

Aalto University
School of Science
Degree programme in Engineering Physics and Mathematics

Estimating the activity of regions of interest using Expectation Maximization algorithm

Bachelor's thesis
29.11.2018

Hilkka Hännikäinen

The document can be stored and made available to the public on the open internet pages of Aalto University.
All other rights are reserved.

Author Hilikka Hännikäinen

Title of thesis Estimating the activity of regions of interest using Expectation Maximization algorithm

Degree programme Engineering Physics and Mathematics

Major Mathematics and System Analysis**Code of major** SCI3029

Supervisor Prof. Harri Ehtamo

Thesis advisor(s) D.Sc. Narayan Puthanmadam Subramarinyam

Date 27.11.2018**Number of pages** 23+5**Language** English

Abstract

The human brain is the most complex organ providing a control center for the human body. Understanding its functional principles would reveal valuable information which could be exploited in the treatment of brain diseases.

The function of the human brain is based on regional functional specialization and collaboration between these regions. This collaboration is called functional connectivity which is estimated from the region activity. The activity of the regions can be estimated from the activity of the brain which can be determined by measuring the electromagnetic fields generated by neurons.

The region activity is constructed by combining the activities of multiple neurons within a certain region. Currently, the methods used for reconstruction are quite ad hoc without any physical relevance. To provide more plausible method for the reconstruction of the region activity, the objective of this work is to examine a method based on the obtained data. In this method, the reconstruction is conducted using weights estimated from the data for each region. The method is implemented using expectation maximization algorithm with a state space model which provides a convenient way to obtain the region activity from Kalman filter.

The implemented algorithm was tested using simulations. Though the errors made in the estimation of weights were in general high ranging between 30–100 %, the region activity could be recovered with smaller estimation error somewhat decently. This implies that some of the assumptions made in the construction of the algorithm need modification. A variant of the algorithm with maximum likelihood estimation was applied on real data producing promising results. Thus, carefully designed and implemented data driven method would provide more accurate approach for estimation of the region activity.

Keywords expectation maximization algorithm, maximum likelihood estimation, functional connectivity, parameter estimation, dimensionality reduction

Tekijä Hilka Hännikäinen

Työn nimi Estimating the activity of regions of interest using Expectation-Maximization algorithm

Koulutusohjelma Teknillinen fysiikka ja matematiikka

Pääaine Matematiikka ja systeemitieteet

Pääaineen koodi SCI3029

Vastuopettaja Prof. Harri Ehtamo

Työn ohjaaja(t) TkT Narayan Puthanmadam Subramarinyam

Päivämäärä 27.11.2018

Sivumäärä 23+5

Kieli englanti

Tiivistelmä

Aivot ovat ihmiskehon monimutkaisin elin, jonka tehtävänä on säädellä kehon toimintoja. Aivojen toimintaperiaatteiden ymmärtäminen tarjoaisi arvokasta tietoa, jota voitaisiin hyödyntää aivope-
räisten sairauksien hoidossa.

Aivojen toiminta perustuu funktionaaliseen erikoistumiseen alueittain sekä yhteistyöhön erilaisiin tehtäviin erikoistuneiden alueiden välillä. Tätä yhteistyötä kutsutaan myös funktionaaliseksi yhdistyneisyydeksi, jota arvioidaan alueiden aktiivisuutta kuvaavasta datasta. Alueen aktiivisuus voidaan määrittää koko aivojen aktiivisuudesta, joka saadaan mittaamalla neuronien synnyttä-
mien sähkömagneettisten kenttien voimakkuutta.

Tietyn alueen aktiivisuus saadaan yhdistämällä alueessa olevien neuronien aktiivisuus. Tällä het-
kellä neuroneiden aktiivisuuden yhdistämiseksi käytetyt menetelmät ovat ad hoc -tyyppisiä ilman
fysikaalisia perusteita. Tämän työn tarkoituksena on tutkia dataan pohjautuvan menetelmän so-
veltuvuutta, jotta aivojen alueiden aktiivisuus voidaan määrittää mahdollisimman luotettavasti ja
tarkasti. Suunnitellussa menetelmässä alueen aktiivisuus saadaan estimoimalla datasta painoker-
toimet jokaiselle alueelle. Menetelmä toteutetaan expectation maximization eli odotusarvon mak-
simointi -algoritmin avulla käyttäen tila-aikamallia, josta alueen aktiivisuus saadaan hyödyntäen
Kalmanin suodinta.

Kehitettyä algoritmia testattiin käyttäen simulaatioita. Vaikka painokertoimien estimoinnissa syn-
tyneet virheet olivat keskimäärin korkeita vaihdellen välillä 30–100 %, alueen aktiivisuus pystyt-
tiin estimoimaan kohtuullisesti pienemmällä virheillä. Tämä viittaa algoritmista käytetyn mallin
oletusten virheellisyyteen. Algoritmista toteutettiin muunneltua, jossa käytettiin suurimman us-
kottavuuden estimointia painokertoimien estimoimiseksi. Muunneltua algoritmia testattiin oike-
alla datalla, mikä tuotti lupaavia tuloksia. Tämän perusteella huolellisesti suunniteltu ja toteutettu
dataan pohjautuva menetelmä alueiden aktiivisuuden estimoimiseksi voisi tuottaa tarkempia tu-
loksia verrattuna nykyisiin menetelmiin.

Avainsanat odotusarvon maksimointi -algoritmi, suurimman uskottavuuden estimointi, funktio-
naalinen yhdistyneisyys, parametrien estimointi

Contents

1	Introduction	1
2	Neural signaling	2
2.1	Structure of a neuron	2
2.2	Generation of a neural signal and approximation for a neural source	3
3	Measuring and modeling the activity	4
3.1	Measurement techniques	4
3.2	The inverse problem in the estimation of activity	5
3.3	Estimating the activity and connectivity between regions of interest	7
4	Methods	9
4.1	State space model	9
4.2	Expectation Maximization algorithm	10
4.3	Kalman Filter	11
4.4	Formulation of the state space model and likelihood function .	12
5	Results	16
6	Conclusions	20
	References	22
A	Derivation of the log-likelihood function	24
B	Kalman prediction, filtering and smoothing equations	25

1 Introduction

The human brain is the most complex organ composed of approximately 100 billion neurons with 1000 – 10000 connections between them (Tortora and Derrickson, 2017). This complexity enables the brain to process a vast amount of electrical signals into meaningful patterns. Thus, the brain can be considered as the control center of the human body (Tortora and Derrickson, 2017). Understanding its functional principles would provide valuable information which could be exploited in the treatment of brain diseases.

The brain organization has two fundamentals. Functional segregation is defined as the anatomical separation of brain whereas functional integration refers to connections appearing as co-operation between brain regions to perform cognitive tasks. Functional integration can be quantified in terms of functional connectivity which is defined as statistical inter-dependencies in the activity of brain regions. (Schoffelen and Gross, 2009)

The activities of the brain regions can be measured using noninvasive measurement techniques like magnetoencephalography (MEG), electroencephalography (EEG) and functional magnetic resonance imaging (fMRI). MEG and EEG measure the magnetic and electric fields, respectively, generated by neurons providing a direct measure of neural activity. In fMRI, the blood-oxygen-level dependent (BOLD) fluctuations are measured yielding an indirect measure of brain activity. While fMRI has good spatial resolution, it suffers from poor temporal resolution. In contrast, MEG and EEG have high temporal resolution but limited spatial resolution. The estimation of underlying brain activity from MEG or EEG recordings constitutes an under-determined inverse problem which has no unique solution. (Hari and Puce, 2017)

The estimation of the brain activity is a little bit problematic. In MEG and EEG recordings, the sensors measure the magnetic or electric fields of multiple neurons. This results in a widespread representation of the fields which complicates the estimation (Hari and Puce, 2017). In addition to this source leakage, the amount of sensors is considerably smaller than the amount of vertices which approximate an assemblage of coherently behaving neurons (Baillet et al., 2001). To concern the source leakage and high dimensionality of the problem, the activities of multiple vertices are combined to form the activity of a region.

There are various approaches to construct the region activity from estimated activity of the brain for example by taking the mean. However, these meth-

ods have some shortcomings. The developed techniques are quite ad hoc and don't have any physical relevance. In order to exploit the collected data in the estimation of region activity, the objective of this thesis is to develop an algorithm which determines weights for vertices within the region. The weights describe how much each vertex contributes to the activity of region. They are estimated using maximum likelihood estimation performed via expectation maximization algorithm.

This thesis is structured as follows. In section 2 the basic functional principles of a human brain are briefly described. In section 3 the magnetoencephalography is presented together with insights into the modeling and estimation of neural activity. In section 4 the methods applied in this work are described followed by the results in section 5. The conclusions are given in section 6.

