Learning brain connectivity from EEG time series

Jitkomut Songsiri
Department of Electrical Engineering
Chulalongkorn University

Research Project Number: ENG-62-007
Chula Engineering Research Support Grant
Faculty of Engineering
Chulalongkorn University
Bangkok
September 27, 2019

Learning brain connectivity from EEG time series

Jitkomut Songsiri, Ph.D.
University of California, Los Angeles

Research Project Number: ENG-62-007
Chula Engineering Research Support Grant
Faculty of Engineering
Chulalongkorn University

Abstract

An estimation of brain dynamical models not only can provide a characteristic of brain dynamics but also lead to a problem of inferring brain networks that explains relationships among brain regions. This research project provides a scheme of discovering a brain connectivity through EEG signals using a Granger concept that is characterized on state-space models. We propose a state-space model for explaining coupled dynamics of the source and EEG signals where EEG is a linear combination of sources according to the characteristics of volume conduction. Our model has a structure that the sparsity pattern of the model output matrix can indicate the position of active and inactive sources. With this assumption, the proposed scheme consists of two main steps: model estimation and model inference to discover brain connectivities. The model estimation consists of performing a subspace identification to obtain a state-space parameter and an active source selection to reduce model complexity. The model inference on brain connectivity relies on the concept of Granger causality (GC) but it requires an additional learning scheme to regard insignificant causalities, which is then proposed to use Gaussian mixture models to cluster between strong and weak causalities. We aim to verify the performance of our method on simulated data sets that represent realistic human brain activities in a fair setting. The ultimate goal of this study is to explore brain networks from real data sets containing EEG signal in a certain condition and discuss the results with previous studies.

Acknowledgment

This research is financially supported by Chula Engineering Research grant. The author would like to thank Nattaporn Plub-in, Poomipat Boonyakitanont, Parinthorn Manomaisaowapak and Anawat Nartkulpat for preparing part of experimental results and contents included in this technical report.

Contents

List of Figures

List of Tables

Abbreviation

Abbreviation	Description
EEG	Electroencephalogram
MEG	Magnetoencephalography
GC	Granger causality
VAR	Vector autoregressive
VARMA	Vector autoregressive moving average
PDC	Partial directed coherence
DTF	Directed transfer function
DARE	Discrete algebraic riccati equation
GMM	Gaussian mixture model
EM	Expectation-Maximization (algorithm)
TP,TN,FP,FN	True positive, True negative, False positive, False negative

1 Introduction

Brain activities can be observed through many modalities, but mostly popular brain data are EEG data and fMRI data. EEG observes brain activities by placing electrodes with a conductive gel on the human scalp. Synchronous activities of neuron groups, called *sources*, cause action potentials that electrodes can detect the currents. The spreading of electrical currents from sources to EEG sensors is complex because different electrical conduction properties of brain environments. As a result, characteristics of EEG data are varied by age, gender and external effects of each person [SC13, MML$^+$04]. EEG has a high temporal resolution where the sampling frequency is in order of several thousand hertz. However spatial resolution in EEG is limited as of now the number of channels is up to order of hundreds. Moreover, the accessibility of EEG is more economical than that of fMRI because EEG hardware costs are significantly lower than those of fMRI hardware. For these reasons, this work preliminarily focuses only EEG signals for exploring human brain function.

In contrast to learning a brain functionality from scalp signals, dynamics of source signals provide more intrinsic interactions among activities inside the brain. Therefore, an approach to estimate source signals from EEG signals is developed by assuming that the source signals are mapped to EEG electrodes through a linear mapping matrix, called a *lead-field matrix*, with additive noise. This problem is known as a *forward problem*, and conversely, a problem of reconstructing source signals from EEG data is called an *inverse problem*. The main goal of this work is to explore a communication in brain networks, called *brain connectivity* which is a relationship between brain region of interests (ROIs) or neuron groups. This relationship can be distinguished into three types by their statistical definitions and interpretations [PS16, Hau12]. The first type is called structural connectivity that can be referred to patterns of anatomical links or a physical wiring connection between neurons observed by diffusion tomography imaging (DT). The second type is called functional connectivity which explains statistical dependencies between remote brain regions explained by correlations, covariances, spectral coherence or phase-locking. The last type is called effective connectivity describing causal interactions of brain regions through a dynamical model [Sak11, PS16]. Examples of effective connectivity can be explained by dynamical causal modeling (DCM), Granger causality, Directed transfer function (DTF), coherence, and partial directed coherence (PDC) [Hau12] whose all require the specification of a model. Granger causality is one of the data-driven widely-used techniques [SBB15] because of its characterization that can be examined on a linear model and will be our main dependence measure used in this research.

In conclusion, we aim to learn a brain connectivity explaining Granger causality pattern among brain regions via EEG signals measured from a certain condition (resting state, task-driven, etc.) A typical approach to achieve this goal consists of two main steps. The first is to propose a dynamical model that explains relationships between source signals and EEG signals where only the latter can be measured from experiment. The model class should be chosen such that i) parameters can be tractably estimated in a practical setting and ii) Granger causality can be inferred consequently once the model is estimated. The second step is to provide a learning scheme of model inference for brain connectivities after the model is trained.

As a practical merit of this study, learning brain connectivity does not only provides us intrinsically insightful information about the brain region interactions but also finds many applications on differentiating conditions of human brain. For example, brain connectivity based-on spectral coherence is used as a biometric classifier for distinguishing a human brain condition [RCV$^+$14]. Various clinical applications on detecting neurological disorders such as epilepsy, Alzheimer or Schizophrenia are described in a survey [Sak11]. These abnormalities are tested based on the use of various connectivity measures including coherence function and Granger causality.

2 Background

2.1 State-space models

Most literature of exploring Granger causality of multivariate time series has relied on the use of vector autoregressive (VAR) models because of its simple Granger causality characterization in model parameters. In this study, we consider a wider class of linear stochastic process in the form of state-space models to explain EEG time series dynamics. We assume that source signals ($x \in \mathbf{R}^m$) is an

output of state-space model whose state variable is $z \in \mathbf{R}^n$ (or what we call a latent), and the EEG signal ($y \in \mathbf{R}^r$) is a linear combination of the source signals, as described in the following equations.

$$z(t+1) = Az(t) + w(t), \tag{1a}$$
$$x(t) = Cz(t) + \eta(t), \tag{1b}$$
$$y(t) = Lx(t) + v(t). \tag{1c}$$

We call $A \in \mathbf{R}^{n\times n}$ the dynamic matrix, $C \in \mathbf{R}^{m\times n}$ an output matrix mapping the latent to source signal, and $L \in \mathbf{R}^{r\times m}$ is the lead-field matrix determined from a head model. The state noise, w, the output noises η, v are zero-mean and assumed to be mutually uncorrelated.

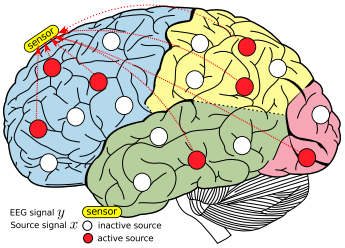


Figure 1: Multivariate EEG signals are modeled as linear combinations of source signals and known as a volume conduction effect.

In EEG applications, the volume conduction explains how the source signal propagates through brain tissues to the EEG signals (here from x to y) and it becomes known that Granger causality learned from y may not be the same pattern as one inferred from x, *i.e.*, spurious effect of Granger causality [dSFK$^+$16]. If model parameters (A, C, L) and noise covariances can be estimated from measurements y then we can consider (1a)-(1b) and conclude a Granger causality in the source signal (x). Such estimation problem can be solved in many ways but we can refer to [PiS18] as a solution. In what follows, we focus on state equations of the source signal only (1a)-(1b) and discuss about how to learn GC of x once all model parameters are estimated.

Moreover, vector autoregressive moving average model (VARMA) is a time series of the form

$$x(t) = \sum_{k=1}^{p} A_k x(t-k) + e(t) + \sum_{k=1}^{q} C_k e(t-k) \tag{2}$$

where A_k's are autoregressive (AR) coefficients and C_k's are moving average (MA) coefficients. VARMA model can be equivalently represented by a state-space equation (1a)-(1b) [CGHJ12]. One of the state-space forms by [Ham94] is given by defining the state variable that contain the lagged values of $x(t)$:

$$z(t) = (x(t), x(t-1), \ldots, x(t-r+1))$$

where $r = \max(p, q+1)$ (the maximum lag order between the AR and MA terms.) Hence, the corresponding state equation is

$$\begin{aligned} z(t) &= \begin{bmatrix} A_1 & A_2 & \cdots & A_r \\ I & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & I & 0 \end{bmatrix} z(t-1) + \begin{bmatrix} I \\ 0 \\ \vdots \\ 0 \end{bmatrix} e(t), \\ y(t) &= \begin{bmatrix} I & C_1 & \cdots & C_{r-1} \end{bmatrix} z(t). \end{aligned} \tag{3}$$

Therefore, VAR process which is a special class of VARMA has a state-space form (3) with $r = p$ but the output equation reduces to

$$y(t) = \begin{bmatrix} I & 0 & \cdots & 0 \end{bmatrix} z(t).$$

2.2 Granger causality (GC)

One topic of interest in time series analysis is to explore relationships among variables. There are many ways to describe statistical relationships between variables in multivariate time series such as correlation function of a random process or coherence function in frequency domain. In this work, we are interested in a concept of exploring *causal* relationships between two variables, x and y, *i.e.*, we determine whether x causes y or y causes x using some statistical definition to explain this causality.

Granger causality is one type of causal relationships. The causality concept is explored by Granger from [Gra69] which analyzes causal relationships between variables in multivariate time series through prediction errors. To this end, assume that a multivariate stationary random processes $x(t)$ consists of $x(t) = (x_1(t), \ldots, x_n(t))$.

Definition 1 *We say x_j is a Granger-cause for x_i, if the variance of the prediction error of x_i conditioning on all components of x is less than variance of the prediction error of x_i conditioning on all components of x except x_j. In other words, x_j helps predict x_i if including x_j in the information set for prediction of x_i yields a decrease in the prediction error [Gra69].*

In order to determine the prediction error of x_i, one can use the best estimator of $x(t)$ in mean squared error (MSE) sense, which is known to be the conditional mean: $\hat{x}(t) = \mathbf{E}[x(t)|y(t-1), \ldots, y(0)]$ where $\{y(\tau)\}_{\tau=0}^{t-1}$ is available information up to time $t-1$. Consequently, the prediction error is obtained from $\varepsilon(t) = x(t) - \hat{x}(t)$ and its covariance is obtained by $\Sigma = \mathbf{E}[\varepsilon(t)\varepsilon(t)^T]$. If we apply Granger causality to learn relationships among variables in a multivariate time series $x(t) = (x_1(t), \ldots, x_n(t))$, then we can consider the prediction of x_i using the past information of other components of $x(t)$, *i.e.*, $x_j(t)$ for $j = 1, \ldots, n$. In order to learn if $x_j(t)$ causes $x_i(t)$ in Granger sense, we distinguish the optimal predictions of $x_i(t)$ into two cases.

- The past information of *all components of* x is included in the prediction. The best prediction of $x_i(t)$ is given (and denoted) by

$$\hat{x}_i(t|t-1) = \mathbf{E}[x_i(t) \mid x_j(t-1), \ldots, x_j(0)], \quad j = 1, \ldots, n.$$

  We use the notation $\hat{x}_i(t|t-1)$ to recognize that the available data for estimating x_i at time t are from the past up to time $t-1$. The prediction of x_i using all the variables is referred to as the **full model**.

- The past information of all components of x *except* x_j is included in the prediction. In this case, the prediction of $x_i(t)$ is denoted by

$$\hat{x}_i^R(t|t-1) = \mathbf{E}[x_i(t) \mid x_k(t-1), \ldots, x_k(0)], \quad \forall k \neq j.$$

  The superscript R denotes that the prediction is obtained from the **reduced model**[1] where we have used *less* information (by excluding x_j from the information set) in order to predict x_i.

From those two optimal predictions, we can assign the corresponding prediction errors and their covariance matrices. The full model has prediction error as $\varepsilon_i(t) = x_i(t) - \hat{x}_i(t|t-1)$ with covariance Σ_{ii}, and the reduced model has the prediction error denoted by $\varepsilon_i^R(t) = x_i(t) - \hat{x}_i^R(t|t-1)$ with covariance Σ_{ii}^R. We have seen that the available data for prediction in the reduced model is less than those of the full model. For this reason, Σ_{ii}^R is always greater than Σ_{ii} or equivalently $\frac{\Sigma_{ii}^R}{\Sigma_{ii}} > 1$ because using more variables in the model provide a better prediction (explained as less prediction error). However, if x_j is indeed has no effect nor can help in the prediction of x_i then including or excluding x_j in the information set has no change in Σ_{ii}^R; that is $\Sigma_{ii}^R = \Sigma_{ii}$. As a result, a measure of Granger causality can be defined as

$$F_{ij} = \log \frac{\Sigma_{ii}^R}{\Sigma_{ii}}$$

to explain if x_j is a Granger cause to x_i. We can say that in general,

$$\frac{\Sigma_{ii}^R}{\Sigma_{ii}} \geq 1, \quad F_{ij} \geq 0.$$

[1] This notation follows the explanation in [BS15]

If $F_{ij} = 0$ then $\Sigma^R_{ii} = \Sigma_{ii}$, and we conclude that $x_j(t)$ is not a Granger cause to $x_i(t)$. In conclusion, we can provide another equivalent definition of Granger causality.

Definition 2 *Let $x(t) = (x_1(t), \ldots, x_n(t))$ be an n-dimensional random process. We say x_j does not Granger cause x_i if and only if*

$$F_{ij} = \log \frac{\Sigma^R_{ii}}{\Sigma_{ii}} = 0. \tag{4}$$

We can examine Granger causality for all possible pairs of x_i and x_j for $i, j = 1, \ldots, n$ in a multivariate time series. This constructs $F = [F_{ij}]$ as a matrix of size $n \times n$, where its (i, j) entry indicates the Granger cause of x_i to x_i and the diagonals of F are not of our interest because x_i must have an influence to itself.

In order to examine F_{ij}, one must determine the covariance of prediction errors from the full and reduced models. It is known that a conditional mean is, in general, a nonlinear function of data in the condition and can be analytically characterized only in some cases of known distribution, or simple models. The following sections then describe how to compute the covariance of prediction errors on two specific models and show that they can be described *analytically* in autoregressive models and can be computed systematically in state-space models. In other words, Granger causality definition can be equivalently characterized in terms of model parameters for those two models used in the prediction.

2.3 Characterization of GC on vector autoregressive models

According to (2), a VAR process is a special case of VARMA when C_k's are all zero. It has been used to explain dynamics in time series because of several reasons; one of which is its simplicity of linear model. One can estimate AR coefficient matrices using ordinary linear least-squares (OLS) which is the best estimator in MSE sense (when assuming data is truly generated from AR model corrupted by Gaussian noise.)

From the Granger causality definition (4), [Lüt05, §2.3] describes the Granger characterization for an n-dimensional AR process $x(t)$ of order p with A_k as AR coefficients, for $k = 1, \ldots, p$. It was shown that $x_j(t)$ does not Granger-cause to $x_i(t)$ if and only if

$$F_{ij} = 0 \iff (A_k)_{ij} = 0, \quad k = 1, \ldots, p \tag{5}$$

As a result, Granger causality between time series in VAR processes can be characterized on coefficients in AR matrices and the conditions are simply linear constraints on AR coefficients. After estimating AR model from data, one can read the common zero locations of estimated AR coefficients of all time lags. The zero pattern then explains the Granger causality among the variables.

