Bayesian Dictionary Learning for EEG Source Identification

Trine Nyholm Kragh & Laura Nyrup Mogensen
Mathematical Engineering, MATTEK

Master's Thesis

Mathematical Engineering
Aalborg University
http://www.aau.dk

AALBORG UNIVERSITY
STUDENT REPORT

Title:
Bayesian Dictionary Learning for EEG Source Identification

Theme:

Project Period:
Fall Semester 2019
Spring Semester 2020

Project Group:
Mattek9b

Participant(s):
Trine Nyholm Kragh
Laura Nyrup Mogensen

Supervisor(s):
Jan Østergaard
Rasmus Waagepetersen

Copies: 1

Page Numbers: 27

Date of Completion:
October 3, 2019

Abstract:

Here is the abstract

Matematik-Teknologi
Aalborg Universitet
http://www.aau.dk

AALBORG UNIVERSITET
STUDENTERRAPPORT

Titel:
Bayesian Bibliotek Læring for EEG Kilde Identifikation

Tema:

Projektperiode:
Efterårssemestret 2019
Forårssemestret 2020

Projektgruppe:
Mattek9b

Deltager(e):
Trine Nyholm Kragh
Laura Nyrup Mogensen

Vejleder(e):
Jan Østergaard
Rasmus Waagepetersen

Oplagstal: 1

Sidetal: 27

Afleveringsdato:
3. oktober 2019

Abstract:

Her er resuméet

Preface

Here is the preface. You should put your signatures at the end of the preface.

Aalborg University, October 3, 2019

Trine Nyholm Kragh
<trijen15@student.aau.dk>

Laura Nyrup Mogensen
<lmogen15@student.aau.dk>

Danish Summary

Dansk resume ?

Contents

Introduction

Introduktion til hele projektet, skal kunne læses som en appetitvækker til resten af rapporten, det vi skriver her skal så uddybes senere. Brug dog stadigvæk kilder.

- kort intro a EEG og den brede anveldelse, anvendelse indenfor høreapperat.
- intro af model, problem med overbestemt system
- Seneste forslag til at løse dette
- vi vil efterviser dette og udvide til realtime tracking
- opbygningen af rapporten

Chapter 1

Motivation

This chapter examines existing literature concerning source localisation from EEG measurements. At first a motivation for the problem is given, considering the application within the hearing aid industry. Further, the state of the art methods are presented followed by a description of the contribution proposed in this thesis.

things to do: - (DONE) phaselock, activities as one source - understreg vigtigheden af at kigge på sources istedet for scalp - specificering af feedback og test, nedjustering af ambitioner, indrage tid aspektet - fjern at vi har focus på attention decoding - objektive tjek af resultater, sammenlign kendte application anvendt direkte på scalp med dem anvendt på vores løsning, kan det forbedre noget i sig selv - sørge for at det ikke ligner correporation med Eriksholm, ref til hjemmeside - updater idledningen - fokus på sammenhængene afsnit - rette - få styr på brugen af, measurements og dentification/seperation/localisation brug kun seperation og localization

1.1 EEG Measurements

Electroencephalography (EEG) is a technique used within the medical field. It is an imaging technique measuring electric signals on the scalp, caused by brain activity. The human brain consist of an enormous amounts of cells, called neurons. These neurons are mutually connected in neural nets and when a neuron is activated, for instance by a physical stimuli, local current flows are produced [16]. This is what makes a kind of neural interaction across different parts of the brain(?).

EEG measurements are provided by a varies number of metal electrodes, referred to as sensors, carefully placed on a human scalp. Each sensor read the present electrical signals, which are then displayed on a computer, as a sum of sinusoidal waves relative to time.
It takes a large amount of active neurons to generate an electrical signal that is

recordable on the scalp as the current have to penetrate the skull, skin and several other thin layers. Hence it is clear that measurements from a single sensor do not correspond to the activity of a single specific neuron in the brain, but rather a collection of many activities within the range of the one sensor. Nor is the range of a single sensor separated from the other sensors thus the same activity can easily be measured by two or more sensors. Furthermore, interfering signals can occur in the measurments resulting from physical movement of e.g. eyes and jawbone [16]. Lastly the transmission of the electric field through the biological tissue to the sensor has an impacted on the signal, this process is called volume conduction[14, p. 68].

This clarifies the mixture of electrical signals with noise that form the EEG measurements. The concept is sought illustrated on figure 1.1.
It will be clear later that it is of highly interest to separate and localize the sources of the neural activities measured on the scalp. Note that a source do not correspond to a single neuron but is typically a collection of synchronized/phase locked active neurons which are generating a constructive interference resulting in a measurable signal on the scalp(?).