2 Neural signaling

2.1 Structure of a neuron

The nervous system of the human body consists of nervous tissue. The nervous tissue comprises of neurons which are nerve cells specialized in relaying electric signals called nerve impulses. Thus neurons form the basis of the organized and controlled function of the human body. (Tortora and Derrickson, 2017)

Neurons possess a great diversity in shape and size but they also share common features in structure. Typically, neurons have three main parts: a cell body, dendrites and an axon which are presented in figure 1. All three parts are surrounded by a cell membrane which is a bilayer of lipids. In neurons, the cell membrane has many essential tasks, especially in the transportation of a neural signal. The cell body contains the essential cellular organelles which ensure the function of the cell. Dendrites and axon are used for communication. Neuron receives information through dendrites which are typically short and highly branched. The information is send forward with axon which is in general long and thin. The axon ends in synaptic terminals which connect the neuron to other cells. (Tortora and Derrickson, 2017)

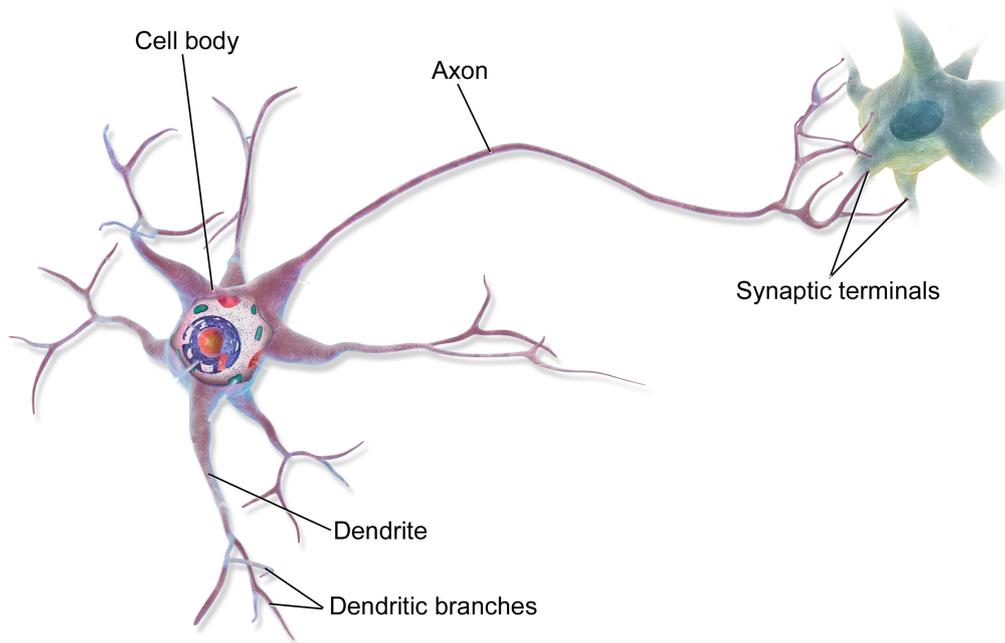


Figure 1: Illustrative picture of the structure of a neuron with the main parts named. Figure reprinted and modified from https://commons.wikimedia.org/wiki/File:Blausen_0657_MultipolarNeuron.png.

2.2 Generation of a neural signal and approximation for a neural source

The main reason for the generation of an extracellular electromagnetic field as a result of neural activity is the electric excitability of the cell membrane (Tortora and Derrickson, 2017). This property enables the neurons to transfer electric currents which superimpose at any given point in the brain (Buzsáki et al., 2012). The superposition of currents leads to a voltage difference called electric potential which gives rise to an electromagnetic field (Buzsáki et al., 2012). The electric currents constituting the electric potential can be divided into intra- and extracellular currents. The intracellular electric currents flowing through the neurons' dendrites and axons are called primary currents. The primary current results in a pile up of charge which is balanced by an extracellular current called volume current flowing through the volume of the brain outside the neuron. (Hämäläinen et al., 1993)

When a primary current is transported, there is an inflow and an outflow of ions through the cell membrane of a neuron. The inflow constitutes an

extracellular sink whereas the outflow of ions comprises a source (Buzsáki et al., 2012). The source and the sink constitute a dipole with a localized electric current depicted in figure 2 with transmembrane fluxes of ions. The approximation of a current dipole can be used in the case of multiple, closely related coherently behaving neurons due to the structure of the uppermost layer of the brain. Neurons in this layer are the main source for the measured neural activity since the dendrites lie almost in parallel with respect to each other as illustrated in figure 2 enabling the constructive superposition of primary currents. The dipole can be described with a location, orientation and amplitude. (Baillet et al., 2001; Hari and Puce, 2017)

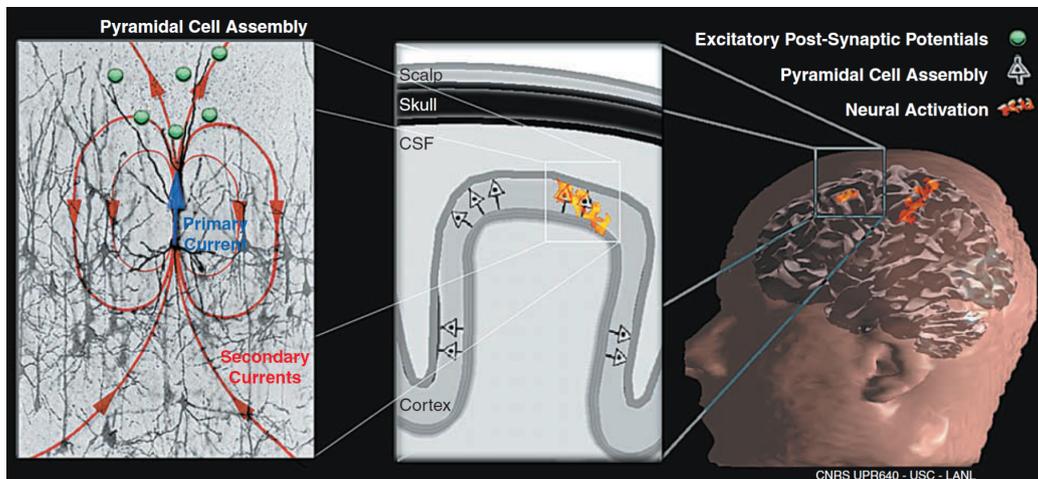


Figure 2: An illustration of the primary and volume currents in neural signal processing with schematic picture of the spatial arrangement of neurons. Figure reprinted from Baillet et al. (2001).

3 Measuring and modeling the activity

3.1 Measurement techniques

Electroencephalography (EEG) and magnetoencephalography (MEG) are commonly used noninvasive measurement techniques based on measuring electromagnetic properties. Both methods possess excellent temporal resolution with limited spatial resolution. EEG and MEG may be combined with other information gathered from the brain, e.g. MRI images to achieve more precise picture about the brain activity. (Hari and Puce, 2017)

The data used in this work is collected using MEG and thus the main focus of the following paragraph is on MEG and its' basic principles. The MEG measurement device consists of a dewar vacuum flask which resembles a giant helmet presented on the left in figure 3. This helmet contains 306 magnetic-field sensors on the inner surface. A schematic illustration of the sensors is depicted on the right in figure 3. Because the magnetic field generated by neural currents is very weak, the sensors have to be extremely sensitive. At the moment, the sensors used are superconducting quantum interference devices i.e. SQUIDs. In the sensors, the magnetic field of the brain generates a current in a coil which is again transformed into a magnetic field sensed by the SQUID. The SQUID is attached to a circuit which transforms the magnetic flux into amplified voltage which is recorded. In generation of the magnetic field based on neural currents, the structure of the cerebral cortex plays a significant role. If the brain is considered as a volume conductor, radially oriented currents don't produce magnetic field outside the conductor. These currents are hence invisible for the measurement devices and only currents with tangential components can be detected making the folded structure important. The superposition of primary currents plays also a significant role making the magnetic fields detectable. The MEG measurement should be conducted inside a shielded room which dampens the Earth's magnetic field and other magnetic fields arising from nearby electrical sources. (Hari and Puce, 2017)

3.2 The inverse problem in the estimation of activity

The objective of the measurements is to estimate the activity of the neural current sources from the collected data which means solving the equation

$$\mathbf{y}_t = \mathbf{G}\mathbf{q}_t + \boldsymbol{\varepsilon}_t , \quad (1)$$

where \mathbf{y}_t is the measured data, \mathbf{G} is the lead field matrix, \mathbf{q}_t is the amplitudes of the sources and $\boldsymbol{\varepsilon}_t$ is the measurement noise. This is an inverse problem. The problem doesn't have a unique solution due to the physical circumstances (Baillet et al., 2001; Hämmäläinen et al., 1993).