2.4 Characterization of GC on state-space models

The generalization of a characterization of GC from VAR model to a state-space model was provided by [BS15] and is summarized here. As our goal here is to learn a GC of the source time series, only state-space equations (1a)-(1b) are considered. The noise covariance matrices in this system are

$$\begin{bmatrix} W & S \\ S^T & N \end{bmatrix} = \mathbf{E} \begin{bmatrix} w(t) \\ \eta(t) \end{bmatrix}^T \begin{bmatrix} w(t) \\ \eta(t) \end{bmatrix},$$

where W is the state noise covariance, N is the measurement noise covariance, and S is the correlation matrix of state and measurement noises. Granger causality concept is to determine relationships between time series from the variance of prediction errors. Since, $\hat{x}(t|t-1)$ is chosen to be the optimal estimator of $x(t)$ in MSE sense, it is a classical result that such optimal predictor of $x(t)$ generated from a state-space model, based on information up to time $t-1$ can be obtained from the Kalman filter.

The Kalman filter finds the conditional mean of state variable $z(t)$ based on all available information $\hat{z}(t|t-1) = \mathbf{E}[z(t)|x(t-1), \ldots, x(0)]$ and the corresponding covariance of state estimation error is $P(t|t-1) = \mathbf{cov}(z(t) - \hat{z}(t|t-1))$. When the filter is applied in asymptotic sense, P converges to steady state and satisfies discrete-time algebraic Riccati equation (DARE):

$$P = APA^T - (APC^T + S)(CPC^T + N)^{-1}(CPA^T + S^T) + W. \tag{6}$$

Asymptotically, the covariance of output estimation error is

$$\Sigma = \mathbf{cov}(x(t) - \hat{x}(t|t-1)) = CPC^T + N.$$

From linear system theory, we note that a positive solution to DARE does not always exist. [Sim06] shows that the DARE solution exists under a condition that (A, C) is detectable (all the unobservable states are stable) which is a weaker condition than observability. Moreover, DARE has a unique solution if (A, W) is also controllable and implies all eigenvalue of $(A - KC)$ lie inside the unit circle. However, the covariance matrix P, which obtained from DARE, is not an optimal solution of Kalman filter because the steady state Kalman gain is not an optimal Kalman gain in each iteration but it still converges to an optimal when $t \to \infty$.

We note that if $x \in \mathbf{R}^m$ then $\Sigma \in \mathbf{R}^{m \times m}$ and it is the output estimation error covariance when predicting x using all lagged components in x (full model). To determine an effect of $x_j(t)$ to $x_i(t)$ in Granger sense, we then consider the *reduced model* introduced by eliminating $x_j(t)$ from the full model, and the reduced model is defined as

$$z(t+1) = Az(t) + w(t), \;\; x^R(t) = C^R z(t) + \eta(t),$$

where the superscript R denotes the variable $x(t)$ with j^{th} component eliminated and C^R is obtained by removing the j^{th} row of C. The optimal prediction of $x(t)$ using all information of x except x_j is then also obtained by applying the Kalman filter to the reduced model. We can solve DARE using (A, C^R, W, N) and obtain P^R, denoted as the state estimation error covariance and the output estimation error covariance of the reduced model is given by

$$\Sigma^R = C^R P^R (C^R)^T + N^R$$

where N^R is obtained from N by removing the j^{th} row and column of N. We also note that Σ^R has size $(m-1) \times (m-1)$. Doing this way, we can test if x_j is a Granger cause to x_i for all $i \neq j$ by using the Granger measure:

$$F_{ij} \equiv F_{x_j \to x_i | \text{all other } x} = \log \left(\frac{\Sigma^R_{ii}}{\Sigma_{ii}} \right), \tag{7}$$

where Σ_{ii} and Σ^R_{ii} are the variance of prediction error of $x_i(t)$ obtained from using the full model and the reduced model, respectively. We can repeat the above step for $j = 1, 2, \ldots, m$, learn Granger causality from data by computing F_{ij} for all (i, j) and construct it as a matrix whose diagonals are not in consideration. Subsequently, a significance testing is performed on the off-diagonal entries of this matrix to discard insignificant entries as zeros. The resulting matrix will be called the Granger causality matrix in this report. Figure 2 illustrates a pattern of GC causality from 10-dimensional VARMA process. The F matrix is represented in black and white color, where black in the (i, j) entry explains that x_j is a Granger cause to x_i, while the white in the (i, j) entry indicates there is no Granger causality from x_j to x_i. For example, if we look at the first row, we see that x_i is caused by x_2 and x_3 only.

It was shown in [BS15] that the F measure in (7) can be characterized in the state-space system matrices as well.

$$F_{ij} = 0 \quad \Leftrightarrow \quad C_i^T (A - KC)^k K_j = 0 \tag{8}$$

for $k = 0, 1, \ldots, n$ where C_i^T is the ith row of C, K_j is the jth column of the Kalman gain, solved from DARE. We have seen that the Granger causality condition for VAR model is linear in AR coefficient matrices. Unlike VAR models, GC condition for state-space models is highly nonlinear in system matrices.

3 Related work

Generally brain connectivity studies apply two main approaches: parametric and non-parametric. A non-parametric approach analyzes EEG signal and compute connectivity measures directly without estimating model parameters. For example, one can compute partial coherence measure from estimated spectrum of brain signals. On the other hand, a parametric approach focuses on estimation

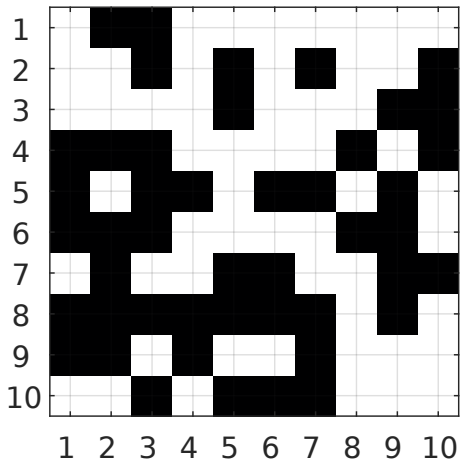


Figure 2: An example of Granger causality pattern encoded in F generated from a sparse VARMA process.

of dynamical models and uses estimated model parameters to infer such connectivity measures. This chapter describes literature on brain connectivity studies that apply the parametric approach. Much of literature has applied typical time series models such as VAR and state-space models. Previous studies can be categorized into two themes: one that infers brain connectivity of scalp signals and the other that concludes a connectivity of the source signal. A conclusion from this survey provides us a guideline to build up our proposed model.

3.1 Connectivity on EEG Signals

This analysis is performed on the scalp EEG signal using a measure of dependence of interest. One typical approach is to fit VAR model to EEG time series and use a measure such as direct transfer function (DTF) as a dependence measure in [GPO12, §4]. The sensor signals are fitted to a VAR model by least-squares estimation and then Granger causality can be obtained by performing significant tests on VAR coefficients. For example, [ACM+07] learned brain connectivity from VAR coefficients using DTF (directed transfer function), PDC (partial directed coherence) and direct DTF (dDTF) from high-resolution EEG data set. Moreover, a state-space framework can be applied to learn connectivity on sensor space, which is introduced in [STOS17]. The state-space model based on switching vector AR (SVAR) model was introduced for non-stationary time series, a characteristic that has been typical for biological signals. The SVAR model was represented in a state-space representation in (3) and the switching parameters were selected by a hidden Markov chain. As a result, the connectivity was learned from PDC that computed from the estimated VAR coefficients.

However, It can be shown that this approach could result in *spurious causality* as mentioned in [HNMN13] where no interactions in the source level may lead to substantial interactions in the scalp level.

3.2 Connectivity on reconstructed sources

The EEG signals cannot explain the true dynamic of neurons inside the brain because of volume conduction effects as shown in Figure 1. An approach of estimating source time series from EEG signals has been developed and is referred to as *source reconstruction* or *source imaging*. The main idea is to estimate $x(t)$ from the lead-field equation:

$$y(t) = Lx(t) + v(t), \tag{9}$$

where $y(t) \in \mathbf{R}^r$ is the EEG data, $x(t) \in \mathbf{R}^m$ is the source signal, $L \in \mathbf{R}^{r\times m}$ is the *lead field matrix* (given) and $v(t) \in \mathbf{R}^r$ is a measurement noise. The lead-field equation (9) can be used to generate artificial EEG signals when $x(t)$ is simulated (known as *forward problem*). On the other hand, constructing the transmitted signal from the measurements in the above linear equation is

often called as an *inverse problem*. In order to solve the inverse problem in practice, we note that the lead field matrix varies upon several factors such as locations of EEG sensors, size or geometry of the head, regions of interest (ROIs) and the electrical conductivity of brain tissues, skull, scalp, etc. [SC13]. Examples of existing methods in source reconstruction are Low resolution tomography (LORETA), the weighted minimum-norm estimate (WMN), the minimum-current estimate, linearly constrained minimum-variance (LCMV) beamforming, sparse basis field expansions (S-FLEX) and the focal underdetermined system solution (FOCUSS) [Hau12, §2], [SC13, HNMN13, LWVS15].

In general, the number of EEG channels is lower than number of sources. Hence, L is generally a fat matrix. As a result, the source imaging problem becomes an underdetermined problem. [MMARPH14] proposed that the source time series matrix is factorized into coding matrix C and a latent source time series $z(t)$, then $x(t) = Cz(t)$ where C is assumed to be sparse. The relationship between sources and sensors is then explained by

$$y(t) = LCz(t) + v(t). \tag{10}$$

The problem of reconstructing x is now to estimate z and C instead. [MMARPH14] applied an ℓ_{21} regularization method by penalizing the rows of the matrix with the 2-norm to induce sparsity pattern in source time series. Then the regularized EEG inverse problem with ℓ_{21}-norm penalty term was proposed as

$$\underset{C,Z}{\text{minimize}} \quad (1/2)\|LCZ - Y\|_F^2 + \lambda\|C_i^T\|_{2,1} + (1/2)\|Z\|_F^2.$$

The problem is non-convex in C and Z (the matrix of latent time series.) An alternating minimization algorithm can be used for solving a bilinear problem by using initial latents $z(0)$ and approximating rank of C from SVD. Another related approach is [WTO16] that applied sLORETA method to estimate source signals x. PCA was used to reduce dimension of the source signals then the principal source signals $\tilde{x}$ were explained $\tilde{x}(t) = Cz(t)$, resulting in a factor model (10) and the dynamics of $z(t)$ was explained by the VAR model. The dynamics of x can then be explained by the VAR model and VAR coefficients are functions of C.

We note that brain connectivities learned from a source reconstruction approach mainly depends on the performance of the source imaging technique. If the source reconstruction does not perform well, learning brain networks from reconstructed sources could lead to a misinterpretation.

3.3 Connectivity inferred from source and EEG coupled dynamics

This approach considers the dynamics of both source and sensor signals concurrently where the estimation of model parameters can infer brain connectivities directly. The work including [Hau12, HTN$^+$10, GHAEC08, CWM12] considered the same dynamical model that the source signals (x) are explained by a VAR process and EEG signal (y) is a linear combination of the sources as

$$x(t) = \sum_{k=1}^{p} A_k x(t-k) + w(t), \quad y(t) = Lx(t).$$

The technique to estimate unknown sources and lead field matrix (L) from only available mixture EEG data is called *blind source separation*. Independent component analysis (ICA) is one of blind source separation technique that was used in [Hau12, HTN$^+$10, GHAEC08]. In the detail, the ICA technique relies on an assumption that the innovation term of process $w(t)$ must be generalized as a non-Gaussian distribution. [GHAEC08] assumed that the innovation term has both sub and super-Gaussian distribution. Initially, PCA was used to reduce the dimension of EEG data with assumption that number of EEG channels was greater than number of sources. Consequently, the principal EEG signals were fitted on VAR model directly and ICA was performed on the VAR innovation term for demixing source VAR coefficients. As a result, DTF was computed from the transfer function of the source in VAR model. However, [GHAEC08] estimated VAR parameters from the sensor signals directly, so the brain connectivity was not sparse due to the volume conduction effect. [Hau12] performed convolutive ICA (CICA) on the innovation term which was assumed to be super-Gaussian hyperbolic secant distributed for ensuring a stable solution. To obtain the sparse source connectivity, model parameters, which are L and A_k's, are estimated using the sum of ℓ_2-regularized maximum-likelihood method. In addition, [Hau12, HTN$^+$10, GHAEC08] assumed that the noise distribution was non-Gaussian, so the decomposition of source signals from ICA had a unique solution. [CWM12] proposed

an idea to perform connectivity analysis via state-space models. The state equation was described by *generalized AR model* where the innovation process has a generalized Gaussian distribution. All state-space model parameters were obtained using from maximum likelihood estimation. As a result, the relationship between sources was explained by PDC computed from estimated VAR coefficients. [CRTVV10] proposed a state-space framework for finding a brain connectivity; however, the sources were assumed to be described by a VAR model. Moreover, [CRTVV10] put some prior information on the lead-field matrix where the cortical regions of interest were known. The dynamical equations are given by

$$x(t) = \sum_{k=1}^{p} A_k x(t-k), \quad y(t) = C\Lambda x(t) + v(t)$$

where C is a known matrix from a prior information on the lead field matrix and Λ is the dipole moment. When formulating the above equation into a state-space form, model parameters including $A_1, \ldots, A_p, \Lambda$ and noise covariance were estimated by expected-maximization (EM) algorithm and then Granger causality can be concluded from the estimated noise covariance. Moreover, a state-space form used in [CRTVV10, CWM12] contains source dynamics described by VAR model and the observation equation represents a relationship between sources and sensors. The state-space parameters were estimated from maximum likelihood estimation using EM. [YYR16] proposed *a one-step state-space model* estimation framework which aims to find the connectivity in ROI level. The state-space model used in [YYR16] was described by

$$z(t+1) = A(t)z(t) + w(t), \quad x(t) = Gz(t) + \eta(t), \quad y(t) = Lx(t) + v(t),$$

where $z(t)$ is a time series for each ROI, $A(t)$ is a VAR coefficient at time t, $x(t)$ is a source time series, G is a binary matrix that determines sources corresponding to its ROIs and $y(t)$ is MEG signal. Hence, the state-space model in [YYR16] is essentially a first-order VAR model. The model parameters and source signals were estimated using EM algorithm and the ROIs connectivity pattern was explained from the zero pattern in VAR coefficients. [CRTVV10] claimed that the state-space framework was less sensitive to noise than two-stage approaches. Another possibility was to employ state-space model to explain brain source connectivity from fMRI data. [PZBC17] proposed a method to estimate a brain connectivity on a linearized dynamic causal model (DCM), described by

$$\dot{x}(t) = \left(A + \sum_{j=1}^{m} u_j(t) B_j\right) x(t) + Cu(t). \tag{11}$$

However, [PZBC17] considered resting-state data which means there was no stimulus signals ($u(t) = 0$). The DCM was discretized with the sampling period of 2 seconds ($h = 2$) and sampled data system is described by

$$x(k+1) = e^{Ah} x(k) + w(k). \tag{12}$$

The nonlinear dynamical model for fMRI data used in [PZBC17] is a Balloon-Windkessel model which is a forth-order state equation that can be linearized by using a Finite impulse response (FIR) as

$$y(k) = \sum_{\tau=0}^{k-1} h(\tau) x(k-\tau). \tag{13}$$

Therefore, the state-space model was described by (12) and (13). Estimation of neural activities was performed using the Kalman filter and Rauch-Tung-Striebel (RTS) smoother. Then, connectivity matrix A was estimated using EM algorithm.

