*, "On the cerebral origin of eeg responses to tms: insights from severe cortical lesions," *Brain Stimulation: Basic, Translational, and Clinical Research in Neuromodulation*, vol. 8, no. 1, pp. 142–149, 2015.

[83] T. Mutanen, H. Mäki, and R. J. Ilmoniemi, "The effect of stimulus parameters on tms–eeg muscle artifacts," *Brain Stimulation: Basic, Translational, and Clinical Research in Neuromodulation*, vol. 6, no. 3, pp. 371–376, 2013.

[84] M. Fecchio, A. Pigorini, A. Comanducci, S. Sarasso, S. Casarotto, I. Premoli, C.-C. Derchi, A. Mazza, S. Russo, F. Resta *et al.*, "The spectral features of eeg responses to transcranial magnetic stimulation of the primary motor cortex depend on the amplitude of the motor evoked potentials," *PloS one*, vol. 12, no. 9, p. e0184910, 2017.

[85] T. Anderson, *An Introduction to Multivariate Statistical Analysis*, ser. Wiley Series in Probability and Statistics. Wiley, 2003.

[86] P. Berg and M. Scherg, "A fast method for forward computation of multiple-shell spherical head models," *Electroencephalography and clinical neurophysiology*, vol. 90, no. 1, pp. 58–64, 1994.

[87] Z. Zhang, "A fast method to compute surface potentials generated by dipoles within multilayer anisotropic spheres," *Physics in medicine and biology*, vol. 40, no. 3, p. 335, 1995.

[88] K. Friston, R. Henson, C. Phillips, and J. Mattout, "Bayesian estimation of evoked and induced responses," *Human brain mapping*, vol. 27, no. 9, pp. 722–735, 2006.

[89] J. Mattout, C. Phillips, W. D. Penny, M. D. Rugg, and K. J. Friston, "Meg source localization under multiple constraints: an extended bayesian framework," *NeuroImage*, vol. 30, no. 3, pp. 753–767, 2006.

[90] C. Phillips, J. Mattout, M. D. Rugg, P. Maquet, and K. J. Friston, "An empirical bayesian solution to the source reconstruction problem in eeg," *NeuroImage*, vol. 24, no. 4, pp. 997–1011, 2005.

[91] J. Lv, D. M. Simpson, and S. L. Bell, "Objective detection of evoked potentials using a bootstrap technique," *Medical engineering & physics*, vol. 29, no. 2, pp. 191–198, 2007.

[92] D. Pantazis, T. E. Nichols, S. Baillet, and R. M. Leahy, "A comparison of random field theory and permutation methods for the statistical analysis of meg data," *NeuroImage*, vol. 25, no. 2, pp. 383–394, 2005.

[93] A. Lempel and J. Ziv, "On the complexity of finite sequences," *IEEE Transactions on information theory*, vol. 22, no. 1, pp. 75–81, 1976.

[94] M. Oizumi, S.-i. Amari, T. Yanagawa, N. Fujii, and N. Tsuchiya, "Measuring integrated information from the decoding perspective," *PLoS Comput Biol*, vol. 12, no. 1, p. e1004654, 2016.

[95] T. Limpiti, B. D. Van Veen, and R. T. Wakai, "Cortical patch basis model for spatially extended neural activity," *Biomedical Engineering, IEEE Transactions on*, vol. 53, no. 9, pp. 1740–1754, 2006.

[96] S. M. Kay, "Fundamentals of statistical signal processing, volume i: Estimation theory (v. 1)," 4 1993. [Online]. Available: http://amazon.com/o/ASIN/0133457117/

[97] R. H. Shumway and D. S. Stoffer, "An approach to time series smoothing and forecasting using the em algorithm," *Journal of time series analysis*, vol. 3, no. 4, pp. 253–264, 1982.

[98] J. G. Dias and M. Wedel, "An empirical comparison of em, sem and mcmc performance for problematic gaussian mixture likelihoods," *Statistics and Computing*, vol. 14, no. 4, pp. 323–332, 2004.

[99] V. A. Lamme and P. R. Roelfsema, “The distinct modes of vision offered by feedforward and recurrent processing,” *Trends in neurosciences*, vol. 23, no. 11, pp. 571–579, 2000.

[100] M. I. Garrido, J. M. Kilner, S. J. Kiebel, and K. J. Friston, “Evoked brain responses are generated by feedback loops,” *Proceedings of the National Academy of Sciences*, vol. 104, no. 52, pp. 20 961–20 966, 2007.

[101] A. Gaillard, “Problems and paradigms in erp research,” *Biological psychology*, vol. 26, no. 1-3, pp. 91–109, 1988.

[102] S. Esser, R. Huber, M. Massimini, M. Peterson, F. Ferrarelli, and G. Tononi, “A direct demonstration of cortical ltp in humans: a combined tms/eeg study,” *Brain research bulletin*, vol. 69, no. 1, pp. 86–94, 2006.

[103] J. McCubbin, T. Yee, J. Vrba, S. Robinson, P. Murphy, H. Eswaran, and H. Preissl, “Bootstrap significance of low snr evoked response,” *Journal of neuroscience methods*, vol. 168, no. 1, pp. 265–272, 2008.