In order to solve the inverse problem, the associated forward problem must also be solved. The forward problem means, in the case of MEG, calculating the generated magnetic field based on distribution of the neural sources and the head model. (Tadel et al., 2016)

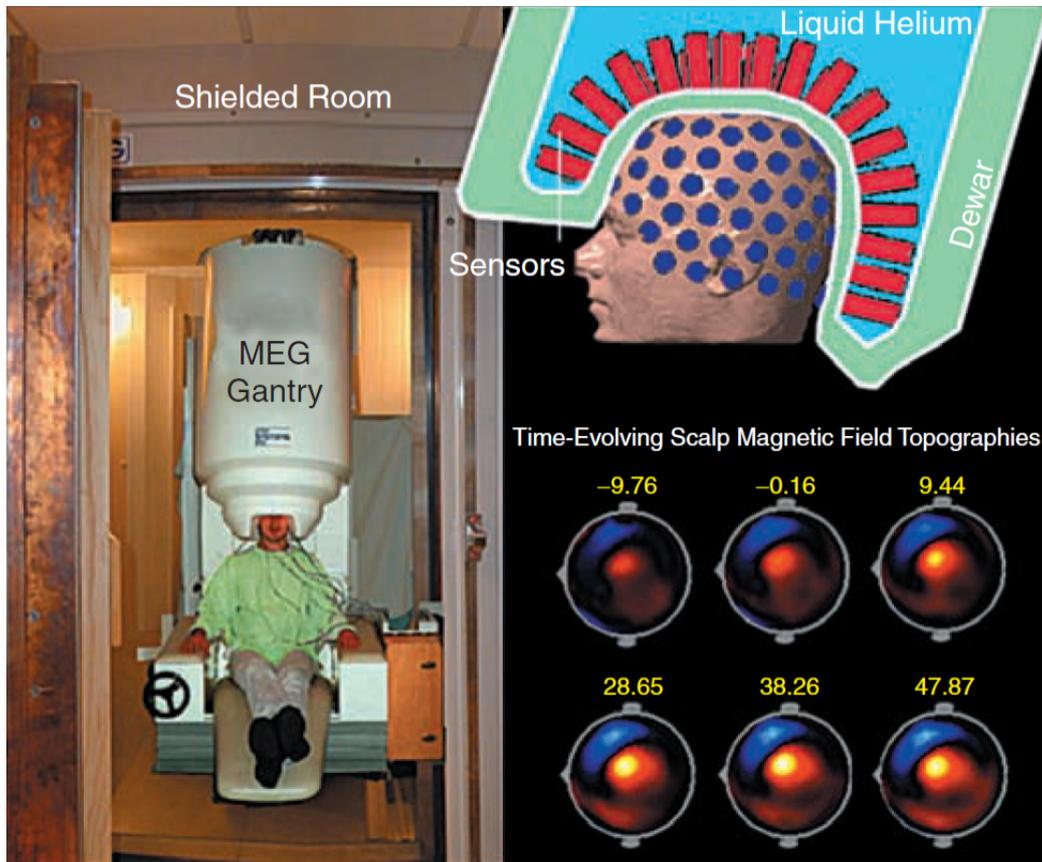


Figure 3: The MEG measurement device is presented inside a shielded room (left) with illustration of the sensors inside the dewar (top). Typical MEG signals are illustrated in the bottom right corner of the figure. Figure reprinted from Baillet et al. (2001).

The neural sources are modeled using current dipole approximation or extending it to be a multipole of current dipoles. To get the complete source space model describing the distribution of activity in the brain, the distribution of neural sources is needed. There are two common approaches to distribute the sources. In dipole fitting models, a few equal current dipoles are placed over the surface with free orientations and amplitudes, possibly with indeterminate locations. The other, called distributed models, is to distribute the neural sources into a dense mesh over the cerebral cortex with constrained locations. The orientations of the current dipoles may be constrained. The whole set of neural sources forms the source space. (Baillet et al., 2001)

When solving the inverse problem, another considerable property is a head

model which takes into account the layered structure and conductivity profile of the brain (Baillet et al., 2001). Conductivity is a material specific quantity which describes how well the material conducts electricity and thus the conductivity profile affects more on EEG than on MEG (Hari and Puce, 2017). The simplest head model assumes the brain to consist of concentric symmetric spherical shells with uniform conductivity. More accurate method is to extract the surface boundaries from an image of the brain and skull and then apply the boundary element method assuming the homogeneity in conductivity. If the assumption of homogeneity is discarded, the finite element method needs to be applied in the calculations. (Baillet et al., 2001)

With the source space and head models, one can calculate the lead field matrix \mathbf{G} appearing in the equation (1). The lead field matrix is a mapping from the source space to the sensor space. In other words, it tells how each dipole contributes to the obtained data. Using the lead field matrix, one is able to solve the inverse problem. The inverse problem produces an estimate for the set of parameters depending on the chosen source space model. The source space model also determines the method applied to obtain the solution. In principle, two types of approaches, parametric and imaging methods, are utilized. Parametric methods enable the estimation of the orientations and locations of the dipoles in addition to the amplitudes and thus they are commonly applied with dipole fitting models. Examples of parametric methods are least-squares estimation using Frobenius norm, beamformer method and multiple signal classification method. In imaging methods, which are applied with distributed models, the only parameters to be estimated are the amplitudes due to constraints on locations and orientations of the current dipoles. This problem is linear and it can be solved using Bayesian approach. (Baillet et al., 2001)

3.3 Estimating the activity and connectivity between regions of interest

Connectivity describes the relations between distinct regions. It can be classified into neuroanatomical, functional and effective connectivity. Neuroanatomical connectivity means the physical connections between regions and it is commonly measured using techniques like diffusion MRI. Functional and effective connectivity are related to temporal correlations during signal processing and thus require measurement techniques, such as MEG, with higher temporal resolution. Functional connectivity is measured in terms of statistical dependencies in the activity of distant regions. Effective connec-

tivity is more focused on trying to find the generative model behind these dependencies. It refers to the dynamic directed influences that one neural system exerts over another. (Sakkalis, 2011; Schoffelen and Gross, 2009)

A schematic chart of methods to determine the functional connectivity is represented in figure 4. The methods are classified into categories based on domain and characteristics. More information about the methods is provided for example in Sakkalis (2011).

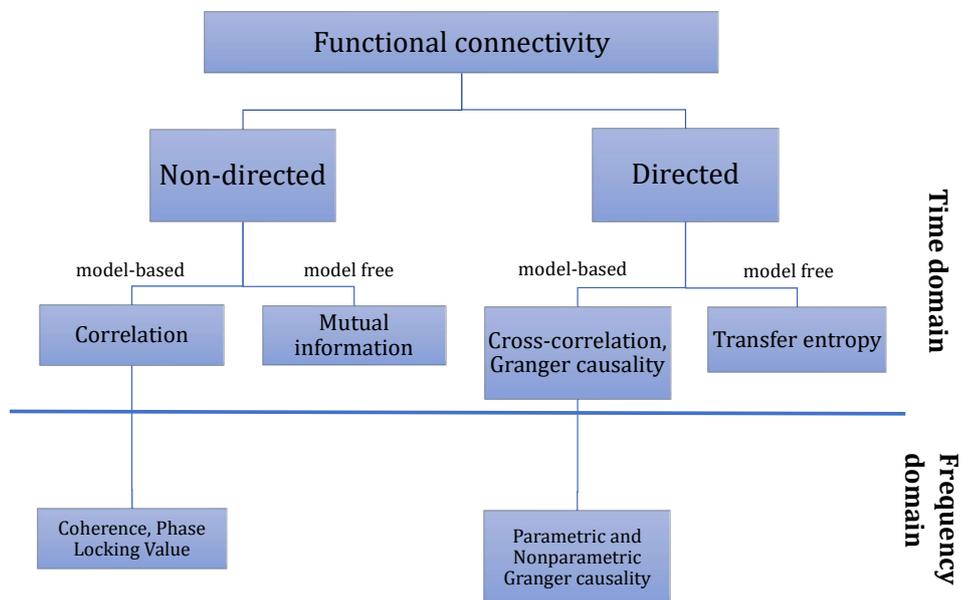


Figure 4: Schematic graph of different methods to estimate functional connectivity.

The estimation of connectivity is in general conducted at the source level. Due to wide-spread representation of the neural fields, estimation of connectivity at sensor level might lead to deceptive interpretation of the connectivity analysis (Schoffelen and Gross, 2009). Spatial leakage also causes the estimates of connectivity to be dependent on each other. In order to obtain independent estimates and reduce the dimensionality of the problem, the estimation is constrained on regions of interest. The regions of interest can be defined based on atlases which are anatomical parcellation maps connecting a region of cerebral cortex to a specific function.

Connectivity estimation methods are in general based on analyzing the time series representing the activity of a region. The region activity can be estimated from the neural source activity for example by taking mean or flipped mean of the source activities within the region or analyzing the principal components. These methods are proven to be convenient in practice and thus applied generally. But as far as we know, there isn't any research on the relevance of these estimation methods. In order to obtain more plausible approach to estimate the region activity and preserve the information present in the data, the aim of this work is to survey the prospects of a data-driven method.

4 Methods

4.1 State space model

State space model can be used to describe a system which has a hidden variable with linear characteristics (Shumway and Stoffer, 2006). It is suitable approach for this work as the neural activity can be modeled as a hidden variable. The model is also capable of modeling systems with uncertainty in both state and measurement dynamics which is quite necessary when considering the physical plausibility of the model. Linearity and Gaussian characteristics of the model variables enable simple way for Bayesian filtering with closed form solutions and lighten up the computational demand (Särkkä, 2013).