To conclude this section, learning brain connectivities from EEG data can be divided into two main approaches. The first approach explore a causality from EEG data directly (sensor space). However, a connectivity between EEG sensors is not an intrinsic connectivity explaining relationships of neuronal activities in the human brain. The second approach, consisting of *two-stage approach* and *coupled models*, is to learn a brain connectivity from source signals (source space). The two-stage approach reconstructs source signals first and often explains source dynamics via VAR models. However, the performance of the two-stage approach is highly dependent of performance of source reconstruction. *Coupled models* are then proposed for explaining dynamics of sources and EEG signals concurrently where brain connectivities are discovered from the estimated model parameters.

3.4 Validation of brain connectivity learned from EEG data

An evaluation of brain connectivity learning method is typically done on a simulated data set while the connectivity results on real data set cannot be validated since the true connectivity on source space is unknown. A recent discussion of brain connectivity validation on EEG data sets can be found in [M$^+$19] where performance evaluation methods can be proposed in many different ways. The first scheme is to validate a method performance on simulated data with a known ground-truth model *e.g.* [Hau12, §5] and [HE16]. Assume that brain source signals obey a VAR process, then the connectivity matrix is expected to be similar to zero pattern of VAR coefficients as shown in (5). Moreover, [HE16] provides a toolbox for simulated pseudo-EEG data with realistic head model based on a bivariate AR model. The second scheme is to validate a method performance on real EEG data. A possible approach is done by comparing connectivity results with previous studies on other modalities such as fMRI. For example, the brain connectivity from resting-state EEG data in [WTO16] was compared with the brain connectivity based on resting-state fMRI.

However, most EEG studies considered task-EEG data such as brain computer interface (BCI) or visual task, so connectivity results are expected to follow the task-based connectivity pattern. [HD$^+$14] validated brain connectivity results from different two-stage approaches on picture and naming recognition EEG task data. The density of connection between occipital, temporal and frontal regions is expected to be found from the task data. Finally, the estimated connectivity was compared with the estimated connectivity matrix from other methods. [YYR16] compared connectivity results with another two-stage approach, which is minimum-norm estimate (MNE) method, from the same data set of Magnetoencephalography (MEG). However, this scheme is mostly applied to simulation data, which can be found in [HTN$^+$10, CWM12].

4 Proposed method

This section describes the methodology of learning Granger causality patterns from EEG time series data. The method consists of five main processes. From the proposed state-space model in (1), the

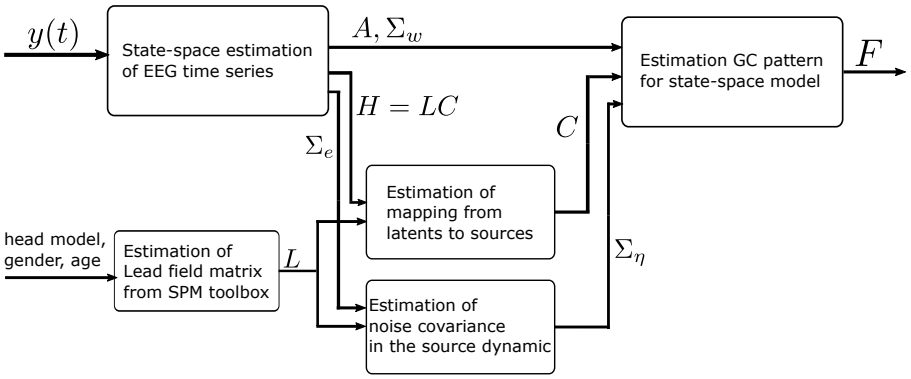


Figure 3: The proposed scheme for learning Granger causality from EEG data.

only available measurement is EEG signal (y). If we substitute the dynamics of source (x) in the EEG forward equation, we have

$$z(t+1) = Az(t) + w(t), \quad y(t) = Hz(t) + e(t) \tag{14}$$

where

$$H = LC, \quad e = L\eta + v. \tag{15}$$

The matrices Σ_w and Σ_e are noise covariances of w and e, respectively. We can view e as a combination of noises corrupted in the latents and source signals, as perceived at the output equation.

Our method assumes that EEG time series y and the lead-field matrix L are available. Note that L can be estimated using a head model, general information about a subject and sensor locations. The

first procedure is to estimate a state-space model to obtain system dynamic matrix and output matrix (A and H). Since we need to learn a connectivity in x, the output matrix in the source equation C needs to be estimated. Moreover, in order to learn a GC pattern, noise covariances are also needed. A process of noise covariance estimation is then proposed. After, a GC matrix F is estimated, it also requires a process of learning significant entries of F as generally entries of F are nonzero.

4.1 Data generation

Generating dynamical models is one of important parts to perform experiments on Granger causality estimation so that we can evaluate the accuracy of estimated GC pattern with the ground-truth model. This step is simple in generating VAR processes as a Granger causality is linearly encoded in VAR parameters. In this section, we explain an approach of generating state-space models where we can control the true GC pattern.

4.1.1 VARMA models with sparse VAR part

In [BS15, BS11], the authors have shown an important result that GC causality of a filtered VAR process is unchanged if the filter is diagonal, stable and minimum-phase. Let $\tilde{x}(t)$ be a p-lagged VAR process where the z transform relation is given by $\tilde{x} = A(z)^{-1}w$ with VAR polynomial:

$$A(z) = I - (A_1 z^{-1} + A_2 z^{-2} + \cdots + A_p z^{-p})$$

We consider $G(z)$ an MIMO (multi-input multi-output) transfer function of the form:

$$G(z) = \begin{bmatrix} \frac{p_1(z)}{q_1(z)} & & & \\ & \frac{p_2(z)}{q_2(z)} & & \\ & & \ddots & \\ & & & \frac{p_n(z)}{q_n(z)} \end{bmatrix} \tag{16}$$

where each of diagonal entries of G is a rational proper transfer function of relative degree q. The minimum-phase property of G suggests that zeros of G must be inside the unit circle, or that the roots of $p_i(z)$ has magnitude less than one. Moreover, stability of G implies that the roots of $q_i(z)$ lie inside the unit circle. As a result, we define $x = G\tilde{x} = G(z)A(z)^{-1}w$. The result from [BS11] shows that x also has the same GC pattern as $\tilde{x}$, which is easily explained from a zero pattern in VAR coefficients. The system transfer function from w to x can be equivalently represented in a state-space form. Therefore, we proposed a procedure to generate state-space equation with sparse GC pattern as follows.

1. Generate sparse $A_1, A_2, \ldots, A_p$ matrices randomly with a common zero pattern. Moreover, the polynomial $A(z)$ must be stable. This is to guarantee that the generated VAR process is stationary. We can do this by randomize stable roots inside the unit circle and compose the polynomial in the diagonal of $A(z)$. Consequently, off-diagonal entries of A_k's are generated randomly in a common (i, j) location. If $A(z)$ is not stable, we randomize off-diagonal entries again.

2. Generate a random diagonal transfer function $G(z)$ with required properties. We can generate stable zeros and poles of $G(z)$ when the order q is given.

3. The transfer function from w to x, the desired source signal, is then given by $H(z) = G(z)A(z)^{-1}$. Convert H into a discrete-time state-space form using `tf2ss` command in MATLAB. We obtain (A, B, C, D) of the state-equation:

$$z(t+1) = Az(t) + Bw(t), \quad x(t) = Cz(t) + Dw(t). \tag{17}$$

   Since H is a proper transfer function, we have $D = 0$.

State-space equation and VARMA models can be interchangably transformed [CGHJ12], so we can refer the generated model (17) as state-space or VARMA model with sparse GC pattern.

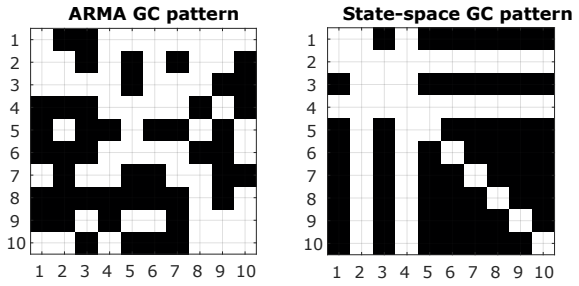


Figure 4: GC pattern the ground-truth models. *(Left.)* VARMA processes with sparse VAR part. *(Right.)* State-space model with sparse rows in C.

Special case of $G(z)$. As suggested in [BS15] is when $G(z)$ in (16) has the form of

$$G(z) = (1 + cz^{-1})^q I = (1 + C_1 z^{-1} + \cdots + C_q z^{-q})I \triangleq C(z)I, \quad |c| < 1$$

One can see that $G(z)$ is an MA polynomial of order q. When $|c| < 1$, $C(z)$ is minimum-phase and C_j's are given by

$$C_j = \begin{pmatrix} q \\ c \end{pmatrix} c^j, \quad j = 1, 2, \ldots, q.$$

In this case, $x = G(z)A(z)^{-1}w = C(z)A(z)^{-1}w = A(z)^{-1}C(z)w$ since $C(z)$ is just a scalar. We can then readily consider x as a VARMA(p, q) process. The AR and MA coefficients in $A(z)$ and $C(z)$ can be used to convert into a state-space form, for example, the Hamilton form (3).

4.1.2 State-space models with sparse rows in C

We could start with randomly generating a stable state-space model and hope this ground-truth model has a non-trivial GC pattern (not dense). However, even though system matrices (A, C) are sparse, it is not necessary that F solved from DARE is also sparse as F is nonlinear in (A, C, K); see (8).

However, if we consider (8), under the following assumptions:

$$C_i^T = 0, \quad S = 0, \quad N \text{ is diagonal}$$

then one can prove that the ith row and jth column of F are zero. As a result, we can randomly generate C with sparse rows and stable A in (1a). Note that when A is not assumed to have a specific structure such as diagonal, the stability constraint (eigenvalues of A lie inside the unit circle) is nonlinear in the entries of A. These random generations are repeated until the stability condition is met. With this generation, we obtain a stable A more easily than generating stable sparse VAR model.

Figure 4 (left) shows an example of GC pattern generated from a VARMA process. The GC pattern was concluded from the structure of F matrix in (7) and must agree with the sparsity pattern of A_k's we have generated. The right figure is a GC matrix generated from state-space models with sparse rows in C. The GC pattern from this approach is more restricted than generating VARMA process with sparse AR part as zeros appear in many blocks of F. Therefore, at this stage of our work, we use sparse VARMA processes which are shown to be equivalent to state-space models, and their GC patterns can be controlled arbitrarily at the AR coefficient generating process.

4.2 State-space estimation of EEG time series

Given the measurement data of $\{y(t)\}_{t=0}^N$, we can estimate state-space parameters A and H in (14) using the *subspace identification method* [OM12] which is available in the system identification toolbox `n4sid` on MATLAB. An estimated state-space model in this toolbox is of the form:

$$\begin{aligned} z(t+1) &= Az(t) + Bu(t) + Ke(t), \\ y(t) &= Hz(t) + Du(t) + e(t), \end{aligned} \tag{18}$$

where u is a deterministic input with the input matrix B, and K is the Kalman gain matrix. Comparing (14) with the model format in (18), we force B and D to be zero using the prediction error method

(PEM) since our model does not have a deterministic input u, which can be done using the command `pem` in MATLAB. In addition, we enforce the stability condition of the estimated model in `n4sid` which guarantees that all eigenvalues of $\hat{A}$ are inside the unit circle. This process gives estimates of A, H, Σ_w and Σ_e.

4.3 Estimation of mapping from latents to sources

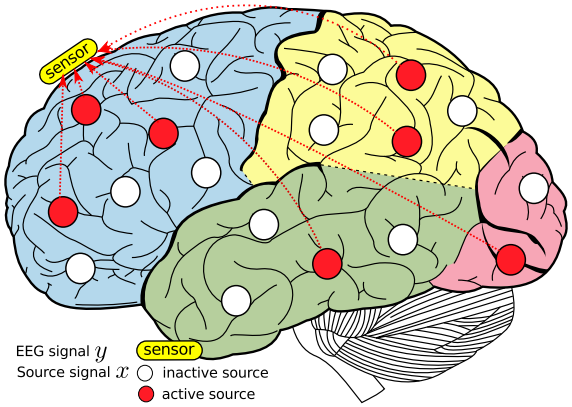


Figure 5: A brain map that explains the activation status of source signals. Red circles represent active sources and white circles represent inactive sources. In other words, EEG signals come from only active sources

This section explain how to estimate C from the information of H estimated from section 4.2 with a prior knowledge of a structure in C. The contents of this section are taken from our publication [PiS18]. Recall from (1b) that the ith source can be interpreted as inactive ($x(t) = 0$) if the ith row of C is entirely zero (in noiseless condition).

In general, the number of EEG channels are less than the number of sources, $r < m$, then $L \in \mathbf{R}^{r\times m}$ is fat matrix. Assume that a lead field matrix L can be estimated from prior knowledge head model. Demixing C from H, with a known fat matrix L, *i.e.*, solving C from $H = LC$, is an underdetermined problem and it leads to non-unique solutions of C.

To overcome this problem, we put some prior in C by assuming that *not all sources are active* because all neuron sources are not activated synchronously. There are only some sources activate corresponding to EEG sensors in a brief time. Consequently, C is *assumed to have some zero rows* corresponding to inactive sources as shown in Figure 5. We therefore propose a problem of estimating C that makes $H \approx LC$ and C contains many zero rows. One way to formulate the problem is to make use of sparse optimization with ℓ_1-norm penalty as mentioned in [HTW15]. The proposed problem is

$$\underset{C}{\text{minimize}} \quad (1/2)\|H - LC\|_F^2 + \lambda \sum_{i=1}^{m} \|C_i^T\|_2 \tag{19}$$

with variable $C \in \mathbf{R}^{m\times n}$. The matrix $H \in \mathbf{R}^{r\times n}$ is the estimated output matrix from state-space estimation, and $L \in \mathbf{R}^{r\times m}$ is the lead-field matrix computed from a head model, and $\lambda > 0$ is a penalty parameter controlling sparsity of rows in C, *i.e.*, when λ is large, C tends to have more sparse rows. When the formulation (19) is rearranged into a vector form of

$$\underset{\beta}{\text{minimize}} \quad (1/2)\|y - X\beta\|_2^2 + \lambda \sum_{i=1}^{m} \|\beta_i\|_2 \tag{20}$$

where $\beta = (\beta_1, \beta_2, \ldots, \beta_m), \beta_k \in \mathbf{R}^n$, it is commonly known as a *group lasso* problem [HTW15, §3.8] where the sum of ℓ_2-norm regularization promotes a *group* sparsity in the estimated variable. If λ is high enough, the solution to (19) is entirely zero. The problem is convex problem because objective function is an affine of norm functions which are convex. There are many numerical methods for solving

this problem where currently we have implemented the Alternating Direction Method of Multipliers (ADMM) described in [PB14, Son13].