[104] B. L. P. Cheung, R. Nowak, H. C. Lee, W. Drongelen, and B. D. Veen, “Cross validation for selection of cortical interaction models from scalp eeg or meg,” *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 2, pp. 504–514, 2012.

[105] K. J. Friston, L. Harrison, and W. Penny, “Dynamic causal modelling,” *Neuroimage*, vol. 19, no. 4, pp. 1273–1302, 2003.

[106] R. Gaillard, S. Dehaene, C. Adam, S. Clémenceau, D. Hasboun, M. Baulac, L. Cohen, and L. Naccache, “Converging intracranial markers of conscious access,” *PLoS Biol*, vol. 7, no. 3, p. e1000061, 2009.

[107] D. Dentico, B. L. Cheung, J.-Y. Chang, J. Guokas, M. Boly, G. Tononi, and B. Van Veen, “Reversal of cortical information flow during visual imagery as compared to visual perception,” *Neuroimage*, vol. 100, pp. 237–243, 2014.

[108] B. Kundu, J.-Y. Chang, B. R. Postle, and B. D. Van Veen, “Context-specific differences in fronto-parieto-occipital effective connectivity during short-term memory maintenance,” *NeuroImage*, vol. 114, pp. 320–327, 2015.

[109] M. Oizumi, L. Albantakis, and G. Tononi, "From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0," *PLoS Comput Biol*, vol. 10, no. 5, p. e1003588, 2014.

[110] M. J. Beal, "Variational algorithms for approximate bayesian inference," *Ph. D. Thesis, University College London*, 2003.

[111] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *arXiv preprint arXiv:1601.00670*, 2016.

[112] C. K. Carter and R. Kohn, "On gibbs sampling for state space models," *Biometrika*, vol. 81, no. 3, pp. 541–553, 1994.

[113] J. Kruschke, *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press, 2014.

[114] D. B. S. Chiappa, "Unified inference for variational bayesian linear gaussian state-space models," in *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference*, vol. 19. MIT Press, 2007, p. 81.

[115] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Foundations and Trends® in Machine Learning*, vol. 1, no. 1-2, pp. 1–305, 2008.

[116] T. Kollo and D. von Rosen, *Advanced multivariate statistics with matrices*. Springer Science & Business Media, 2006, vol. 579.

[117] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis, Third Edition (Chapman & Hall/CRC Texts in Statistical Science)*, 3rd ed. Chapman and Hall/CRC, 11 2013.

[118] R. M. Neal, "Bayesian learning for neural networks," Ph.D. dissertation, University of Toronto, 1994.

[119] D. P. Wipf and S. S. Nagarajan, "A new view of automatic relevance determination," in *Advances in neural information processing systems*, 2008, pp. 1625–1632.

[120] J. Drugowitsch, "Variational bayesian inference for linear and logistic regression," *arXiv preprint arXiv:1310.5438*, 2013.

[121] W. Penny and S. Roberts, “Bayesian multivariate autoregressive models with structured priors,” *IEE Proceedings-Vision, Image and Signal Processing*, vol. 149, no. 1, pp. 33–41, 2002.

[122] J.-Y. Chang, A. Pigorini, F. Seregni, M. Massimini, L. Nobili, and B. Van Veen, “Sparse multivariate autoregressive models with exogenous inputs for modeling intracerebral responses to direct electrical stimulation of the human brain,” in *Signals, Systems and Computers, 2013 Asilomar Conference on*. IEEE, 2013, pp. 803–807.

[123] D. J. MacKay, “Bayesian methods for adaptive models,” Ph.D. dissertation, California Institute of Technology, 1991.

[124] C. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 10 2006. [Online]. Available: http://amazon.com/o/ASIN/0387310738/

[125] E. B. Fox, “Bayesian nonparametric learning of complex dynamical phenomena,” Ph.D. dissertation, Massachusetts Institute of Technology, 2009.

[126] B. M. Yu, K. V. Shenoy, and M. Sahani, “Derivation of kalman filtering and smoothing equations,” *Saatavissa: http://wwwnpl. stanford. edu/˜ byronyu/papers/derive_ks. pdf. Hakupäivä*, vol. 7, p. 2010, 2004.

[127] U. Paquet, “Bayesian inference for latent variable models,” University of Cambridge, Computer Laboratory, Tech. Rep., 2008.

[128] P. D. Hoff, “Simulation of the matrix bingham–von mises–fisher distribution, with applications to multivariate and relational data,” *Journal of Computational and Graphical Statistics*, vol. 18, no. 2, pp. 438–456, 2009.

[129] B. D. O. Anderson and J. B. Moore, *Optimal Filtering (Dover Books on Electrical Engineering)*. Dover Publications, 2005.

[130] S. Gibson and B. Ninness, “Robust maximum-likelihood estimation of multivariable dynamic systems,” *Automatica*, vol. 41, no. 10, pp. 1667–1682, 2005.

[131] J. Nocedal and S. Wright, *Numerical optimization*. Springer Science & Business Media, 2006.

[132] G. H. Golub, M. Heath, and G. Wahba, “Generalized cross-validation as a method for choosing a good ridge parameter,” *Technometrics*, vol. 21, no. 2, pp. 215–223, 1979.

[133] K. Sekihara and S. S. Nagarajan, *Adaptive spatial filters for electromagnetic brain imaging*. Springer Science & Business Media, 2008.