The state space model consists of two models first of which illustrates the dynamics of the latent variable. Let the latent variable of the system be $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_t)$. The variable \mathbf{x} is now the state variable which fully describes the state of the whole system. It is assumed to follow the equation

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \boldsymbol{\vartheta}_t, \quad (2)$$

where \mathbf{A} is the evolution matrix and $\boldsymbol{\vartheta} \sim \mathcal{N}(0, \mathbf{S})$ is the process noise. The noise is assumed to be white and Gaussian. Equation (2) is called the dynamic model which states the Markovian and autoregressive properties of the state variable \mathbf{x} . The initial value of \mathbf{x} is assumed to follow some normal distribution $\mathbf{x}_0 \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. (Shumway and Stoffer, 2006)

The second model in the state space model determines the relation between the state variable \mathbf{x} and the observed variable \mathbf{y} . This relation is described

through equation

$$\mathbf{y}_t = \mathbf{Z}\mathbf{x}_t + \boldsymbol{\varepsilon}, \quad (3)$$

where \mathbf{Z} is the matrix defining the deterministic linear mapping from the state space to the measurement space and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{Q})$ is the measurement noise which is also assumed to be white and Gaussian. Equation (3) represent the measurement model. (Shumway and Stoffer, 2006)

Gathering the equations together, the state space model is

$$\begin{aligned} \mathbf{x}_t &= \mathbf{A}\mathbf{x}_{t-1} + \boldsymbol{\vartheta}_t, \boldsymbol{\vartheta} \sim \mathcal{N}(0, \mathbf{S}), \\ \mathbf{y}_t &= \mathbf{Z}\mathbf{x}_t + \boldsymbol{\varepsilon}_t, \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{Q}), \\ \mathbf{x}_0 &\sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \end{aligned}$$

4.2 Expectation Maximization algorithm

Expectation maximization algorithm is an iterative method to obtain maximum likelihood estimates from incomplete data formally developed by Dempster et al. (1977). It has gained popularity since then due to its applicability in a wide range of fields. The incompleteness can mean for example that some observations are missing or the model includes variables whose values have not been obtained (Schön, 2009). These variables are called hidden or latent variables (Chen and Gupta, 2010). The applicability in a case of incomplete data makes the expectation maximization suitable for this work.

In the set up of expectation maximization algorithm, there are the observed data $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_T)$ and some latent variable $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T)$ both of which have some parametric densities $p(\mathbf{y}|\boldsymbol{\theta})$ and $p(\mathbf{x}|\boldsymbol{\theta})$. The parametric density functions can be considered as likelihood functions describing the plausibility of the parameter $\boldsymbol{\theta}$. The observed data \mathbf{y} is a realization of the latent variable \mathbf{x} through some deterministic function. To make the obtained data as likely as possible, we wish to find the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ by maximizing the likelihood function $p(\mathbf{y}|\boldsymbol{\theta})$. For convenience, log of the likelihood function $\ln p(\mathbf{y}|\boldsymbol{\theta})$ can be considered. Maximization of the log-likelihood function might be a difficult problem so another closely related maximization problem is considered. Assuming that necessary information about the latent variable \mathbf{x} is available, it can be shown that maximizing the joint log-likelihood function $\ln p(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta})$ is equivalent to maximizing $\ln p(\mathbf{y}|\boldsymbol{\theta})$. Because the latent variable \mathbf{x} is actually unknown, expectation of the joint

log-likelihood function conditioned to the current information has to be considered. Current information is considered to be the obtained data \mathbf{y} and previous guess for the parameter value $\boldsymbol{\theta}_k$. (Chen and Gupta, 2010; Schön, 2009)

To formulate the expression for expectation of the joint log-likelihood function as an integral, the distribution of \mathbf{x} conditioned to the current information $p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}_k)$ needs to be constructed. This is called the E-step in the expectation maximization framework. Using the expression obtained in E-step and integrating over possible values of \mathbf{x} , the expectation of the joint log-likelihood is

$$\mathbb{E}_{\mathbf{x}}[\ln p(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta})|\mathbf{y}, \boldsymbol{\theta}_k] = \int \ln p(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta})p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}_k)d\mathbf{x} . \quad (4)$$

This equation is the objective function of the optimization problem. Maximizing the equation will give the next estimate for the parameter values

$$\hat{\boldsymbol{\theta}}_{k+1} = \arg \max_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{x}}[\ln p(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta})|\mathbf{y}, \boldsymbol{\theta}_k] .$$

Performing maximization is called the M-step. (Chen and Gupta, 2010)

The two steps, E- and M-step, described above are repeated until the solution convergences. Convergence can be qualified by calculating for example changes in the estimates or in the objective function (Schön, 2009). In the beginning of the algorithm, one needs to pick an initial guess for the values of the parameters to be estimated. This will possibly affect the obtained solution because the expectation maximization algorithm always converges but there is no guarantee that the solution is a global maximum. Hence, it is recommended to use multiple initial guesses for the parameters to attain an estimate close to the global optimum. (Chen and Gupta, 2010)

4.3 Kalman Filter

To obtain the expectation of the joint log-likelihood in (4) the estimate of the complete data \mathbf{x} is needed. One way to produce the estimate is to use filtering methods which can be applied to estimate a time dependent, indirectly observed systems. Optimal filtering is a statistical inversion problem where the unknown state of the system is defined from the given measurements in a statistically optimal way. In practice, this means calculating the posterior distribution of all the states given all the measurements. As the number of measurement increases, the calculation becomes intractable. The

idea behind filtering is to consider some marginal distribution instead of the complete distribution. The marginal distributions considered are called prediction, filtering and smoothing distributions. Prediction distribution is the distribution of the future state given the current measurement whereas filtering distribution is the distribution of the current state given the current measurement. Smoothing distribution describes the distribution of some state given measurements on a certain interval which extends also to the future. (Särkkä, 2013)

Deciding to use a state space model, a prior information about the system can be included in the filtering procedure meaning a Bayesian filter can be applied. Due to linear and Gaussian characteristics presented in equations (2) and (3), a convenient means to implementing a Bayesian filter is a Kalman filter developed by Kalman (1960). The Kalman filter provides a recursive method with closed form solutions to the Bayesian filtering problem where the system dynamics are discrete time, linear and Gaussian (Särkkä, 2013).

4.4 Formulation of the state space model and likelihood function

The state space model for the brain activity and its dynamics is formulated as follows. In the following equations, the activity of the regions of interest is denoted with $\mathbf{u} = \mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n$, which is chosen to be the state variable in the model. The measured quantity in the model will be the obtained data denoted with $\mathbf{y} = \mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n$. The activity of the neural sources is denoted as $\mathbf{q} = \mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_n$. Each of the presented variables is discretely time dependent denoted by the subscript where n is defined to be the number of time points in the measurement. Hence, the column vectors \mathbf{u}_t , \mathbf{q}_t and \mathbf{y}_t have the dimensions $p \times 1$, $q \times 1$ and $s \times 1$.

The dynamics of the region activity are described as a probabilistic Markov sequence through equation (Yang et al., 2016)

$$\mathbf{u}_t = \mathbf{A}\mathbf{u}_{t-1} + \boldsymbol{\vartheta}_t, \boldsymbol{\vartheta}_t \sim \mathcal{N}(0, \mathbf{S}), \quad (5)$$

where \mathbf{A} is a $p \times p$ matrix describing the dependency of \mathbf{u}_t and \mathbf{u}_{t-1} , and $\boldsymbol{\vartheta}_t$ is a $p \times 1$ column vector modeling the uncertainty drawn from a Gaussian distribution with zero mean and \mathbf{S} as a $p \times p$ covariance matrix. The initial probability distribution for \mathbf{u}_0 is assumed to be Gaussian: $\mathbf{u}_0 \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. (Särkkä, 2013)

To construct the measurement model, the relation between the region activity and dipole activity is determined with equation

$$\mathbf{q}_t = \mathbf{L}\mathbf{u}_t + \boldsymbol{\xi}_t, \quad \boldsymbol{\xi}_t \sim \mathcal{N}(0, \mathbf{R}), \quad (6)$$

where \mathbf{L} is the $q \times p$ matrix containing weights for individual source points within each region, $\boldsymbol{\xi}_t$ is the $q \times 1$ column vector for the uncertainty in the activity of the source points, and \mathbf{R} its $q \times q$ covariance matrix. The source points are assumed to be independent and thus the covariance matrix \mathbf{R} is a diagonal matrix. It is constructed so that a variance σ^2 for the dipole activity is specified for each region of interest and source points outside all the regions. Thus each entry $R_{i,i}$ is determined by where the i th source point belongs to.

Combining (6) with the equation (1) introduced in the inverse problem results in measurement equation of the form

$$\mathbf{y}_t = \mathbf{G}\mathbf{L}\mathbf{u}_t + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \sim \mathcal{N}(0, \boldsymbol{\Omega}), \quad (7)$$

where \mathbf{G} is the $s \times q$ lead field matrix and the $s \times 1$ column vector $\boldsymbol{\eta}_t$ is equal to $\mathbf{G}\boldsymbol{\xi}_t + \boldsymbol{\varepsilon}_t$ with $s \times s$ covariance matrix $\boldsymbol{\Omega} = \mathbf{G}\mathbf{R}\mathbf{G}^\top + \mathbf{Q}$. In previous definition, the $\boldsymbol{\varepsilon}_t$ is the measurement noise which is also a $s \times 1$ column vector with \mathbf{Q} as $s \times 1$ covariance matrix.