The penalty parameter affects the sparsity of C and therefore, a model selection criterion such as BIC can be applied to trade-off between the model fitting and complexity. The BIC score [HTW15] is given by $-2\mathcal{L}(\beta) + k \log N$ where $\mathcal{L}$ is the loglikelihood function, k is the number of effective parameters and N is the number of data samples in the estimation. The BIC score of the problem (19) reduces to

$$\text{BIC}(\lambda) = -nm - n \log\det \hat{\Sigma}(\lambda) + k(\lambda)\log(n) \tag{21}$$

where $\hat{\Sigma}(\lambda) = (1/n)(H - L\hat{C}(\lambda))(H - L\hat{C}(\lambda))^T$. We note that BIC is a function of λ because for each λ, it corresponds to a sparsity of rows in C. Hence, $k(\lambda)$ decreases as λ decreases. We choose the estimated C that yields the lowest BIC score.

To vary λ in a range and choose one that minimizes BIC, we consider an analytical form of a critical value that results in the zero solution. In other words, there exists λ_{max} such that if $\lambda > \lambda_{\max}$ then the solution C to (19) is entirely zero [Son13, §4.3]. The expression of $\lambda_{\max}$ depends on the problem data, which are H and L only, hence it can be computed beforehand. Therefore, we can varying the value of λ in (19) by starting from $10^{-4}\lambda_{\text{max}}$ to λ_{max}. As a result, the solutions C from (19) vary from densest to sparsest solutions.

Uniqueness of the solution. The estimation formulation (19) apparently relies on a parameter H that is solved from a subspace identification. However, it is known that state-space system parameters are not unique due to a similarity transformation. If the estimated H is associated with another coordinate, denoted by $\tilde{H}$, we would question if solving (19) using $\tilde{H}$ could lead to a different solution of C or not. Moreover, even C is not unique, we should examine whether the sparsity of rows in C is unique or not. If it were not, we would interpret results on selecting active sources differently.

We provide an analysis of uniqueness in the sparsity of C under some mild assumption. If H is projected to another coordinate, that is $\tilde{H} = HU$ where U is a *nonsingular orthogonal* matrix and denoted as a transformation matrix, then the solution from (19) could change but its rows has *the same zero pattern*.

Proposition 3 *If U is orthogonal, then the rows of $\tilde{C}$, solved from*

$$\textit{minimize}_{\tilde{C}} \quad (1/2)\|HU - L\tilde{C}\|_F^2 + \lambda \sum_{i=1}^{n} \|\tilde{C}_i^T\|_2 \tag{22}$$

has the same zero pattern as the solution C of (19)*. Moreover, $\tilde{C} = CU$.*

Proof. Since T is orthogonal, we have $U^TU = UU^T = I$, and that $\|UX\|_F^2 = \|U\|_F^2$ for any matrix X. Then we can write the cost objective of (22) as

$$\begin{aligned}(1/2)\|HU - L\tilde{C}\|_F^2 + \lambda \sum_{i=1}^{m} \|\tilde{C}_i^T\|_2 &= (1/2)\|(H - L\tilde{C}U^T)U\|_F^2 + \lambda \sum_{i=1}^{m} \|\tilde{C}_i^T U^T\|_2 \\ &= (1/2)\|(H - L\tilde{C}U^T)\|_F^2 + \lambda \sum_{i=1}^{m} \|\tilde{C}_i^T U^T\|_2\end{aligned}$$

If we let

$$C = \begin{bmatrix} C_1^T \\ \vdots \\ C_m^T \end{bmatrix} = \tilde{C}U^T$$

then the cost objective of (22) is the same as (19). The matrix C that minimizes (19) corresponds to $\tilde{C} = CU$ that minimizes (22). Moreover, we see that $\tilde{C}_i^T = C_i^T U$. Hence, the ith row of $\tilde{C}$ is zero if and only if the ith row of C is.

4.4 Estimation of noise covariance in the source dynamic

The GC estimation from state-space model parameters explained in Section 2.4 requires information of noise covariances (both state and measurement noises). Consider our methodology in the diagram 3 and the model equations (1a) and (1b). Currently, we have estimated A, Σ_w from subspace identification, and we have factored C. Then it is left to estimate Σ_η (the measurement noise covariance at the source equation) in order to solve a GC matrix via the Riccati equation.

We then propose that, the relation of noises at EEG and source equations given by (15) leads to a relation of noise covariance as'

$$\Sigma_e = L\Sigma_\eta L^T + \Sigma_v. \tag{23}$$

The covariance of e is obtained in Section 4.2 from the subspace identification. The lead field matrix can be obtained from a head model (as part of our assumptions). Therefore, it is left to estimate the unknown Σ_η and Σ_v. Consider the dimensions of all these matrices, where they are symmetric and positive definite, *i.e.*, $\Sigma_e \in \mathbf{S}^r$ and $\Sigma_\eta \in \mathbf{S}^m$ and $\Sigma_v \in \mathbf{S}^r$, and that $r < m$ (the number of EEG channels is typically less than the number of sources.) A straightforward formulation is to solve

$$\begin{array}{ll} \text{minimize} & \|\Sigma_e - L\Sigma_\eta L^T - \Sigma_v\|_F \\ \text{subject to} & \Sigma_\eta \succeq 0, \Sigma_v \succeq 0, \end{array}$$

with variables $\Sigma_\eta \in \mathbf{S}^m$ and $\Sigma_v \in \mathbf{S}^r$ and $\|\cdot\|_F$ denotes the Frobenius norm. If the optimal value the above problem is zero, it means we can solve (23) exactly. However, since $r < m$, it is possible that even we can have the zero optimal value, but the solution Σ_η is not unique since we have more degree of freedoms in the variables. We then further restrict some constraints on the variables, by assuming that these noise covariances are *diagonal*, meaning that each of noise vectors η and v is mutually uncorrelated. Following this assumption, we propose the estimation problem:

$$\begin{array}{ll} \text{minimize} & \|\Sigma_e - L\Sigma_\eta L^T - \Sigma_v\|_F \\ \text{subject to} & \Sigma_\eta \succeq 0, \Sigma_v \succeq 0, \\ & \Sigma_\eta = \alpha I, \quad \Sigma_v \text{ is diagonal} \end{array} \tag{24}$$

with variables α and Σ_v. Imposing the diagonal structure in the variables certainly results in a chance of nonzero optimal value (we have not solved (23) exactly). However, if the true noise covariances have these diagonal structures, the problem (24) can alleviate the chance of non-unique solutions (but we have not guaranteed either.) The cost objective of (24) is a composite of a norm function with linear transformation in the variables. Moreover, the constraints are linear in the variables and the positive definite cone constraints and hence, the constraint set is a convex set. For these two reasons, the problem (24) is convex and can be solved efficiently.

4.5 Learning significant Granger causality pattern

From the methods explained in Section 4.2 through Section (4.4), we are ready to estimate a GC pattern from EEG data. To summarize here, the GC characterization on state-space models is used for finding GC of $x(t)$ that described from (1a)-(1b):

$$\begin{aligned} z(t+1) &= Az(t) + w(t), \\ x(t) &= Cz(t) + \eta(t). \end{aligned}$$

Available information needed for GC estimation are

- the state transition matrix A,
- the state noise covariance matrix Σ_w from subspace identification obtained from Section 4.2,
- the latent-to-source mapping matrix C solved from the method in Section 4.3,
- the noise covariance in the source dynamic, Σ_η, obtained from the method in Section 4.4.

The computation of GC that characterized on state-space model follows the details in Section 2.4. It consists of solving DARE (6) for the *full model* and *reduced model*. The DARE solutions, denoted by P can be obtained by using `dare` command in MATLAB.

Full model. We solve DARE for full model by using $(A, C, \Sigma_w, \Sigma_\eta)$ as input parameters to obtain P (as the covariance error of $\hat{z}$) and then the covariance of source estimation error is obtained from $\Sigma = CPC^T$ with dimension $m \times m$.

Reduced model. There are m components in x. Suppose we would like to test if x_j (for a given $j \in \{1, \ldots, m\}$, is a cause the other $m-1$ variables, we solve DARE for a reduced model when x_j is removed. Thus, we perform the following steps:

- Remove x_j from (1b). This equivalent to deleting the j^{th} row of C and obtain C^R as the output matrix in the reduced model.
- The noise covariance of η in the reduced model is Σ_η^R which is obtained by deleting the j^{th} row and column of Σ_η.
- Solve DARE using $(A, C^R, \Sigma_w, \Sigma_\eta^R)$ as input parameters to obtain P^R (the covariance error of $\hat{z}$ in the reduced model.) Compute the covariance error of $\hat{x}$ as $\Sigma^R = C^R P^R (C^R)^T$, that has size $(m-1) \times (m-1)$.
- Compute F_{ij} from (7) for $i = 1, \ldots, m$, except $i = j$ (since $F_{jj} \neq 0$ and we are not interested in learning causality from x_j to itself.)

The above steps are then repeated for $j = 1, \ldots, m$ and we obtain the GC matrix, F where its zero entries indicate the pair of variables that contain no causality.

We can learn a Granger causality pattern of a time series by a statistical test on the Granger causality measure described in (7). If one uses VAR models, the test becomes the log-likelihood ratio test for a nested VAR model. Moreover, GC inference of VAR models can be represented in many equivalent forms other than VAR parameters such as autocovariance sequence and cross-power spectral density where a MATLAB toolbox for this test is available [BS14]. However, it was stated in [BS15] that as the GC inference measure in (7) does not have a theoretical asymptotic distribution, a significance testing can be alternatively done through permutation or bootstrapping methods. In their experimental results, however, it was suggested that the test statistics is well-approximated by a Γ distribution.

This section describes a scheme of learning significant GC pattern proposed in our work [PiS19] (the following contents are taken out from this publication.) Firstly, we construct a set of the sample means of F estimated. Let N_0 be the number trials of EEG time series. Subspace identification is performed to each of these trials to estimate A and C and the noise covariances in (1a)-(1b). We estimated GC matrices (F) by solving Riccati equation explained in Section 2.4. We then take the average over N_1 samples out of these N_0 samples of F and obtain $\bar{F}$ in the amount of $N_2 = N_0/N_1$ samples. As the distribution of all entries in F is unknown, we can apply the central limit theorem to claim that if N_1 is large enough, the entries in $\bar{F}$ approach a Gaussian distribution. When all (i, j) entries of $\bar{F}$ are pooled together ($\bar{F}$ is vectorized), their histogram is shown in Figure 6 and we recognize that it consists of several components of Gaussian distributions having different means and variances. The Gaussian with the lowest mean corresponds to small (i, j) entries in $\bar{F}$ implying that these entries should be regarded as no Granger causality. Other Gaussian components refer to entries in $\bar{F}$ having significant magnitudes, so Granger causality exists in these (i, j) entries.

Following this observation, we propose to fit a *Gaussian Mixture Model (GMM)* [HTF09] to samples of vectorized $\bar{F}$. Denote Y the random variable of vectorized $\bar{F}$. The GMM model takes the form of

$$Y = Z_1 Y_1 + Z_2 Y_2 + \cdots + Z_K Y_K$$

where K is the number of Gaussian components, Y_k is the Gaussian variable with parameter (μ_k, σ_k^2) for $k = 1, \ldots, K$ and $(Z_1, \ldots, Z_K)$ is the latent variable having a multinomial distribution, *i.e.*, possible values of Z are

$$Z = (1, 0, \ldots, 0), (0, 1, 0, \ldots, 0), \ldots, (0, 0, \ldots, 0, 1)$$

and each of the value above is associated with a multinomial pmf $\pi = (\pi_1, \pi_2, \ldots, \pi_K)$. Whenever it is given that $Z = \mathbf{e}_k$ (a standard unit vector), Y is distributed by the kth Gaussian model. The pdf of Y is given by

$$f(y) = \pi_1 f_1(y; \mu_1, \sigma_1^2) + \cdots + \pi_K f_K(y; \mu_K, \sigma_K^2)$$

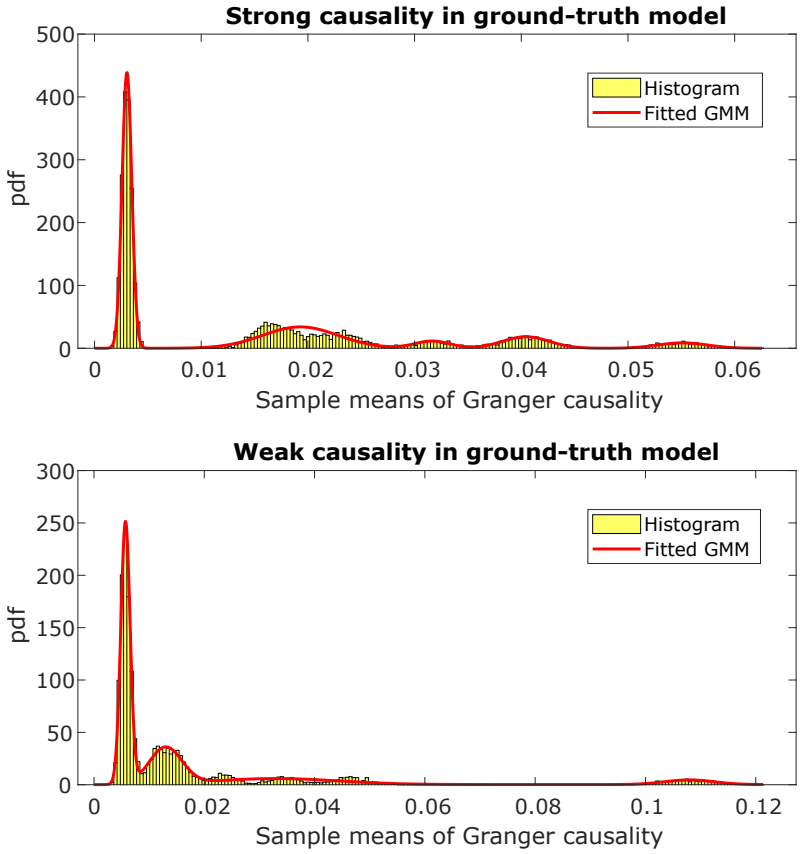


Figure 6: Examples of fitting GMM to vectorized $\bar{F}$ in the training set.

where f_k is the Gaussian density for $k = 1, \ldots, K$. We assume that the Gaussian components are sorted according to the lowest to highest μ_k. After fitting GMM with EM (Expected-Maximization algorithm), we can classify each of (i, j) entries of $\bar{F}$ into two groups: zero Granger causality (null) and significant Granger causality (causal) by the following methods:

- The entries clustered by the posterior probabilities into the first Gaussian component with parameter (μ_1, σ_1^2) are regarded as *null*. The entries clustered into the other $K - 1$ components are classified as *causal*. Here, the posterior probabilities are computed as

$$f_k(y|Z = k; \mu_k, \sigma_k^2) P(Z = \mathbf{e}_k; \mu_k, \sigma_k^2)$$

  for $k = 1, \ldots, K$.

- We consider the first two distributions: f_1 and f_2 where they should be regarded as the pdfs of null and causal entries, respectively. We determine a threshold y_c such that

$$\log f_1(y_c; \mu_1, \sigma_1^2) = \log f_2(y_c; \mu_2, \sigma_2^2).$$

  If a sample of entry in $\bar{F}$ is greater than y_c, we classify it as causal entry and null otherwise.

The scheme we have explained can be summarized in Figure 7. The implementation of our approach also involves choosing the number of components in GMM. We have considered three indices: BIC, relative change of BIC [MP04, §6.9] and Silhouette score. The first lowest number of components having the relative change of BIC less than a threshold is selected. The Silhouette score ranges from -1 to 1 indicating the degree to which the data are well-clustered [KR09, §2.2]. AIC score is not in consideration here because it tends to choose a complex model. We will examine the performances of our approach based on these three options.