The waves resulting from EEG measurement have been classified into four groups according to the dominant frequency. The delta wave ($0.5 - 4$ Hz) is observed from infants and sleeping adults, the theta wave ($4 - 8$ Hz) is observed from children and sleeping adults, the alpha wave ($8 - 13$ Hz) is the most extensively studied brain rhythm, which is induced by an adult laying down with closed eyes. Lastly the beta wave ($13 - 30$ Hz) is considered the normal brain rhythm for normal adults, associated with active thinking, active attention or solving concrete problems [14, p. 11]. An example of EEG measurements within the four categories are illustrated by figure 1.2.

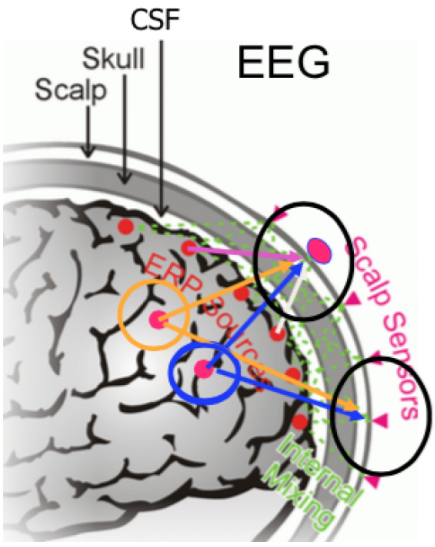

Figure 1.1: Illustration of volume conduction, source [5](we will make our own figure here instead)

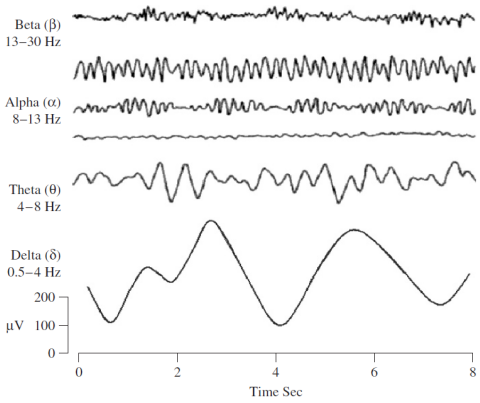

Figure 1.2: Example of time dependent EEG measurements within the four defined categories, source [14]

EEG measurements are widely used within the medical field and especially research of the cognitive processes in the brain. Diagnosis and management of neurological disorders such as epilepsy is one example.
Due to EEG being non-invasive and fast it is widely used to study of the dynamical behaviour of the brain. Neural activity can be measured within fractions of a second after a stimuli has been provided [16, p. 3]. When a person is exposed to a certain stimuli, e.g. visual or audible, the measured activity is said to result from evoked potential.
Over the past two decades, especially functional integration has become an area of interest, that is the interplay between functionally segregate brain areas [15](evt. friston 2011). This concerns the identification and localisation of the single cortical sources causing the united signal measured by a sensor.

The hearing aid industry is one example where this research is highly prioritised. At Eriksholm research center which is a part of the hearing aid manufacture Oticon cognitive hearing science is a research area within fast development. One main purpose is to make it possible for a hearing aid to identify the attended sound source by reading the signal from the brain of the user, which is where the EEG measurements are used [1], [6]. It is essentially the well known but unsolved cocktail problem which is sought improved by use of EEG. The focus of the research considers the correlation between EEG measurements and the sound source rather than identification of the activated source from the EEG. Hence this localisation approach regarding hearing aids is of interest. Furthermore, a real-time application to provide feedback from

EEG measurements would be essential.

When considering the issue of identifying and localising the activated sources from the EEG measurements, one known option is to model the observed data by the following linear system

$$\mathbf{Y} = \mathbf{AX},$$

where $\mathbf{Y} \in \mathbb{R}^{M \times N_d}$ is the EEG measurements from M sensors at N_d data points, $\mathbf{A} \in \mathbb{R}^{M \times N}$ is an unknown mixing matrix and $\mathbf{X} \in \mathbb{R}^{N \times N_d}$ is the actual activation of sources within the brain. The i^{th} column of $\mathbf{A}$ represent the relative projection weights from the i^{th} source to every sensor [5]. From this model the aim is to identify both $\mathbf{A}$ and $\mathbf{X}$ given the measurements $\mathbf{Y}$. This set up is in general referred to as the EEG inverse problem.