Equations (7) and (5) with prior probability distribution for \mathbf{u}_0 construct the state space model:

$$\begin{aligned} \mathbf{y}_t &= \mathbf{G}\mathbf{L}\mathbf{u}_t + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \sim \mathcal{N}(0, \boldsymbol{\Omega}), \\ \mathbf{u}_t &= \mathbf{A}\mathbf{u}_{t-1} + \boldsymbol{\vartheta}_t, \quad \boldsymbol{\vartheta}_t \sim \mathcal{N}(0, \mathbf{S}), \\ \mathbf{u}_0 &\sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \end{aligned} \quad (8)$$

The primary aim of this work is to survey the estimation of the weight matrix \mathbf{L} from the given data. Instead of assigning for example equal weights for the dipoles in a region of interest, the weights would be adjusted so that the dipoles which have the highest activity would have a larger weight coefficients than other dipoles. Based on the processing of a neural signal, the weights are assumed to have Gaussian spread around the dipole with higher activity. Hence, dipoles close the center of activity would have more effect on the activity of the region than dipoles further away from the centre. To incorporate this prior of spatial smoothness on \mathbf{L} , an additional penalty term is included. The columns of the weight matrix \mathbf{L} are assumed to follow a multivariate Gaussian distribution with zero mean and $\mathbf{Q}_{\text{pen}}^m$ as covariance

matrix where $m \in 1, \dots, p$. The covariance matrix $\mathbf{Q}_{\text{pen}}^m$ is constructed for each region separately according to equation

$$(\mathbf{Q}_{\text{pen}}^m)_{i,j} = \phi \mathbf{n}_i^\top \mathbf{n}_j \exp(-\lambda \|\mathbf{r}_i - \mathbf{r}_j\|_2), \quad i, j \in 1, \dots, \text{dipoles}$$

where \mathbf{n}_k is the normal vector, \mathbf{r}_k is the position vector of the k th dipole in the parcel, and $\phi, \lambda \in \mathbb{R}_+$ are hyperparameters.

Using the constructed state space model and the formulated penalty matrices, one is able to derive the joint log-likelihood function of the activity of the regions of interest \mathbf{u} and the observed data \mathbf{y} . The complete derivation is presented in appendix A.

For simplicity, the only parameter to be estimated is the weight matrix \mathbf{L} . Thus, taking the expectation of the derived joint log-likelihood conditioned to the obtained data and previous estimates for the parameters, and moving the sums and expectation inside the trace results in

$$\begin{aligned} \mathbb{E}[\ln p(\mathbf{u}, \mathbf{y}) | \mathbf{Y}, \mathbf{L}_{k-1}] &= \Xi_{k-1} - \frac{1}{2} \text{Tr}\{\Sigma^{-1}(\hat{\mathbf{u}}_{0|n} \hat{\mathbf{u}}_{0|n}^\top \\ &\quad - \hat{\mathbf{u}}_{0|n} \boldsymbol{\mu}^\top - \boldsymbol{\mu} \hat{\mathbf{u}}_{0|n}^\top + \boldsymbol{\mu} \boldsymbol{\mu}^\top)\} \\ &\quad - \frac{1}{2} \text{Tr}\{\mathbf{S}^{-1} \sum_{t=1}^n (\boldsymbol{\Phi}_{t,t} - \boldsymbol{\Phi}_{t,t-1} \mathbf{A}^\top \\ &\quad - \mathbf{A} \boldsymbol{\Phi}_{t,t-1}^\top + \mathbf{A} \boldsymbol{\Phi}_{t-1,t-1} \mathbf{A}^\top)\} \tag{9} \\ &\quad - \frac{1}{2} \text{Tr}\{\boldsymbol{\Omega}^{-1} \sum_{t=1}^n (\mathbf{y}_t \mathbf{y}_t^\top - \mathbf{y}_t \hat{\mathbf{u}}_{t|n}^\top \mathbf{L}_{k-1}^\top \mathbf{G}^\top \\ &\quad - \mathbf{G} \mathbf{L}_{k-1} \hat{\mathbf{u}}_{t|n} \mathbf{y}_t^\top + \mathbf{G} \mathbf{L}_{k-1} \boldsymbol{\Phi}_{t,t} \mathbf{L}_{k-1}^\top \mathbf{G}^\top)\} \\ &\quad - \frac{1}{2} \sum_{m=1}^p \text{Tr}\{(\mathbf{Q}_{\text{pen}}^m)^{-1} (\mathbf{L}_{*,m})_{k-1} (\mathbf{L}_{*,m}^\top)_{k-1}\}, \end{aligned}$$

where

$$\Xi_{k-1} = -\frac{1}{2} (\ln[\det 2\pi \boldsymbol{\Sigma}] + \text{T} \ln[\det 2\pi \mathbf{S}] + \text{T} \ln[\det 2\pi \boldsymbol{\Omega}_{k-1}] + \sum_{m=1}^p \ln[\det 2\pi \mathbf{Q}_{\text{pen}}^m]),$$

and abbreviations for the expectations are,

$$\begin{aligned}
\hat{\mathbf{u}}_{t|n} &= \mathbb{E}[\mathbf{u}_t | \mathbf{y}_n, \boldsymbol{\theta}_{k-1}], \\
\mathbf{P}_{t,s|n} &= \mathbb{E}[(\mathbf{u}_t - \hat{\mathbf{u}}_{t|n})(\mathbf{u}_s - \hat{\mathbf{u}}_{s|n})^\top | \mathbf{y}_n, \boldsymbol{\theta}_{k-1}], \\
\boldsymbol{\Phi}_{t,t} &= \hat{\mathbf{u}}_{t|n} \hat{\mathbf{u}}_{t|n}^\top + \mathbf{P}_{t|n}, \\
\boldsymbol{\Phi}_{t,t-1} &= \hat{\mathbf{u}}_{t|n} \hat{\mathbf{u}}_{t-1|n} + \mathbf{P}_{t,t-1|n}, \\
\boldsymbol{\Phi}_{t-1,t-1} &= \hat{\mathbf{u}}_{t-1|n} \hat{\mathbf{u}}_{t-1|n} + \mathbf{P}_{t-1|n}.
\end{aligned}$$

The expectations $\hat{\mathbf{u}}_{t|n}$ and $\mathbf{P}_{t,s|n}$ are obtained using Kalman smoother. The Kalman prediction, filtering and smoothing equations are presented in appendix B.

The equation (9) is the objective function of the maximization problem. Instead of maximizing the obtained equation, the related problem of minimizing the negative of the objective function is considered because the optimization problems are equivalent. Hence, the objective function is $\mathcal{L}_{\mathbf{L}}(\mathbf{u}, \mathbf{y}) = -2 \mathbb{E}[\ln p(\mathbf{u}, \mathbf{y} | \mathbf{Y}, \mathbf{L}_{k-1})]$. The objective function is nonlinear and nonconvex.

The minimization problem is unconstrained, so gradient descent method is applied. In gradient descent method, a minimizing sequence of parameter values is formed by moving some amount defined by an adjustable value called step size into the direction of the negative gradient which points the direction of steepest descent of the objective function (Boyd and Vanderberghe, 2004). To construct the update equation for \mathbf{L} , the negative gradient of the objective function is calculated with respect to \mathbf{L} using Petersen and Pedersen (2012),

$$\begin{aligned}
\frac{\partial \mathcal{L}_{\mathbf{L}}(\mathbf{u}, \mathbf{y})}{\partial \mathbf{L}} &= \sum_{t=1}^n (-\mathbf{G}^\top \boldsymbol{\Omega}^{-1} \mathbf{y}_t \hat{\mathbf{u}}_{t|n}^\top - \mathbf{G}^\top (\boldsymbol{\Omega}^{-1})^\top \mathbf{y}_t \hat{\mathbf{u}}_{t|n}^\top \\
&\quad + \mathbf{G}^\top (\boldsymbol{\Omega}^{-1})^\top \mathbf{G} \mathbf{L}_{k-1} (\mathbf{P}_{t|n} + \hat{\mathbf{u}}_{t|n} \hat{\mathbf{u}}_{t|n}^\top) + \mathbf{G}^\top \boldsymbol{\Omega}^{-1} \mathbf{G} \mathbf{L}_{k-1} (\mathbf{P}_{t|n} + \hat{\mathbf{u}}_{t|n} \hat{\mathbf{u}}_{t|n}^\top)), \\
&\quad + \sum_{i=1}^p 2(\mathbf{Q}_{\text{pen}}^{\mathbf{m}})^{-1} \mathbf{L}_{*,\mathbf{m}_{k-1}}
\end{aligned}$$

that should be equal to zero. The equation doesn't have a closed form solution for \mathbf{L} so backtracking line search is implemented with gradient descent method. Backtracking line search is an inexact line search method in which the step size is chosen so that it minimizes the objective function for a certain, predefined amount. (Boyd and Vanderberghe, 2004)

5 Results

The expectation maximization algorithm with minimization procedure described in the previous section is implemented using MATLAB R2017a. The code can be found in <https://github.com/HilkkaH/Estimating-the-activity-of-regions-of-interest-using-EM-algorithm->. A sample data set for left auditory stimulus and the minimum norm solution for the inverse problem from MNE-library for Python are exploited in both simulations and experiments with real data.