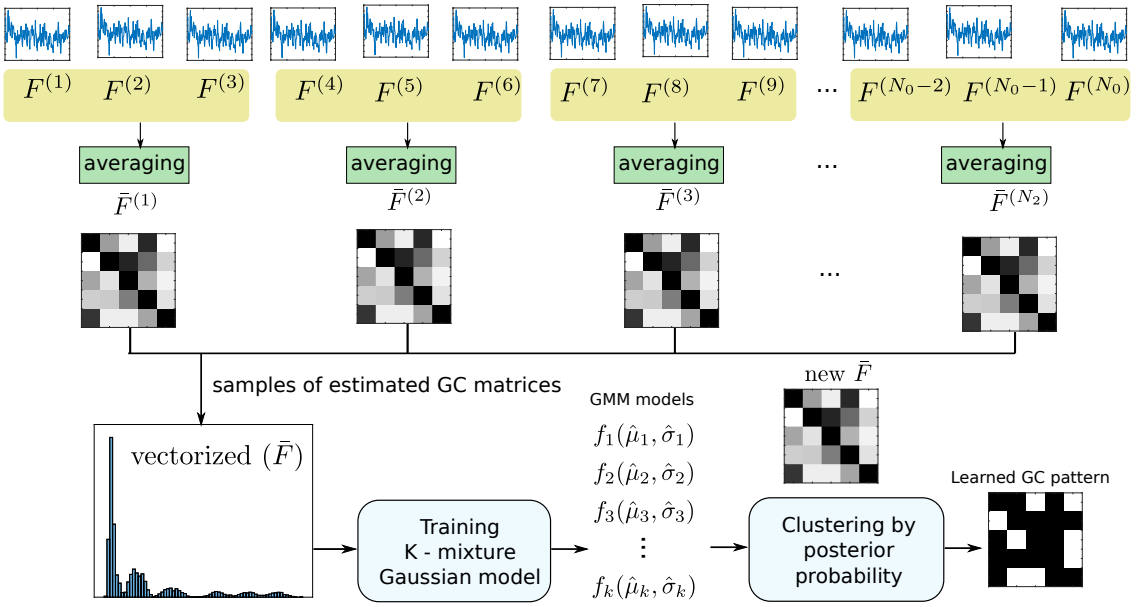


Figure 7: The proposed scheme of learning causality pattern from estimated GC matrices (F).

4.6 Performance evaluation

We aim to evaluate the performance of our method on simulated EEG data set first. In data generating process, we can set up model dimensions (n, m, r) and a ground-truth sparsity pattern on GC matrix associated with such models. In an estimation process, one needs to assume the model dimension; here let $(\tilde{m}, \tilde{n})$ be the number of sources and latents in the estimation which could be larger or smaller than (m, n), while r (the number of EEG channels) is certainly known. Then it leads to a condition in an evaluation procedure since the estimated matrix F of size $\tilde{m} \times \tilde{m}$ has a different dimension from the ground-truth matrix F. Secondly, when considering the performance of detecting causality in each of (i, j) entries in F, this is a kind of binary classification problem. A sparsity pattern of estimated F could come from two possibilities: one from regarding inactive sources reflected on sparse rows of C, and the other is from the attempt to learn zeros in the ground-truth matrix F.

To classify the estimated F_{ij} as zero or nonzero, we define nonzero as positive and zero as negative. Hence, performance indices for Granger causality significance test are

- True positive (TP): correctly identified nonzeros in F
- True nagative (TN): correctly identified zeros in F
- False positive (FP): incorrectly identified nonzeros in F
- False negative (FN): incorrectly identified zeros in F

From these measures, we also apply the following traditional classification measures:

$$\text{Accuracy (ACC)} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{25}$$

$$\text{True positive rate (TPR)} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{26}$$

$$\text{True negative rate (TNR)} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \tag{27}$$

$$\text{False positive rate (FPR)} = \frac{\text{FP}}{\text{TN} + \text{FP}} = 1 - \text{TNR}, \tag{28}$$

$$\text{False negative rate (FNR)} = \frac{\text{FN}}{\text{TP} + \text{FN}} = 1 - \text{TPR}. \tag{29}$$

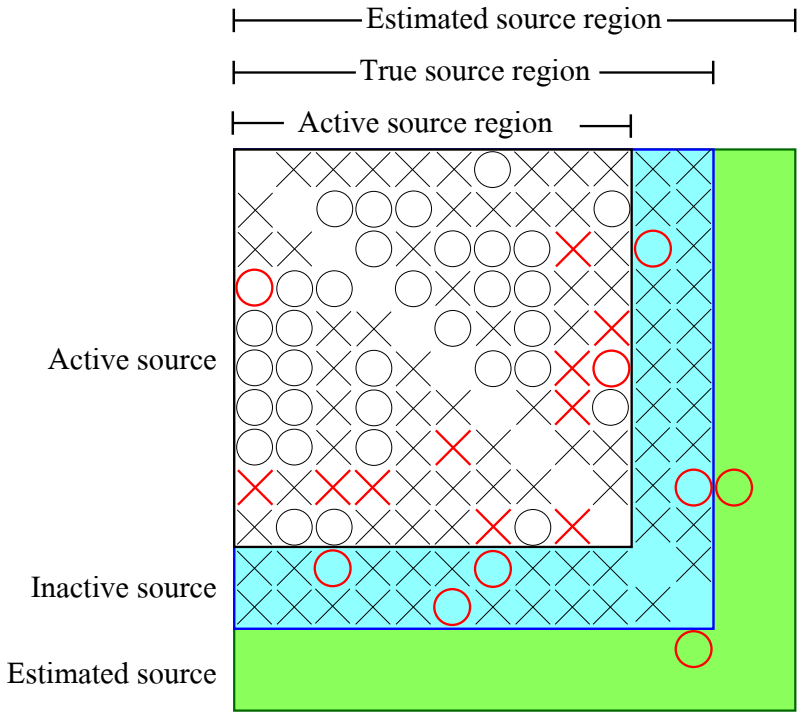


Figure 8: Example of Granger causality evaluation including active source regions, true source region and estimated source region. Black ○ are TP; Red ○ are FP; Red × are FN and black × are TN.

		True condition F	
		Condition $F \neq 0$	Condition $F = 0$
Predicted F	Predicted $F \neq 0$	True positive (***TP***)	False positive (***FP***) **Type I error**
	Predicted $F = 0$	False negative (***FN***) **Type II error**	True negative (***TN***)

Figure 9: Granger causality classification performance indices.

Recall that a ground-truth model used to generate data (1a)-(1c) is given by

$$z(t+1) = Az(t) + wt), \quad x(t) = Cz(t) + \eta(t), \quad y(t) = Lx(t) + v(t),$$

where $z \in \mathbf{R}^n, x \in \mathbf{R}^m, y \in \mathbf{R}^r$. From the issue that the number of estimated sources $(\tilde{m})$ may not be equal to the number of sources in ground-truth model (m), we describe how to calculate the classification measures in a fair setting. In this study, we assume that $\tilde{m} > m$ since we can overestimate the number of sources and we expect the source selection procedure to remove inactive sources at the end. By this assumption, $\hat{F} \in \mathbf{R}^{\tilde{m} \times \tilde{m}}$, which is a matrix with a bigger size than the true Granger causality matrix $F \in \mathbf{R}^{m \times m}$.

Figure 8 shows all three square regions involved in the evaluation process. We start with the **true source region (T)** that contains all the sources in a ground-truth model, and since not all sources are active, a subset called **active source region (A)** consists of all the true active sources where we can reorder the source coordinates so that active sources contain in this region. We define the **estimated source region (E)** as the set of all sources considered in an estimated model. By the assumption that $\tilde{m} > m$, then the true source region must lie inside the estimated source region. By these notations, the set $T - A$ contains all inactive sources in the ground-truth model (highlighted in the blue color),

and $E - T$ (green area) represents possible Granger causality that occurred in estimated sources that do not exist in the ground-truth model. The circles $\circ$ denotes the predicted nonzero GC (nonzeros in $\hat{F}$), and the black circles are TP and while the red circles are FP. The cross signs denote the predicted zero GC (zeros in $\hat{F}$), and the red crosses are FN while the black crosses are TN.

Hence, when we evaluate an estimated GC matrix $\hat{F} \in \mathbf{R}^{\tilde{m} \times \tilde{m}}$, the following properties hold on the regions shown in Figure 8.

- TP and FN only exist inside the active regions because nonzeros in F in a ground-truth model can only exist in this region.
- TPR is equal in all regions because the numbers of TP are equal in all regions.
- If all active sources are correctly classified then there is no FP in the true source region and the estimated source region.
- Predicted nonzeros in the green region are regarded as FP since there are no true sources there.
- A fair comparison should be tested on the true source region.
- ACC and TNR between regions cannot be compared because the numbers of negatives are different in those regions.
- FP and FN on the estimated source region can only evaluated when a method is tested on a simulated data sets as the ground-truth models and hence the true source region are known.

From above reasons, the performance on the active true source region reflects how high the method can achieve in TPR. An overall performance of a method can be worse when evaluated on the true source region since if the method predicts any nonzero in the inactive source region, it must be FP. A good method should yield a high TNR on the blue area. Lastly, the performance evaluated on the estimated source region can only drop if the method introduces unnecessary predicted nonzeros in the green area. This arises from two possibilities: error from the source selection algorithm or error from learning significant GC entries.

5 Simulation Results

This section illustrates experimental results of each process in our scheme shown in Figure 3. In what follows, we refer $F_{ij} = 0$ as null and $F_{ij} > 0$ as causal entries.

5.1 Generating EEG data

The parameters of ground-truth models of source signals according to (1a) and (1b) are generated from sparse VARMA models of order $(3, 2)$ described in Section 4.1. The lead field matrix, L, is computed based on the three-shell spherical head model with ICBM152 anatomical template using *brainstorm toolbox* [T$^+$11]. We select 28 number of brain sources from regions of interest (ROIs) based on Automated Anatomical Labeling (AAL) template relied on T1 MRI image from MNI152 template. Some of ROIs associates with emotional memory pathways including Amygdala, Frontal area, Motor cortex and Occipital cortex. Moreover, additional ROIs including Angular, Insula, Lingual, Putamen and Thalamus are added. The EEG channels are in 10-10 and 10-20 placement system, respectively. Center of mass for each ROI is chosen to be the position of sources. The unit of lead field matrix L is microvolts per nanoamp-meter, $\mu V/(nA - m)$; the unit of EEG data is microvolts, μV, and the unit of source signals is nanoamps-meter, $nA - m$. The noise variance of w, η and v are $10^{-4}, 10^{-4}$ and 10^{-2} respectively. The number of state variables (or latents z), sources, and EEG channels in the ground-truth models are denoted by n, m, r, respectively. In the model estimation process, m and n must be set and we use notations of $\tilde{n}, \tilde{m}$. Therefore, $\tilde{L}$ also denotes the lead-field matrix used in the estimation process which has size of $r \times \tilde{m}$.

5.2 Granger causality estimation of state-space models

Experiment setting. We generate time series data from a VARMA model with $n = 20, m = 15$ and 1000 data points. The generated ground-truth models can be categorized into two groups: weak and strong causal models. The weak causal model represents the causality that have small values in F and the strong causal model corresponds to a relatively high values in F. Hence, we expect that strong causality should be more easily captured than the weak causality. In this experiment, the number of trials (N_0) is $20,000$ and the number of samples to calculate $\bar{F}$ (N_1) is 20. Therefore, the number of $\bar{F}$ matrices (N_2) is $1,000$ and we use 10-fold cross validation to split the data set into training and test sets. We perform model estimation The model parameters are estimated using training data set and by the subspace identification method explained in Section 4.2. The computation of F and learning its significant entries follows the details in Section 4.5. The number of GMM components is varied from 1 to 10 and is chosen based on three scores: BIC, relative BIC and silhouette score. After we obtain a GMM model used for clustering entries in F, we evaluate the accuracy on the test set.

Result. The histogram of vectorized F in Figure 6 appears to have multi-modal shape where the number of modes (around 10) suggests us to vary GMM components from 1 to 10. Three criterion choose different numbers of GMM components as shown in Table 1. Silhouette score tends to choose less components, while BIC is prone to choose more components and the relative BIC performs in between. Table 2 shows that GMM model performs a clustering in a good accuracy when the number of components is suitably chosen and this is obtained by using the relative change of BIC as a criterion. We achieve more accuracy when the ground-truth models have strong causality since the GMM components of high mean can be well separated from the first component. When comparing the accuracies between two methods of classifying null from causal entries, we find that the clustering method based on comparing posterior probabilities given all estimated Gaussian components performs better than using a threshold obtained from the log-likelihood ratio test between the first two Gaussian distributions fitted by GMM.

Table 1: The number of mixture components selected by `BIC`, `rBIC` (relative change in BIC), and `Silh` (Silhouette score).

	$N_0 = 2000$			$N_0 = 10000$		
Ground-truth model	BIC	rBIC	Silh	BIC	rBIC	Silh
Weak causality	6-9	4-7	2	7-10	4-7	2
Strong causality	6-8	3-5	2-6	6-10	3-6	2-7

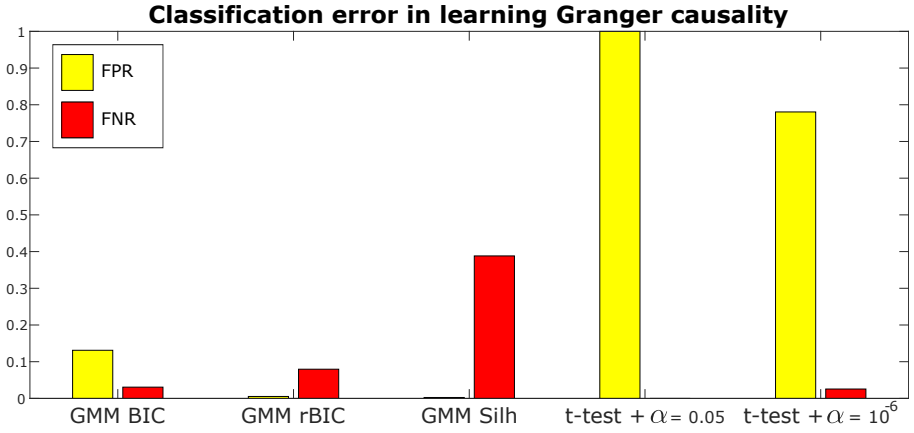


Figure 10: A comparison of Granger causality detection errors between the GMM and t-test.

Figure 10 shows the classification errors in details and a comparison with t-test. As in average sense, Table 1 suggests that BIC tends to choose a highest number of GMM components, it could create an unnecessarily more Gaussian modes capturing small entries of $\bar{F}$. Therefore, it happens that we

Table 2: The accuracy of classifying Granger causality as *null* or *causal* tested on data generated from two types of ground-truth models. The approach is based on clustering method using GMM where the number of mixture components is selected by `BIC`, `rBIC` (relative change in BIC), and `Silh` (Silhouette score).