Simulations are used to test the validity and performance of the algorithm. The data for simulations are generated using the lead field matrix \mathbf{G} from the sample data set and identity matrix as the evolution matrix \mathbf{A} . The region activity \mathbf{u} is initialized as a multivariate normal distribution with zero mean and identity matrix as covariance. The dipole and measurement noise covariance matrices are adjusted so that the amount of noise is relatively small. The penalty matrices are constructed using values 10 and 1 for hyperparameters ϕ and λ , respectively.

The weight matrix \mathbf{L} is designed to have a Gaussian amplitude distribution and is generated in the following manner: The dipole closest to the geometric center of the region is assigned to be the most active point within the region and thus having the highest weight. Other weights are drawn as normalized probabilities from a normal distribution with zero mean and some variance. The variance is used to describe the spread of activity around the most active dipole. Later, this variance is referred as radius. Visualization of \mathbf{L} with variance of 0.02 for the superior temporal area on the left hemisphere is shown in figure 5.

To take into account the influence of dipole orientation, \mathbf{L} is multiplied element-wise with a matrix containing values $\{-1, 1\}$ indicating the orientation of the dipole.

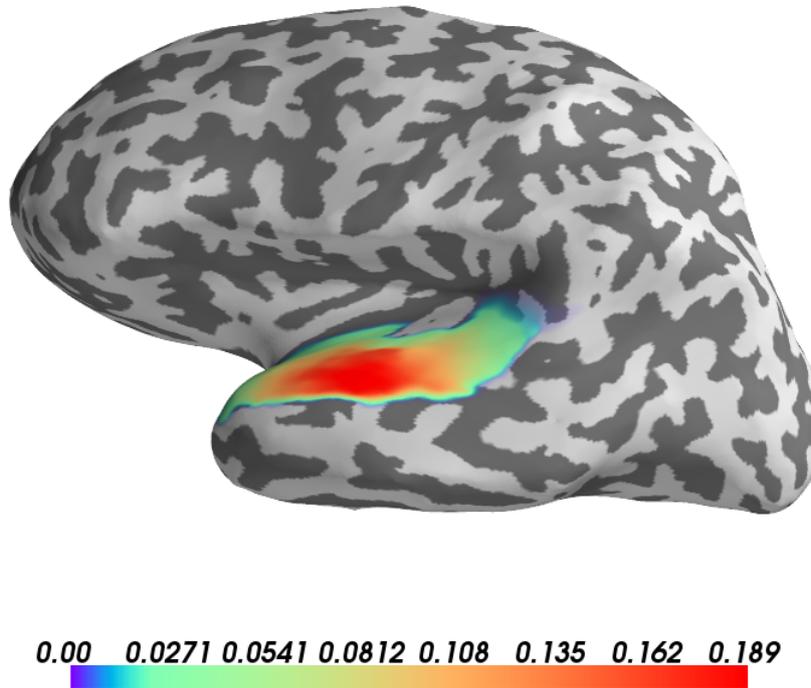


Figure 5: Visualization of the weight matrix \mathbf{L} on the surface of the brain for superiortemporal area on the left hemisphere.

The accuracy of the algorithm is tested in two different manners. At first, the estimation is initialized with the true values of \mathbf{L} to obtain a possible lower bound for the estimation error and secondly, the estimation is initialized with \mathbf{L} drawn from standard normal distribution to attain a limit for the estimation error. The tests are performed for seven different radii between 0.07 – 0.01 using a hundred randomly drawn signals. The expectation maximization algorithm is let to run one hundred iterations with constrained orientation and the estimation errors of \mathbf{L} and \mathbf{u} are measured. The means and standard deviations of the relative estimation errors for the superiortemporal areas as a function of the radius are presented in figure 6.

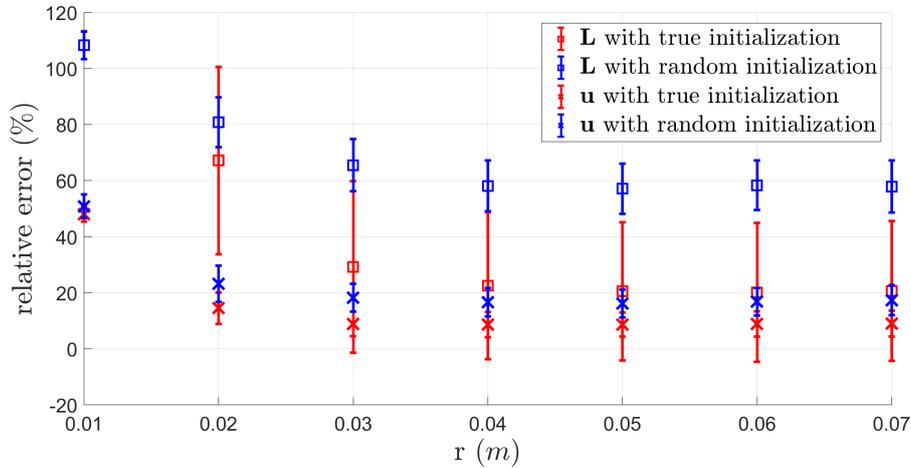


Figure 6: The relative estimation errors of \mathbf{L} and \mathbf{u} with two different initialization manners.

Overall, the relative estimations errors of \mathbf{L} are high. The minimum error is around 30 % in the largest radius of 0.07 but increases to be over 100 % in the smallest radius of 0.01. This implies that the model is possibly invalid or not accurate enough to capture really localized activity. The chosen estimation method might also have an effect on the obtained errors especially on smaller radius when random initialization gives more accurate result than initializing with the true values. The difference in the obtained estimation errors with two different initialization methods shows that the algorithm is dependent on the initial values set for the estimated parameters. The deviation in the errors implies that the shape of the estimated signal and thus the amount of noise has an effect on the estimation result. Though the relative estimation errors of \mathbf{L} are high, the relative estimation errors of the signal \mathbf{u} are relatively small with smaller variances. The relative error of \mathbf{u} behaves similarly with the relative error of \mathbf{L} starting from error of around 10 % and increasing up to 60 %. The smaller errors are possibly due to the robustness of the Kalman filter, but it also implies that different combinations of the weights can present the same signal with decent accuracy. In conclusion, the implemented algorithm works tolerably when the radius is greater representing an activation of a larger patch on the cortex. Unfortunately, the greater radii with values between 0.03–0.07 are quite large in the scale of brain.

To survey further the possibilities of determining the weights using data driven method, maximum likelihood estimation for the parameters is conducted on real data. For the initial guess of \mathbf{L} , the dipole closest to the

geometric center of the area is chosen to be most active point and the variance is chose to be 0.05. The dipole noise matrix \mathbf{R} was determined to be zero. Before applying the algorithm, the data is whitened. The region activity \mathbf{u} is calculated as a minimum norm solution to equation (6) based on the estimated \mathbf{L} . The obtained signals for superiortemporal areas on left and right hemispheres are presented together with signals obtained using mean, flipped mean and PCA in figure 7.

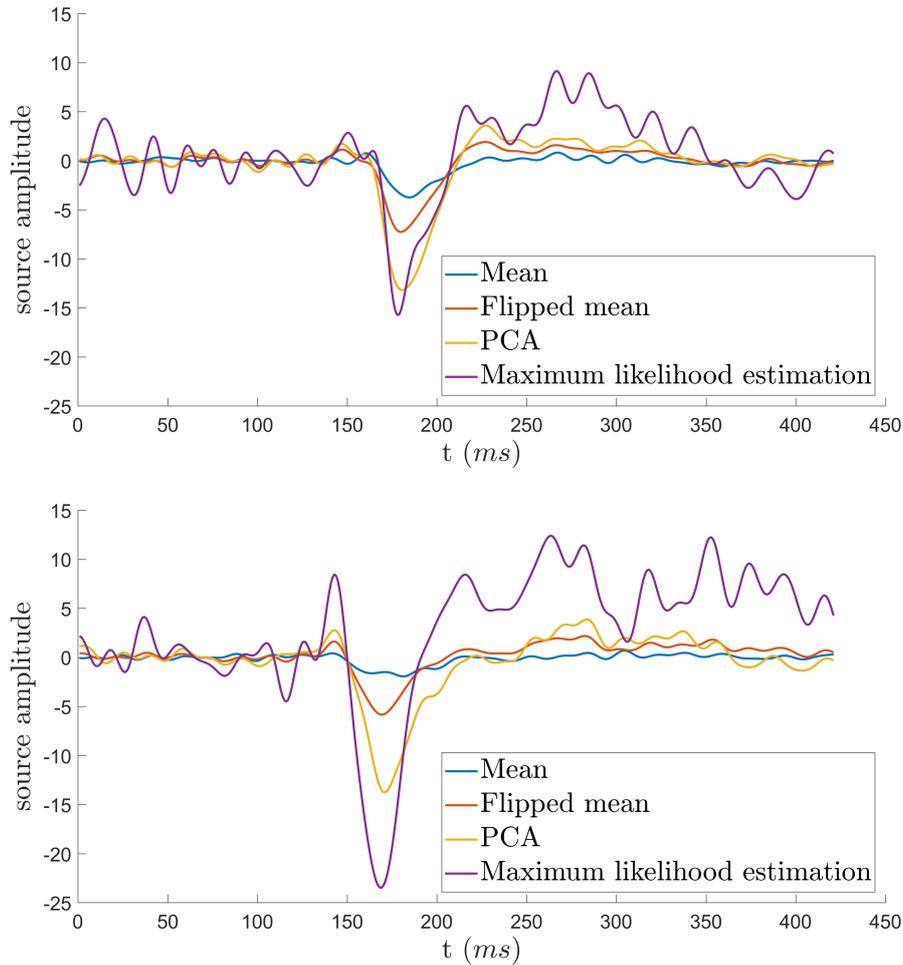


Figure 7: The minimum norm estimate for the region activity calculated using maximum likelihood estimation compared to other estimates for region activity obtained with different methods for the superiortemporal area on left (top) and right (bottom) hemisphere.

The maximum likelihood estimation produces good results for both hemi-

spheres since the obtained shape of the estimated signal is in accordance with other signals. The signal representing the activity of the superior temporal area on the right hemisphere has higher peak than the activity of the left superior temporal area due to the left stimulus used in the measurement. The activity of the right superior temporal area has also significantly higher peak compared to obtained activities with other methods indicating the maximum likelihood estimation is able to capture better the activation of underlying neural sources. However, the calculated signals are not really smooth and have quite a lot of fluctuations before and after the peak.

6 Conclusions

The activity of the brain can be recorded using noninvasive measurement techniques such as EEG or MEG which measure the electromagnetic fields generated by neurons. From the collected data, the underlying neural activity can be estimated by solving an inverse problem. The region activity can be constructed based on the estimated neural activity using various methods. Only validation for these developed methods is their convenience in practice and thus in this thesis, the possibility implement a data driven method to estimate the region activity is surveyed.

The method examined estimates the region activity using a weight matrix which describes how much each vertex contributes to the activity of the region. The weight matrix is obtained as maximum likelihood estimate from the data using expectation maximization algorithm with a state space model. The signal representing the region activity is obtained via Kalman filter.

The performance of the implemented algorithm was first tested using simulations. Though the calculated relative estimation errors of the weight matrix were overall quite high, the underlying signal could be decently recovered with smaller relative estimation error. The obtained errors were approximately proportional to the inverse of the radius representing the spread of activity. Thus the algorithm can provide any valid information only on larger values of the radius. To test the possibility to estimate the weights from the data, a simple maximum likelihood estimation algorithm is implemented on real data. The minimum norm estimate for the region activity produced promising results when considering the shape and amplitude of the signal.

The major defect in the algorithm is the inability to capture the underlying activity accurately. One prospect of improving the performance of the algo-

rithm could be developing a different model which could describe the localized activity better. The assumptions made in the state space model are valid but the significant amount of noise in the problem hinders the detectability of the activity. Though the state space model fulfilled the requirements of modeling these different process noises, the projections of noises made the algorithm unstable causing severe estimation errors. Hence eventually, the dipole noise had to be left out from the model. The modeling and estimation of different process noises would require more subtle methods.

The amount of estimable weights per region that can be estimated is over a hundred, on average. This amount is quite high and might complicate the estimation. To cover this problem the amount of parameters should be reduced or find a way to include additional information. The additional information currently included in the algorithm is the designed penalty matrices. The validity of this design isn't tested and it can be unsuitable. The covariance could be designed using some other approach, for example the Matérn class of kernels.

The estimation was conducted using gradient descent method with backtracking line search. The gradient descent has in general proven to be a robust estimation method but the complexity of the objective function might cause an issue. It is highly nonlinear with respect to the estimated weights and the curvature of the function is unknown. The objective function might be convex only locally and thus have multiple local minima. Algorithm can only detect a local minimum and even multiple initializations might not lead to a global minimum. So more advanced optimization algorithm designed for nonlinear functions would improve the accuracy of the algorithm.

Based on the results from maximum likelihood estimation, carefully designed and implemented data driven method would provide more accurate approach for estimation of the region activity. This could improve the quantification of functional integration and hence provide more detailed information about the functional principles of the brain.

References

- Sylvain Baillet, John C. Mosher, and Richard M. Leahy. Electromagnetic Brain Mapping. *IEEE Signal Processing Magazine*, 18(6):14–30, November 2001.
- Stephen Boyd and Lione Vanderberghe. *Convex Optimization*, chapter 9, pages 455–483. Cambridge University Press, Cambridge, United Kingdom, 2004.
- György Buzsáki, Costas A Anastassiou, and Christof Koch. The origin of extracellular fields and currents — EEG, ECoG, LFP and spikes. *Nature reviews neuroscience*, 13(6):407, 2012.
- Yihua Chen and Maya R. Gupta. EM Demystified: An Expectation-Maximization Tutorial. Technical report, Department of Electrical Engineering, University of Washington, Seattle, Washington, USA, February 2010.
- A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, 39:1–38, 1977.
- Riitta Hari and Aina Puce. *MEG-EEG Primer*, chapter 1–6, pages 3–87. Oxford University Press, 2017.
- Matti Hämäläinen, Riitta Hari, Risto J. Ilmoniemi, Jukka Knuutila, and Lounasmaa Olli V. Magnetoencephalography theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of Modern Physics*, 65(2):413–497, 1993.
- R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.
- K. B. Petersen and M. S. Pedersen. The Matrix Cookbook, November 2012. URL <http://www2.imm.dtu.dk/pubdb/p.php?3274>. Version 20121115.
- V. Sakkalis. Review of advanced techniques for the estimation of brain connectivity measured with EEG/MEG. *Computers in Biology and Medicine*, 41:1110–1117, 2011.
- Thomas B. Schön. An Explanation of the Expectation Maximization Algorithm. Technical report, Department of Electrical Engineering, Linköpings Universitet, Linköping, Sweden, August 2009.

- Jan-Mathijs Schoffelen and Joachim Gross. Source connectivity analysis with MEG and EEG. *Human Brain Mapping*, 30(6):1857–1865, June 2009.
- Robert H. Shumway and David S. Stoffer. *Time Series Analysis and Its Applications with R Examples*, chapter 6, pages 324–404. Springer Science+Business Media, United States, 2. edition, 2006.
- Simo Särkkä. *Bayesian Filtering and Smoothing*, chapter 1–4. Cambridge University Press, 2013.
- Francois Tadel, Elizabeth Bock, John C. Mosher, Richard M. Leahy, and Sylvain Baillet. Tutorial 20: Head Model, 2016. URL <https://neuroimage.usc.edu/brainstorm/Tutorials/HeadModel>.
- Gerard J. Tortora and Bryan H. Derrickson. *Principles Of Anatomy and Physiology*, chapter 12, 14, pages 353–389, 417–444. John Wiley, 15 edition, 2017.
- Ying Yang, Elissa Aminoff, Michael Tarr, and Kass E Robert. A state-space model of cross-region dynamic connectivity in MEG/EEG. In *Advances in neural information processing systems*, pages 1234–1242, 2016.

A Derivation of the log-likelihood function

The probability function for obtaining \mathbf{u} and \mathbf{y} can be written in the form

$$p(\mathbf{u}, \mathbf{y}) = p(\mathbf{y}|\mathbf{u})p(\mathbf{u}) \quad (10)$$

using Bayes' rule. Probabilities in the equation above can be expressed as

$$\begin{aligned} p(\mathbf{y}|\mathbf{u}) &= \prod_{t=1}^n p(\mathbf{y}_t|\mathbf{u}_t), \\ p(\mathbf{u}) &= p(\mathbf{u}_0) \prod_{t=1}^n p(\mathbf{u}_t|\mathbf{u}_{t-1}), \end{aligned} \quad (11)$$

where n is the number of time points. Probability functions for $p(\mathbf{u}_t|\mathbf{u}_{t-1})$ and $p(\mathbf{y}_t|\mathbf{u}_t)$ can be obtained using the state space model (8):

$$\begin{aligned} p(\mathbf{y}_t|\mathbf{u}_t) &= p(\mathbf{y}_t - \mathbf{G}\mathbf{L}\mathbf{u}_t) \sim \mathcal{N}(0, \mathbf{\Omega}), \\ p(\mathbf{u}_t|\mathbf{u}_{t-1}) &= p(\mathbf{u}_t - \mathbf{A}\mathbf{u}_{t-1}) \sim \mathcal{N}(0, \mathbf{S}). \end{aligned} \quad (12)$$