$N_0 = 2000$

		Weak causality	**Strong causality**
BIC	Thresholding	82.20	96.74
	Clustering	91.40	98.78
rBIC	Thresholding	92.22	97.41
	Clustering	**92.85**	**99.15**
Silh	Thresholding	64.02	83.68
	Clustering	71.90	85.83

$N_0 = 10000$

		Weak causality	**Strong causality**
BIC	Thresholding	82.20	96.74
	Clustering	85.25	99.20
rBIC	Thresholding	92.22	97.41
	Clustering	**92.90**	**99.15**
Silh	Thresholding	64.02	83.68
	Clustering	73.75	85.83

misclassify some entries as causal while it is supposed to be null, leading to the highest false positive in this case. On the other hand, the silhouette score chooses the lowest number of GMM components, so the model lacks of flexibility to explain detailed characteristics in multi-modal shapes of the histogram. Then it is likely to misclassify causal entries as null, resulting in a high false negative. The performance of GMM using relative change of BIC is therefore in between the two previous methods.

If we consider a conventional significance test as t-test, we found that the t-test in element-wise manner is not suitable for testing $H_0 : F_{ij} = 0$ and $H_1 : F_{ij} > 0$ via computing the t-score as $t = \bar{F}_{ij}/\mathrm{SD}(F_{ij})/\sqrt{N_1}$ where N_1 is the number of samples used to compute $\bar{F}_{ij}$. Theoretically, GC matrix is always nonnegative, $F \geq 0$ but due to estimation error, the estimated F_{ij} is always perturbed from zero even the (i, j) is truly a null entry. From the t-score calculation, we find that $\mathrm{SD}(F_{ij})$ is small and even smaller when N_1 is large, so the score can be very high. Consequently, the test result from t-test reject H_0 most of the times, which reported in a high FPR, until the significant level α is very low as shown in Figure 10. In conclusion, t-test should not be practically applied as the true distribution of F_{ij} is not normal.

5.3 Selecting active sources

In this experiment we show the performance of classifying active sources in two cases: i) $m = 28, \tilde{m} = 28, r = 65$, and ii) $m = 28, \tilde{m} = 28, r = 19$. These two cases show how the method performs when the number of EEG sensors, r, increases (more measurement data are obtained) while assuming that the number of sources, m is known. The objective of this experiment is to estimate model parameters in (1b) that promote sparsity pattern in *rows* of C. The estimation procedure consists of two steps: a state-space estimation of EEG time series as explained in Section 4.2 and the estimation of C given in Section 4.3.

Simulated EEG data. The parameters of ground-truth source models are generated from the sparse VARMA(3,2) models with dimension $n = 30, m = 28$ and only 10 sources are active. We consider two EEG placement systems: i) 10-10 EEG placement when $r = 65$ and ii) 10-20 EEG placement when $r = 19$. The noise variance of w, η and v are $10^{-4}, 10^{-4}$ and 10^{-2} respectively. Ten different ground-truth models were used to generate time series trials in this experiment.

Experiment setting: In the estimation process, we need to set the model configuration in two cases: *Case I.* $\tilde{m} = 28$ and $r = 65$: the number of estimated sources is less than the number of EEG sensors

and *Case II*. $\tilde{m} = 28$ and $r = 19$: the number of estimated sources is greater than the number of EEG sensors. Varying λ in the formulation (19) gives us various sparsity patterns of rows in $\hat{C}$. In all 10 trials, the range of λ chosen by BIC lies approximately in $[10^{-6}\lambda_{\max}, 10^{-4}\lambda_{\max}]$.

Results Figure 11 shows an example of estimated C. The left bar shows the zero pattern of rows in the ground-truth C, while the middle figure shows the zero pattern of $\hat{C}$ using 65 EEG sensors ($r = 65$) and the right figure shows $\hat{C}$ when using 19 EEG sensors ($r = 19$). The true active sources correspond to nonzero row vectors C_i^T where $i =$ 1,3,5,6,12,14,18,20,25 and 28. If r is large (using more EEG channels), rows having strong magnitudes of $\hat{C}_i^T$'s are similar to the pattern of the true active sources shown in the left bar of Figure 11. However, $\hat{C}$ still contains falsely nonzero rows (and have very small coefficients) that do not exist in the ground-truth C; regarded as spurious active sources.

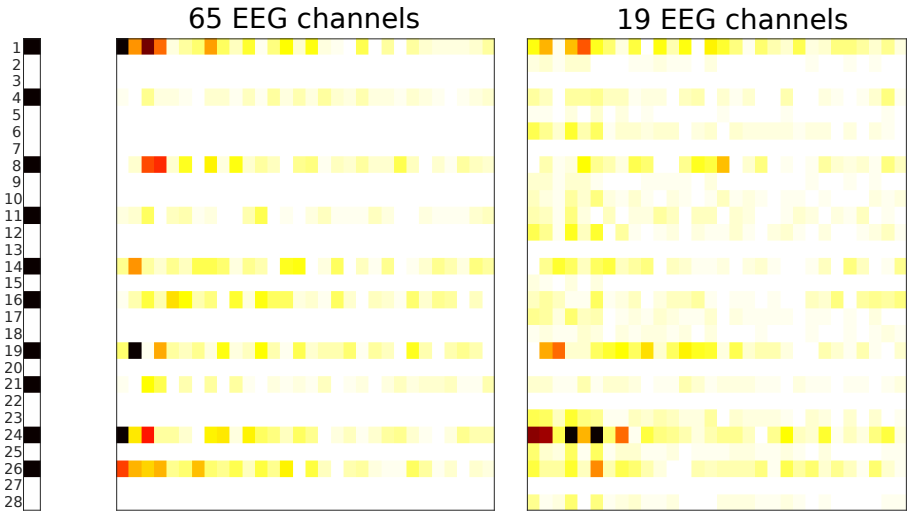


Figure 11: Example of zero patterns of estimated C. The color scale is proportional to magnitudes of C_{ij}'s (white color refers to magnitude of zero). The black rows in the left vertical bar indicate the nonzero rows of the true C.

Classification performance of the method which depends on λ is shown in Figure 12. Each point on ROC curves refers to a classification result from a value of λ, where TP (TN) correspond to correctly identified nonzero (zero) rows in $\hat{C}$. The bottom left and top right corners of ROC correspond to $\lambda = \lambda_{\max}$ (sparsest $\hat{C}$) and $\lambda = 0$ (densest $\hat{C}$) respectively. Using more EEG channels yields a better classification performance (we can almost reach 100% accuracy by some value of λ) and this can be explained from the formulation (19) since r refers to the number of rows in H, equivalent to the number of samples in a regression problem. However, λ suggested from BIC does not yield the best performance on ROC curve , but it tends to choose the λ that provide a relatively denser solution in $\hat{C}$. Even though the pattern of estimated $\hat{C}$ in Figure 11 is similar to the true active sources, but it still contains very small coefficients. This result may occur from the estimation process in subspace identification, which contain estimation errors in H. In conclusion, our method can be used to select the active sources using estimated model parameters from subspace identification. The more number of EEG channels can help to improve the performance of our method.

5.4 Learned Granger causality

This experiment aims to show overall performance of our method that combines all the steps including state-space estimation, source selection, noise covariance estimation, learning significant GC patterns shown in Figure 3. We consider a realistic scenario where the number of estimated source could be larger than the number of true source $\tilde{m} > m$ and see if the method can disregard the sources that do not exist in the ground-truth model.

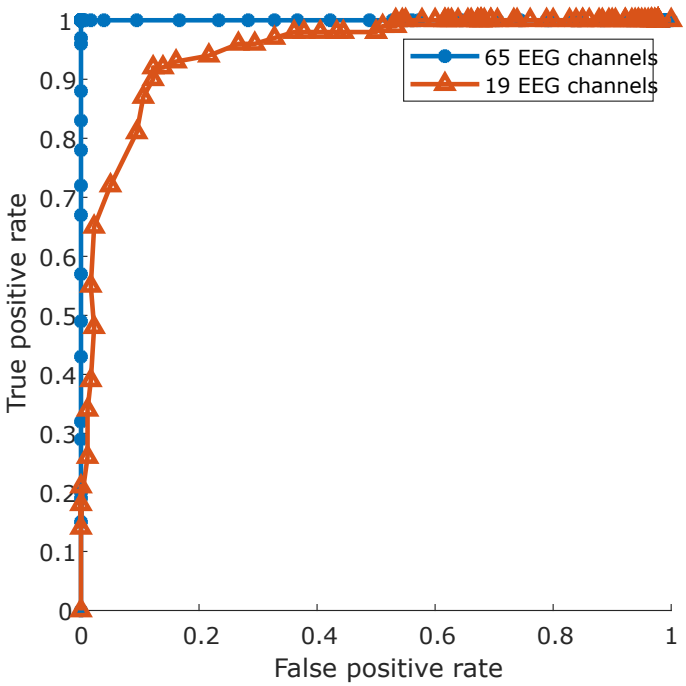


Figure 12: Receiver operation characteristic (ROC) of active and inactive source classification under various settings in number of EEG channels (r).

Experiment setting: Firstly, ground-truth sparse VARMA(3,2) models are generated with dimensions of $n = 15, m = 10, r = 25$. The density of nonzeros in the ground-truth GC matrices are varied to 25%, 50% and 70% (by controlling sparsity of AR coefficients in VARMA models.) These three density values refers to ground-truth model 1,2, and 3, respectively. The number of active and inactive sources are both set to be 5. The EEG sensor is based on the extended 10-20 system. Variance of noises $w, \eta(t)$ and v are set to be 10^{-2}, 10^{-2} and 10^{-4} respectively. Assume that all active sources in ground-truth models are contained in estimated sources. The lead-field matrix L is calculated from the realistic head model using *brainstorm toolbox*. Active sources are randomly selected from major ROIs including Amygdala, Angular, Frontal lobes, Hippocampus, Lingual, Occipital, Motor area, Thalamus, Frontal, Parietal and Temporal. The position of EEG sensors follows the EEG 10-20 system. We generate $10,000$ trials of EEG time series having time points of 1000, from the three ground-truth models. Following our estimation methodology, we obtain $10,000$ estimated Granger causality matrices to be evaluated. When learning significant GC by GMM, the number of GMM components is choosen from the relative change of BIC.

Result: In estimation process, it requires assuming $\tilde{n}$ and $\tilde{m}$ (dimensions of z, latents, and x, sources). For $\tilde{n}$, it is chosen from the value that achieves a high average fitting over EEG 25 channels and from many trials of time series. Figure 13 shows that the number of estimated latents that provides the best fitting are in range from 4 to 7. As a result, we set $\hat{n} = 5$.

The performance of GC estimation follows the description in Section 4.6 where the true GC matrix is extended to have same size as the estimated GC matrix ($\tilde{m} \times \tilde{m}$) shown in Figure 14 so that we can make a comparison with the estimated GC. It shows that strong causality in the ground-truth model are observed in estimated model, *i.e.*, darker dots in the estimated F appears in the same location as those in the true F. The white area of the true F in the middle column contain no nonzeros entries of F because there are no true sources there. We see from the right column that our method of source selection performs well as it never detects nonzero in that region. However, we observe both misclassified zeros and nonzeros in active source regions, then we show numerical overall performances of learning GC from three ground-truth models in Table 3 where we can discuss performance results in three aspects: i) choice of performance measures, ii) the density of ground-truth models and iii) the evaluation region. When evaluated on the true source and estimated source regions, our method

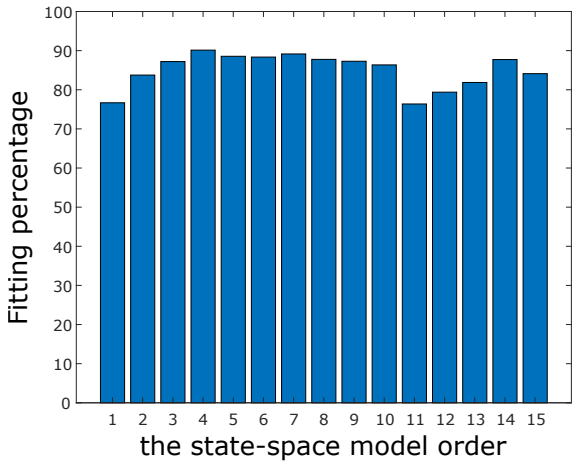


Figure 13: Example of fitting percentage of estimated model from Subspace identification averaged over 25 EEG channels.

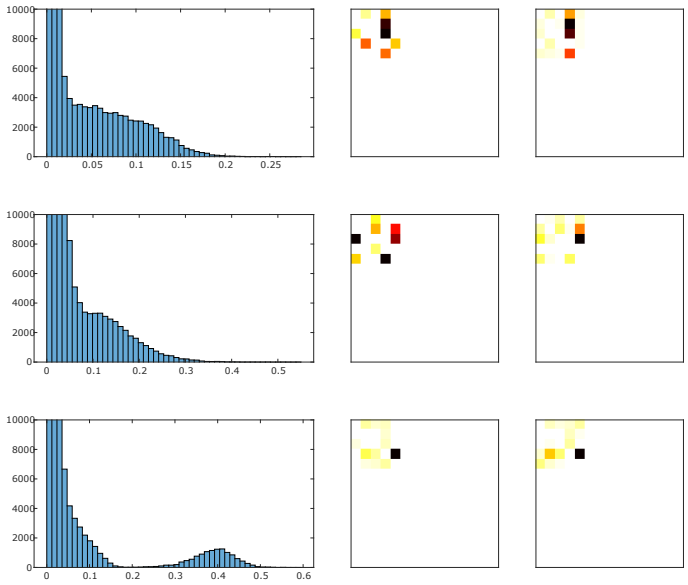


Figure 14: Example of GC learned from simulated EEG data over 10,000 trials. *(Left column.)* The histograms of vectorized GC matrices. *(Middle column.)* The GC pattern of the ground-truth model. *(Right column.)* The average of estimated GC patterns.

achieves TNR higher than TPR for model 2 and 3 (sparser models) which means the method tends to predict non-causality better. The density of ground-truth GC affects the results in the way that the method is prone to perform better when the ground-truth models are sparse. The nominal performance should be evaluated on the true source region and we achieve ACC = 88.19%, TPR = 80.05% and TNR = 89.02%. Considering how the performance is changed on the estimated source region, we first note that TPRs must be the same. The nominal value of TNR is improved when evaluated on the estimated source region; showing that our method of source selection does not create much of spurious causality in the region that does not contain the true sources. If we focus on the active source region, the drop of TNR from its nominal can suggest us that the method of learning significant GC still requires some room for improvement as it appears that portions of FP are created from the method in the active source area. The averaged performances over all models is also illustrated in Figure 15.

Table 3: The averages (%) of accuracy (ACC), true positive rate (TPR), and true negative rate (TNR) of estimated Granger causality patterns over $120 - 180$ trials.