Combining equations (10), (11) and (12) gives

$$p(\mathbf{u}, \mathbf{y}) = p(\mathbf{u}_0) \prod_{t=1}^n p(\mathbf{u}_t|\mathbf{u}_{t-1}) \prod_{t=1}^n p(\mathbf{y}_t|\mathbf{u}_t),$$

and taking the natural logarithm results in

$$\ln p(\mathbf{u}, \mathbf{y}) = \ln p(\mathbf{u}_0) + \ln \sum_{t=1}^n p(\mathbf{u}_t|\mathbf{u}_{t-1}) + \ln \sum_{t=1}^n p(\mathbf{y}_t|\mathbf{u}_t). \quad (13)$$

If variable \mathbf{x} follows a multivariate normal distribution $\mathcal{N}(\boldsymbol{\omega}, \mathbf{\Gamma})$, the probability density function is given by

$$p(\mathbf{x}) = (\det 2\pi\mathbf{\Gamma})^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\omega})^\top \mathbf{\Gamma}^{-1}(\mathbf{x} - \boldsymbol{\omega})\right).$$

Using this definition with (12), including also the penalty term for \mathbf{L} , taking natural logarithm of the distributions and plugging these into (13) gives the

log-likelihood function for \mathbf{u} and \mathbf{y} :

$$\begin{aligned}
\ln p(\mathbf{u}, \mathbf{y}) = & -\frac{1}{2} \ln[\det 2\pi\boldsymbol{\Sigma}] + (\mathbf{u}_0 - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{u}_0 - \boldsymbol{\mu}) \\
& -\frac{1}{2} \sum_{t=1}^n (\ln[\det 2\pi\mathbf{S}] + (\mathbf{u}_t - \mathbf{A}\mathbf{u}_{t-1})^\top \mathbf{S}^{-1} (\mathbf{u}_t - \mathbf{A}\mathbf{u}_{t-1})) \\
& -\frac{1}{2} \sum_{t=1}^n (\ln[\det 2\pi\boldsymbol{\Omega}] + (\mathbf{y}_t - \mathbf{G}\mathbf{L}\mathbf{u}_t)^\top \boldsymbol{\Omega}^{-1} (\mathbf{y}_t - \mathbf{G}\mathbf{L}\mathbf{u}_t)) \\
& -\frac{1}{2} \sum_{m=1}^p (\ln[\det 2\pi\mathbf{Q}_{\text{pen}}^m] + \mathbf{L}_{*,m}^\top (\mathbf{Q}_{\text{pen}}^m)^{-1} \mathbf{L}_{*,m}).
\end{aligned} \tag{14}$$

In order to derive more simpler form, Ξ is defined to be equal to all constants in (14),

$$\Xi = -\frac{1}{2} (\ln[\det 2\pi\boldsymbol{\Sigma}] + T \ln[\det 2\pi\mathbf{S}] + T \ln[\det 2\pi\boldsymbol{\Omega}] + \sum_{m=1}^p \ln[\det 2\pi\mathbf{Q}_{\text{pen}}^m]).$$

Using the definition above and applying the trace trick

$$\mathbf{x}^\top \mathbf{D} \mathbf{x} = \text{Tr}\{\mathbf{x}^\top \mathbf{D} \mathbf{x}\} = \text{Tr}\{\mathbf{D} \mathbf{x} \mathbf{x}^\top\},$$

equation (14) can be written as,

$$\begin{aligned}
\ln p(\mathbf{u}, \mathbf{y}) = & \Xi - \frac{1}{2} \text{Tr}\{\boldsymbol{\Sigma}^{-1} (\mathbf{u}_0 - \boldsymbol{\mu}) (\mathbf{u}_0 - \boldsymbol{\mu})^\top\} \\
& -\frac{1}{2} \sum_{t=1}^t \text{Tr}\{\mathbf{S}^{-1} (\mathbf{u}_t - \mathbf{A}\mathbf{u}_{t-1}) (\mathbf{u}_t - \mathbf{A}\mathbf{u}_{t-1})^\top\} \\
& -\frac{1}{2} \sum_{t=1}^t \text{Tr}\{\boldsymbol{\Omega}^{-1} (\mathbf{y}_t - \mathbf{G}\mathbf{L}\mathbf{u}_t) (\mathbf{y}_t - \mathbf{G}\mathbf{L}\mathbf{u}_t)^\top\} \\
& -\frac{1}{2} \sum_{m=1}^p \text{Tr}\{(\mathbf{Q}_{\text{pen}}^m)^{-1} \mathbf{L}_{*,m} \mathbf{L}_{*,m}^\top\}.
\end{aligned}$$

B Kalman prediction, filtering and smoothing equations

To calculate the Kalman smoother, the Kalman prediction and filtering have to be calculated first. The equations to calculate the prediction and filtering

are

$$\begin{aligned}\hat{\mathbf{u}}_{t|t-1} &= \mathbf{A}\hat{\mathbf{u}}_{t-1|t-1} \\ \mathbf{P}_{t|t-1} &= \mathbf{A}\mathbf{P}_{t-1|t-1}\mathbf{A}^\top + \mathbf{S} \\ \hat{\mathbf{u}}_{t|t} &= \hat{\mathbf{u}}_{t|t-1} + \mathbf{K}_t(\mathbf{y}_t - \mathbf{G}\mathbf{L}\hat{\mathbf{u}}_{t|t-1}) \\ \mathbf{P}_{t|t} &= (\mathbf{I} - \mathbf{K}_t\mathbf{G}\mathbf{L})\mathbf{P}_{t|t-1},\end{aligned}$$

where

$$\mathbf{K}_t = \mathbf{P}_{t|t-1}\mathbf{L}^\top\mathbf{G}^\top(\mathbf{G}\mathbf{L}\mathbf{P}_{t|t-1}\mathbf{L}^\top\mathbf{G}^\top + \mathbf{\Omega})^{-1}$$

is the Kalman gain. With predicted and filtered values of the signal one can calculate the backward smoother with following equations:

$$\begin{aligned}\hat{\mathbf{u}}_{t-1|n} &= \mathbf{J}_{t-1}^{-1}(\hat{\mathbf{u}}_{t|n} - \hat{\mathbf{u}}_{t-1|t-1}) + \hat{\mathbf{u}}_{t|t-1}, \\ \mathbf{P}_{t-1|n} &= \mathbf{J}_{t-1}^{-1}(\mathbf{P}_{t|n} - \mathbf{P}_{t-1|t-1})(\mathbf{J}_{t-1}^\top)^{-1} + \mathbf{P}_{t|t-1},\end{aligned}$$

where

$$\mathbf{J}_{t-1} = \mathbf{P}_{t-1|t-1}\mathbf{A}^\top(\mathbf{P}_{t|t-1})^{-1}.$$

The Kalman smoother is initialized with $\hat{\mathbf{u}}_{n|n}$ and $\mathbf{P}_{n|n}$, obtained from the Kalman filter. With definition of \mathbf{K}_t one can also calculate the covariance for one time step lag using equation

$$\mathbf{P}_{t-1,t-2|n} = \mathbf{P}_{t-1|t-1}\mathbf{J}_{t-2}^\top + \mathbf{J}_{t-1}(\mathbf{P}_{t,t-1|n} - \mathbf{A}\mathbf{P}_{t-1|t-1}\mathbf{J}_{t-2}^\top),$$

with

$$\mathbf{P}_{n,n-1|n} = (\mathbf{I} - \mathbf{K}_n\mathbf{G}\mathbf{L})\mathbf{A}\mathbf{P}_{n-1|n-1},$$

as initial guess. In all presented equations above, the values used for parameters in $\boldsymbol{\theta}$ at the j th iteration round are the previous estimates for the parameters $\boldsymbol{\theta}_{j-1}$.

Abbreviations and nomenclature

η	Sum of measurement noise and projected dipole noise
μ	Mean of the initial distribution of the state variable
Ω	Covariance of the sum of the measurement noise and projected dipole noise
Σ	Covariance of the initial distribution of the state variable
θ	A vector containing all the parameters
ε	Measurement noise
ϑ	State noise
ξ	Dipole noise
λ	Hyperparameter for Euclidean distance between two dipoles
\mathbb{Z}	Matrix representing the linear mapping from the state to the measurement
\mathbf{A}	Evolution matrix
\mathbf{G}	The lead field matrix
\mathbf{L}	Weight matrix
$\mathbf{Q}_{\text{pen}}^m$	Penalty matrix corresponding to m th region
\mathbf{Q}	Measurement noise covariance
\mathbf{R}	Dipole noise covariance
\mathbf{S}	State noise covariance
\mathbf{u}	Activity of the regions of interest
\mathbf{x}	The state variable, complete data
\mathbf{y}	Measured data, the incomplete data
ϕ	Hyperparameter for orientation of the dipole
EEG	electroencephalography

fMRI functional magnetic resonance imaging

MEG magnetoencephalography

MRI magnetic resonance imaging

p Number of regions

q Number of vertices in the source space

s Number of sensors