Models	Estimated source region			True source region			Active source region		
	ACC	TPR	TNR	ACC	TPR	TNR	ACC	TPR	TNR
Model 1	75.95	82.97	75.69	75.89	82.97	75.28	73.53	82.97	69.15
Model 2	98.42	87.50	98.82	96.48	87.50	97.23	85.80	87.50	85.00
Model 3	96.53	70.77	98.11	92.18	70.77	95.38	68.73	70.77	66.53
Averaged	90.11	80.05	90.55	88.19	80.05	89.02	77.10	80.05	75.35

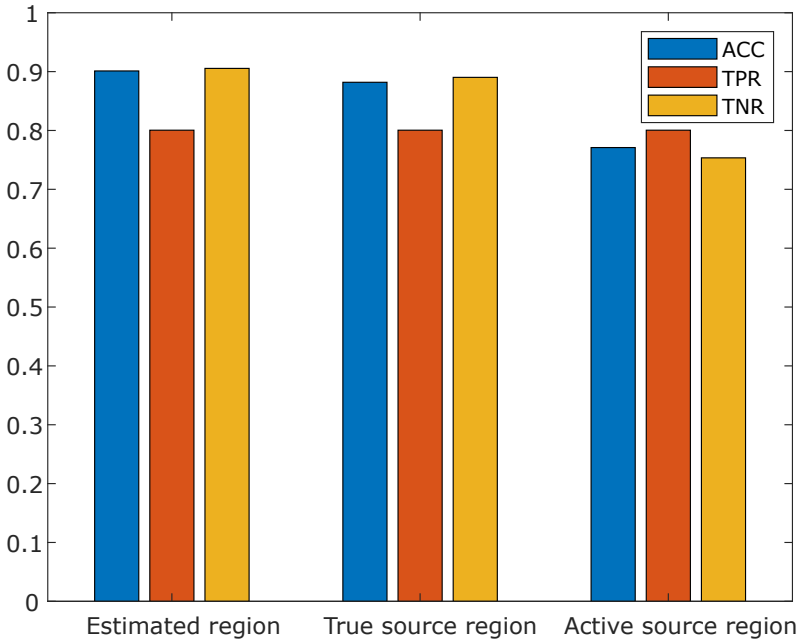


Figure 15: The average performance measures over from the three ground-truth models.

6 Application to real EEG data

In this section, we perform an experiment on real EEG data sets and compare the findings with the previous studies that also explore brain connectivities on this data set with other modalities, since the true connectivity is unknown.

Data description. We consider on task-EEG data containing a steady state visual evoked potential (SSVEP) EEG data. The data are recorded from a healthy volunteer with flickering visual stimulation at 4 Hz using extended 10-20 system with 30 EEG channels. EEG data were recorded total 298 seconds and contain three blocks of stimulation and each of stimulation blocks contains 44.7 seconds. Each of EEG data trial is sampled using $1,000$ data points from stimulation period (three blocks with $11,126$ data points for each block), so we obtained 30 EEG data trials. More information about this data set can be found in [DEK+11, PLGMBB+18].

Experiment setting. In estimation process, the state variable dimension (n) is selected from the suggested model order selection in subspace identification toolbox (`n4sid`). The lead field matrix is assumed to be computed from realistic SPM human head model. The selection of brain sources follows the details in [PLGMBB+18] which includes the most actively ranked generators of Occipital lobe, Temporal lobe and Frontal lobe. We sample three sources from each of six ROIs including

- left Occipital lobe (OL-L), right Occipital lobe (OL-R),
- left Temporal lobe (TL-L), right Temporal lobe (TL-R),

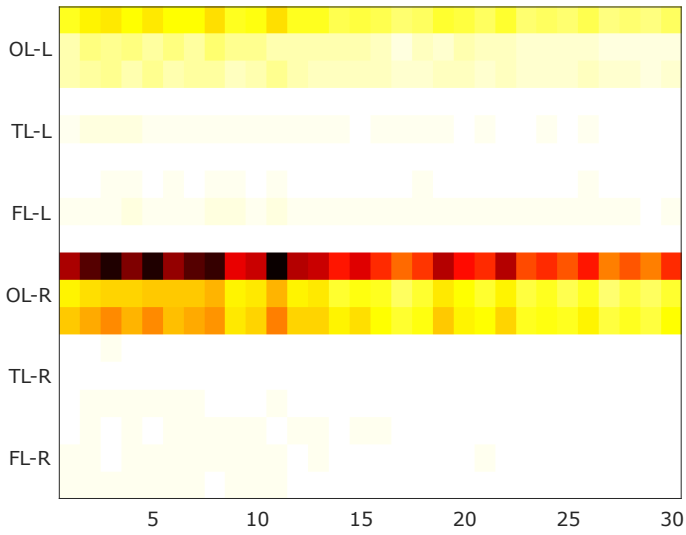


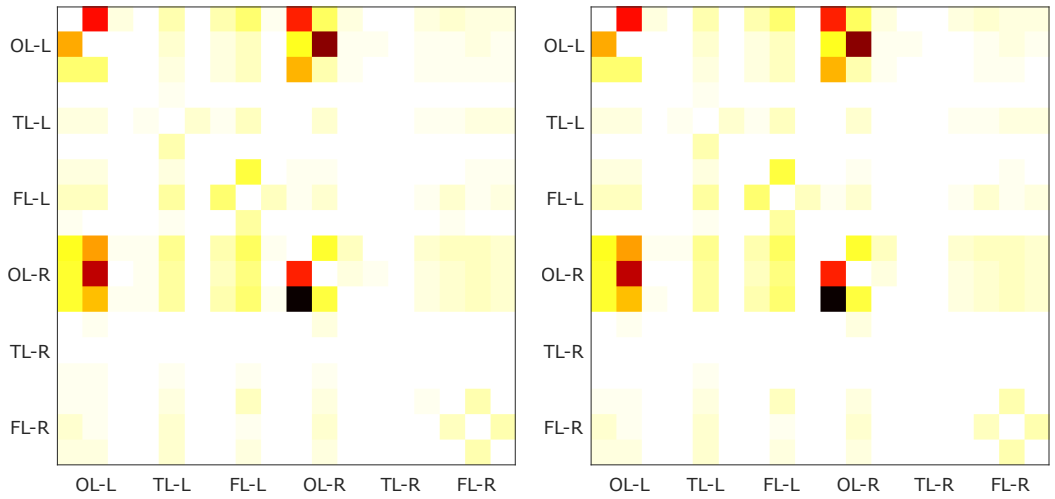
Figure 16: A color scale of $\hat{C}$ from source selection process using λ chosen from BIC. The estimated $\hat{C}$'s are averaged over 30 trials.

- left Frontal lobe (FL-L), right Frontal lobe (FL-R),

then total number of considered sources $(\tilde{m})$ is $3 \times 6 = 18$.

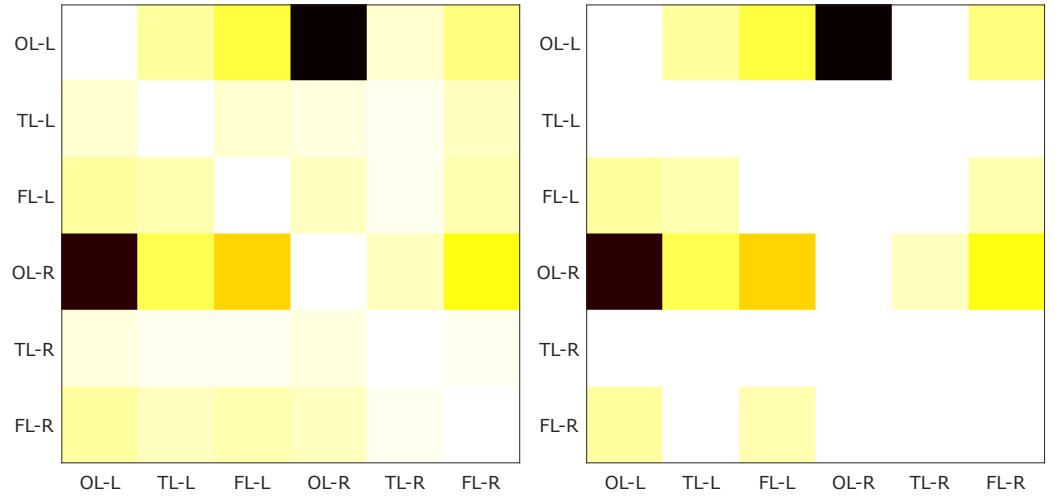
Result. We expect to observe the relationship between OL and OR from using SSVEP data. Figure 16 shows the performance of source selection where each row correspond to each source in the ROI. Our method learned that the most active sources are from occipital lobes (OL-L and OL-R). The finding that activities from occipital lobes are outstanding from other sources is no surprising as SSVEP EEG data were obtained in a paradigm that visual cortices were stimulated where it is known that OL is an area responsible for visual processing center. This agrees with [DEK+11] that showed that the activated regions occured in the visual cortex area. We validate our connectivity result with [PLGMBB+18] which also revealed that the strong causality between OL are observed in SSVEP data. However, causalities in the areas of FL and TL are rarely observed in Figure 16. Moreover, we observe that some of FL is a Granger cause to OL as shown in rows of OL-L and OL-R corresponding to columns of FL-L and FL-R. However, a causality from TL is not detected in our result.

Finally, the average Granger causality based on ROIs is shown in Figure 18. The ROI-based connectivity matrix reveals the strongest causality is found between occipital lobes as expected from visual stimulation. Moreover, a causality between FL and OL is detected. To interpret this result, we found that the connectivity between OL and FL also existed when using other modalities such as the human SSVEP fMRI in [S+07] and rat SSVEP EEG in [L+15].



(a) average of estimated GC matrices. (b) significant GC matrices.

Figure 17: Average of estimated GC from SSVEP EEG data.



(a) average estimated ROI-based GC. (b) significant ROI-based GC.

Figure 18: Average of estimated GC ROI-based from SSVEP EEG data. *(Left.)* Before clustering process by GMM. *(Right.)* After after clustering process by GMM.

7 Conclusion

This project studies about an estimation of linear dynamical models for EEG time series and aims to use the model parameters to infer causalities among source signals, or brain activities, in a human brain. The model equations explain coupled dynamics of source signals and scalp EEG signals where only EEG can be measured. The definition of relationships among variables follows the idea of Granger causality (GC) that has been well-established and applied on vector autoregressive (VAR) models. This work extends the VAR models to a more general class, a state-space equation which can be equivalently transformed to a vector autoregressive moving average (VARMA) model. The GC characterization on state-space equation is highly nonlinear in system parameters and can be numerically computed via solving the discrete Riccati algebratic equation. The result is called a GC matrix of size $m \times m$ where m is number of sources, the variables we aim to find a connectivity among them.

In order to estimate such GC matrices, we have proposed a statistical learning scheme consisting of i) state-space estimation using subspace identification, ii) source selection to classify inactive from active sources, iii) estimation of noise covariances, and iv) learning significant entries in the GC matrices.

In estimation process, one needs to assume model dimensions which may not correctly agree with that of the ground-truth model, we also propose descriptions of evaluation method in different cases. The main contributions of this work are an estimation formulation of classifying active sources based on ℓ_{21}-regularized regression, a formulation of estimating noise covariances based on a semidefinite programming, a scheme of learning significant entries of GC matrix based on Gaussian mixture model (GMM), and lastly, a combined scheme of all four procedures.

The estimation of mapping from latents to source, or our source selection process performs acceptably well according to the obtained ROC curve. The highest performance can be achieved if the penalty parameter λ is chosen appropriately, while currently, applying a model selection criterion, BIC, to choose λ may not yield the optimal performance yet. The performance is shown to be improved if one uses a higher number of EEG channels as it corresponds to having more data samples in the estimation. The scheme of learning GC significance using GMM requires many EEG trials to obtain multi-trial of estimated GC matrices, so that its sample mean can be approximated by a Gaussian distribution. As a result, GMM can be used to cluster different modes in the vectorized GC matrix. The results showed that clustering insignificant entries using posterior probabilities achieves the accuracy of $92 - 99\%$ for a moderate sample size setting and it is obtained when the number of GMM modes is chosen by the relative change of BIC. The overall performance when combining all the procedures achieve the accuracy of $75 - 96\%$ when evaluated on the true source region. The accuracies can be vary upon the density of sparsity level in the ground-truth model. Moreover, we conclude that our source selection method is likely to detect inactive source correctly as TNR is not deteriorated from the true source region to the estimated source region (where there is no source there.) However, our scheme of learning GC significance using GMM could be further improved since the overall performance decreases (due to a drop in TNR) when evaluating on the active source region. The performance of our method on real data set is evaluated on SSVEP EEG data whose setting is to stimulate human brain in visual cortex area. One of our results is consistent to this setting and previous studies in the sense that a strong causality is found between occipital lobes which are known to be related to a task of visual processing.

Many practical concerns and limitations of the method can be concluded. Firstly, it requires an approximate of the lead-field matrix (L) which needs information about sensor position, source position, and a head model. In our opinion, the latter appears to be most uncertain parameter as different subjects would correspond to different head models but this information is unlikely to be exactly known. Secondly, the clustering process using GMM requires multi-trial of EEG time series, which may not be easily obtained in practice. If one has a few trials of time series with a certain length, it is more beneficial to use long data points to improve estimation results (in subspace identification and source selection process), rather than chopping data in to several trials so that GMM can apply. In many situations, it is also possible to obtain a long multi trial of EEG data as the sampling rate of EEG is very high.

References

[ACM+07] L. Astolfi, F. Cincotti, D. Mattia, M. G. Marciani, L. A Baccala, F. de Vico Fallani, S. Salinari, M. Ursino, M. Zavaglia, L. Ding, et al. Comparison of different cortical connectivity estimators for high-resolution EEG recordings. *Human brain mapping*, 28(2):143–157, 2007.

[BS11] L. Barnett and A.K. Seth. Behaviour of Granger causality under filtering: theoretical invariance and practical application. *Journal of neuroscience methods*, 201(2):404–419, 2011.

[BS14] L. Barnett and A. K. Seth. The MVGC multivariate Granger causality toolbox: a new approach to Granger-causal inference. *Journal of neuroscience methods*, 223:50–68, 2014.

[BS15] L. Barnett and A. K. Seth. Granger causality for state-space models. *Physical Review E*, 91(4):1–6, 2015.

[CGHJ12] J. Casals, A. García-Hiernaux, and M. Jerez. From general state-space to VARMAX models. *Mathematics and Computers in Simulation*, 82(5):924–936, 2012.

[CRTVV10] Bing Leung Patrick Cheung, Brady Alexander Riedner, Giulio Tononi, and Barry D Van Veen. Estimation of cortical connectivity from EEG using state-space models. *IEEE Transactions on Biomedical engineering*, 57(9):2122–2134, 2010.

[CWM12] J. Chiang, Z. Jane Wang, and M. J. McKeown. A generalized multivariate autoregressive (GMAR)-based approach for EEG source connectivity analysis. *IEEE Transactions on Signal Processing*, 60(1):453–465, 2012.

[DEK+11] A.D. Duru, S.B. Erdogan, I. Kasikci, A. Bayram, A. Ademoglu, and T. Demiralp. Investigaton of the neuronal efficacy and EEG source power under steady-state visual stimulation. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 6576–6579. IEEE, 2011.

[dSFK+16] F. Van de Steen, L. Faes, E. Karahan, J. Songsiri, P.A. Valdes-Sosa, and D. Marinazzo. Critical comments on EEG sensor space dynamical connectivity analysis. *Brain Topography*, pages 1–12, 2016.

[GHAEC08] G. Gómez-Herrero, M. Atienza, K. Egiazarian, and J. L. Cantero. Measuring directional coupling between EEG sources. *Neuroimage*, 43(3):497–508, 2008.

[GPO12] R. E. Greenblatt, M. E. Pflieger, and A. E. Ossadtchi. Connectivity measures applied to human brain electrophysiological data. *Journal of Neuroscience Methods*, 207(1):1–16, 2012.

[Gra69] C. WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 37:424–438, 1969.

[Ham94] J. D Hamilton. *Handbook of Econometrics*, volume 4. Elsevier, 1994.

[Hau12] S. Haufe. *Towards EEG Source Connectivity Analysis*. PhD thesis, Technische Universität Berlin, Germany, 2012.

[HD+14] M. Hassan, O. Dufor, et al. EEG source connectivity analysis: from dense array recordings to brain networks. *Brain Topography*, 9(8):1–15, 2014.

[HE16] S. Haufe and A. Ewald. A simulation framework for benchmarking EEG-based brain connectivity estimation methodologies. *Brain Topography*, pages 1–18, 2016.

[HNMN13] S. Haufe, V. V Nikulin, K. Müller, and G. Nolte. A critical assessment of connectivity measures for EEG data: a simulation study. *Neuroimage*, 64:120–133, 2013.

[HTF09] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, 2nd edition, 2009.

[HTN+10] S. Haufe, R. Tomioka, G. Nolte, K. Müller, and M. Kawanabe. Modeling sparse connectivity between underlying brain sources for EEG/MEG. *IEEE Transactions on Biomedical Engineering*, 57(8):1954–1963, 2010.

[HTW15] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall/CRC, 2015.

[KR09] L. Kaufman and P. J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

[L+15] F. Li et al. The enhanced information flow from visual cortex to frontal area facilitates ssvep response: evidence from model-driven and data-driven causality analysis. *Scientific Reports*, 5:1–11, 2015.

[Lüt05] H. Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer Science & Business Media, 2005.

[LWVS15] X. Lei, T. Wu, and P. Valdes-Sosa. Incorporating priors for EEG source imaging and connectivity analysis. *Frontiers in Neuroscience*, 9(284):1–12, 2015.

[M+19] D. Marinazzo et al. Controversies in EEG source imaging and connectivity: Modeling, validation, benchmarking. *Brain Topography*, 32:1–3, 2019.

[MMARPH14] J. Montoya-Martínez, A. Artés-Rodríguez, M. Pontil, and L. K. Hansen. A regularized matrix factorization approach to induce structured sparse-low-rank solutions in the EEG inverse problem. *EURASIP Journal on Advances in Signal Processing*, 19:97, 2014.

[MML+04] C.M. Michel, M.M. Murray, G. Lantz, S. Gonzalez, L. Spinelli, and R. de Peralta. EEG source imaging. *Clinical Neurophysiology*, 115(10):2195–2222, 2004.

[MP04] G. McLachlan and D. Peel. *Finite mixture models*. John Wiley & Sons, 2004.

[OM12] P. Van Overschee and BL. De Moor. *Subspace identification for linear systems: Theory—Implementation—Applications*. Springer Science & Business Media, 2012.

[PB14] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239, 2014.

[PiS18] N. Plub-in and J. Songsiri. State-space model estimation of EEG time series for classifying active brain sources. In *2018 11th Biomedical Engineering International Conference (BMEiCON)*, pages 1–5. IEEE, 2018.

[PiS19] N. Plub-in and J. Songsiri. Estimation of granger causality of state-space models using a clustering with gaussian mixture model. In *To appear in the Proceedings of IEEE International Conference on Systems, Man, and Cybernetics (IEEE SMC)*. IEEE, 2019.

[PLGMBB+18] D. Paz-Linares, E. Gonzalez-Moreira, J. Bosch-Bayard, A. Areces-Gonzalez, M.L. Bringas-Vega, and P.A. Valdes-Sosa. Neural connectivity in M/EEG with hidden hermitian Gaussian graphical model. *arXiv preprint arXiv:1810.01174*, pages 1–34, 2018.

[PS16] A. Pruttiakaravanich and J. Songsiri. A Review on Exploring Brain Networks from fMRI Data. *Engineering Journal*, 20(3):1–28, 2016.

[PZBC17] G. Prando, M. Zorzi, A. Bertoldo, and A. Chiuso. Estimating effective connectivity in linear brain network models. *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 5931–5936, 2017.

[RCV+14] D. La RoccaX, P. Campisi, B. Vegso, P. Csertir, G. Kozmann, F. Babiloni, and F. Fallani. Human brain distinctiveness based on EEG spectral coherence connectivity. *IEEE Transactions on Biomedical Engineering*, 61(9):2406–2412, 2014.

[S+07] R. Srinivasan et al. fMRI responses in medial frontal cortex that depend on the temporal frequency of visual input. *Experimental Brain Research*, 180(4):677–691, 2007.

[Sak11] V. Sakkalis. Review of advanced techniques for the estimation of brain connectivity measured with EEG/MEG. *Computers in biology and medicine*, 41(12):1110–1117, 2011.

[SBB15] A.K. Seth, A.B. Barrett, and L. Barnett. Granger causality analysis in neuroscience and neuroimaging. *Journal of Neuroscience*, 35(8):3293–3297, 2015.

[SC13] S. Sanei and J. A. Chambers. *EEG Signal Processing*. John Wiley & Sons, 2013.

[Sim06] D. Simon. *Optimal State Estimation: Kalman, H_∞ , and Nonlinear Approaches*. John Wiley and Sons, 2006.

[Son13] J. Songsiri. Sparse autoregressive model estimation for learning Granger causality in time series. In *Proceedings of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3198–3202, 2013.

[STOS17] S.B. Samdin, C.M. Ting, H. Ombao, and S.H. Salleh. A unified estimation framework for state-related changes in effective brain connectivity. *IEEE Transactions on Biomedical Engineering*, 64(4):844–858, 2017.

[T+11] F. Tadel et al. Brainstorm: a user-friendly application for MEG/EEG analysis. *Computational Intelligence and Neuroscience*, 2011.

[WTO16] Y. Wang, C. Ting, and H. Ombao. Modeling effective connectivity in high-dimensional cortical source signals. *IEEE Journal of Selected Topics in Signal Processing*, 10(7):1315, 2016.

[YYR16] M. Tarr Y. Yang, E. Aminoff and K. E Robert. A state-space model of cross-region dynamic connectivity in MEG/EEG. *Advances in Neural Information Processing Systems 29*, pages 1234–1242, 2016.

Appendix: Lead field matrix toolbox

A lead field matrix L is a matrix transformation from sources to scalp EEG signals. We describe how to generate lead field matrix from realistic prior information such as MRI template, Brain tissue map and others. Main templates that we used in this work include ICBM template and MNI152 template.

- ICBM (International Consortium for Brain Mapping) template is the average of T1 weighted MRI data from single subject with 27 obtained data. The template is aligned within the stereotaxic space which based on Talairach-Tournoux brain atlas.
- MNI152 (Montreal Neurological Institute) template is an improvement brain template from ICBM by using MRI scans from 152 subjects to locate structures inside the brain.

The *brainstorm toolbox* in MATLAB is used to generate lead field in our experiments. Prior knowledges for lead field matrix calculation are

- *Head model*: the model of human's head from ICBM152
- *Brain tissue map* or *TPM*: the model of brain tissue from ICBM152
- *MRI template*: we use T1 images from MNI152 template
- *Sensor placement system*: the position of sensor placement from 10-20 system
- *ROI template*: the template of regions of interest based on Automated Anatomical Labeling (AAL) template which relies on T1 MRI image from MNI152 template

Firstly, subject's anatomy information is added including head model and brain tissue map as shown in Figure 19 Consequently, reference MRI data is added to the subject's anatomy information. In this

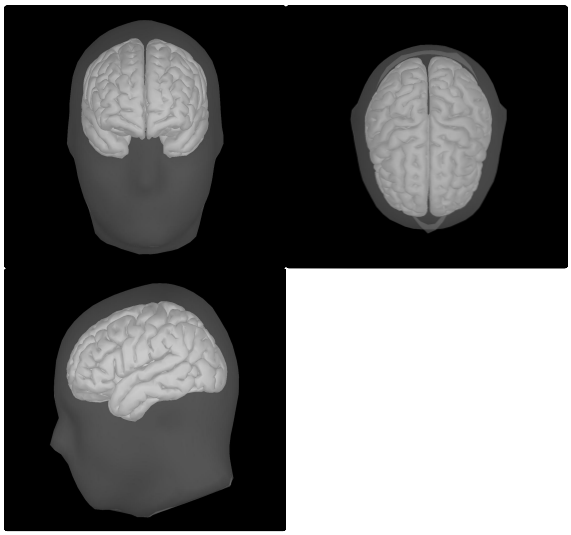


Figure 19: The brain tissue map with a subject head model.

work, we use T1 MRI image from MNI152 which allows us to extract regions of interest (ROI) easily (details are described later). A preprocessing of MRI data, such as slice timing, realignment, is needed for adding MRI data to the subject's information. Moreover, coregistration is performed to adjust coordinates of brain tissue map to MRI data as shown in Figure 20. Next, we fit sensor placement from 10-20 system template to the subject's head model as shown in Figure 21. Then, position of each sensor on subject's head model can be obtained. However, coordinate system of anatomy data is different. The original coordinate system for brain data is *Talairach atlas*. Talairach coordinate system relies on two points: AC (Anterior commissure) point and PC (Posterior commissure) point. *MRI* and MNI coordinate system are used to index voxels in the space of MRI volume. SCS (Subject Coordinate

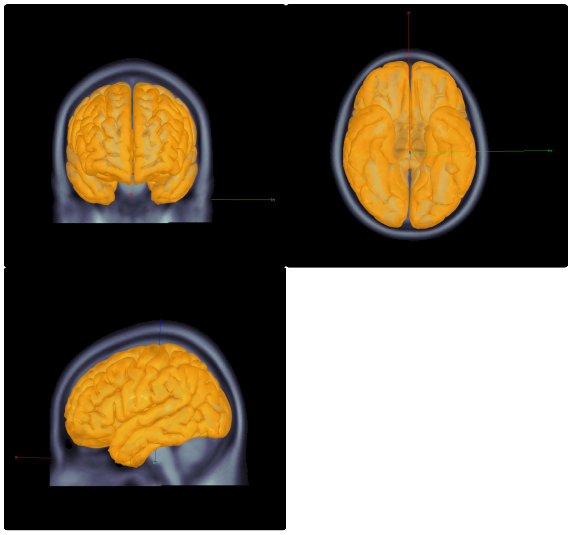


Figure 20: The result of coregistration brain tissue map with the MRI data.

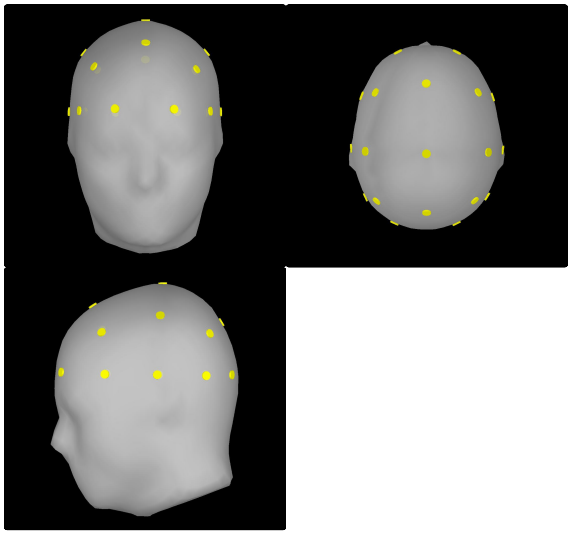


Figure 21: 10-20 system sensor placement based on the subject's head model.

system) is the coordinate system based on Nasion, LPA (left pre-auricular point) and RPA (right pre-auricular point) of subject. MRI data relies on MNI coordinate system but sensor placement system is associated with SCS coordinate system. The position of source is obtained from *Marsbar toolbox*, an additional toolbox in SPM which can extract ROI position easily. ROI data is based on avg152T1 MRI template and the coordinate system relies on MNI coordinate system. The coordinate transform is needed to transform from MNI coordinate system to SCS coordinate system. Consequently, positions of all sources are added then we can compute the lead field matrix from all information.

The result is 3-dimensional lead field matrix gain, $\mathcal{L}$, which describe the propagation of each source to each sensor taking a form of

$$\mathcal{L} = \begin{bmatrix} (L_x)_{11} & (L_y)_{11} & (L_z)_{11} & \cdots & (L_x)_{1m} & (L_y)_{1m} & (L_z)_{1m} \\ (L_x)_{21} & (L_y)_{21} & (L_z)_{21} & \cdots & (L_x)_{2m} & (L_y)_{2m} & (L_z)_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ (L_x)_{r1} & (L_y)_{r1} & (L_z)_{r1} & \cdots & (L_x)_{rm} & (L_y)_{rm} & (L_z)_{rm} \end{bmatrix} \tag{30}$$

where m is a number of sources, r is a number of EEG sensors and subscript x, y, z are the direction

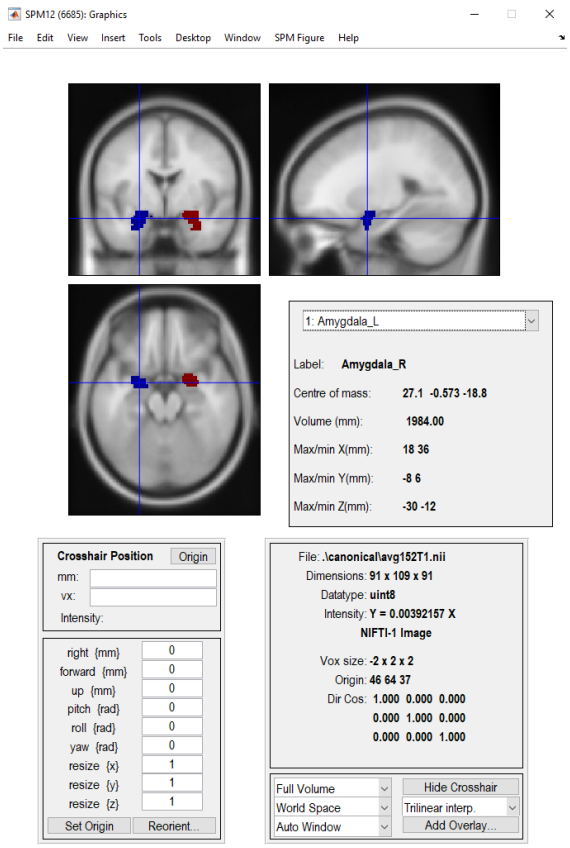


Figure 22: An example of ROI data from marsbar toolbox in SPM12.

of electrical propagation in each axis. The forward problem is described by

$$y(t) = Lx(t),$$

$$\begin{bmatrix} y_1(t) \\ y_2(y) \\ \vdots \\ y_r(t) \end{bmatrix} = \begin{bmatrix} L_{11} & L_{12} & \cdots & L_{1m} \\ L_{21} & L_{22} & \cdots & L_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ L_{r1} & L_{r2} & \cdots & L_{rm} \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(y) \\ \vdots \\ \vdots \\ x_m(t) \end{bmatrix}.$$

where $y_i(t)$ is the ith EEG signal, $x_j(t)$ is the jth source signal and L_{ij} is a lead field gain from jth source to ith EEG sensor. As a result, we can compute the matrix L from $\mathcal{L}$ by using direction normal vectors from each source to each sensor.

$$L_{ij} = \begin{bmatrix} (L_x)_{ij} & (L_y)_{ij} & (L_z)_{ij} \end{bmatrix} \begin{bmatrix} (e_x)_{ij} \\ (e_y)_{ij} \\ (e_z)_{ij} \end{bmatrix} \tag{31}$$

where $e_{ij} = \begin{bmatrix} (e_x)_{ij}^T & (e_y)_{ij}^T & (e_z)_{ij}^T \end{bmatrix}^T$ is the direction normal vector from jth source to ith EEG sensor.

The unit of lead-field gain matrix is volt per amp meter (V/A-m). In general, the unit of EEG data is μV then the suggested unit for lead field matrix is μV/μA-mm.