To solve the EEG inverse problem the concept of compressive sensing makes a solid foundation including sparse signal recovery and dictionary learning. Independent Component Analysis (ICA) is a common applied method to solve the inverse problem [12], [11], here statistical independence between source activity is assumed.
Application of ICA have shown great results regarding source separation of high-density EEG. Furthermore, an enhanced signal-to-noise ratio of the unmixed independent source time series processes allow essential study of the behaviour and relationships between multiple EEG source processes [8].
However a significant flaw to this method is that the EEG measurements are only separated into a number of sources that are equal or less than the number of sensors. This means that the EEG inverse problem can not be over-complete. That is an assumption which undermines the reliability and usability of ICA, as the number of simultaneous active sources easily exceed the number of sensors [5]. This is especially a drawback when low-density EEG are considered, that is EEG equipment with less than 32 sensors. Improved capabilities of low-density EEG devices are desirable due to its relative low cost, mobility and ease to use.

This makes a foundation to look at the existing work considering the over-complete inverse EEG problem.

1.2 Related Work and Our Contribution

As mentioned above ICA has been a solid method for source localisation in the case where a separation into a number of sources equal to the number of sensors was accepted. To overcome this issue an extension of ICA was suggested, referred to as the ICA mixture model, instead of identifying one mixing matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$ this

approach learns N_{model}(number of sources? or datapoints) different mixing matrices $\mathbf{A}_i \in \mathbb{R}^{M \times M}$. The method was further adapted into the Adaptive Mixture ICA (AMICA) showed successful results regarding identification of more sources than available sensors [13]. However an assumption of no more than M simultaneously active sources has to be made which is still an essential limitation, especially when considering low-density EEG.
Other types of over-complete ICA algorithms have been proposed to overcome the problem of learning over-complete systems. One is the Restricted ICA (RICA), an efficient method used for unsupervised learning in neural networks [17]. Here the hard orthonormal constraint in ICA is replaced with a soft reconstruction cost.
In 2015, [3], O. Balkan et. al. suggested a new approach also targeting the identification of more sources than sensors regarding EEG. The suggested method referred to as Cov-DL is a covariance based dictionary learning algorithm. The point is to transfer the forward problem into the covariance domain which has higher dimensionality than the original EEG sensor domain. This can be done when assuming the scalp mixing is linear and using the assumed natural uncorrelation of sources within a certain time-window. The Cov-DL algorithm stands out from the other straight forward dictionary learning methods as it does not relay on the sparsity of active sources thus it is a further advantage when low-density EEG is considered.
Cov-DL was found to outperform both AMICA and RICA, thus it is considered the state of the art within the area of source identification.

It is essential to note that the Cov-DL algorithm do only learn the mixing matrix $\mathbf{A}$, the projection of sources to the scalp sensors, and not the explicit source activity time series $\mathbf{X}$.
For this purpose a multiple measurement sparse bayesian learning (M-SBL) algorithm was proposed in [4] also by O. Balkan et. al., also targeting the case of more active sources than sensors [4]. Here the mixing matrix which is now assumed known should fulfil the exact support recovery conditions. Though the method was proven to outperform the recently used algorithm M-CoSaMP even when the defined conditions was not fulfilled.

The two state of the art methods for source identification will make the foundation of this thesis. This thesis propose an algorithm which uses the investigated methods on synthetic EEG data and real EEG data. Furthermore, the purpose is to extend the algorithm to perform on EEG measurement in real-time in order to investigate the possibility of providing useful feedback depending on the real-time results of the algorithm. For this, analysis of results in different sound environments such as noisy and noise-less cases and cases of directional noise.
The overall purpose of the real-time performance is to provide results that can be useful to the hearing aid industry, considering the development of self-adaptive hear-

ing aids. By this extension and associated analysis we seek to extend the existing results within the area.

.....
nb. husk at dette kapitel skal vise et helt system og hvor henne i det system vi kigger nærmere og kommer ind med vores bidrag. Det skal gøres klart hvilke områder vi vælger at ligge vores kræfter i.

Vi skal lige have styr på brugen af source identification,localisation og separation.

Chapter 2

Problem Statement

From the motivation 1 and related work it was stated that EEG measurement of the brain activity could be a new contribution within the hearing aid industry to develop hearing aid with better performing in situations as the cocktail party problem. By solving the EEG inverse problem of the low-density EEG device to localise and identifying the sources of the brain activity the results could be used to guide and adapt the hearing aids performance such as move the beamformer in the direction of interest. This lead to the following problem statement.

How can sources of activation within the brain be localised from the EEG inverse problem, in the case of less sensors than sources and how can such algorithm be extended to a real-time application useful within the hearing aid development?

From the problem statement some clarifying sub-questions have been made.

- How can the over-complete EEG inverse problem be solved by use of compressive sensing included domain transformation?
- How can Cov-DL be used to estimate the mixing matrix $\mathbf{A}$ from the over-complete EEG inverse problem?
- How can M-SBL be used to estimate the source matrix $\mathbf{X}$ from the over-complete EEG inverse problem?
- How can the estimates from the over-complete EEG inverse problem be interpreted, regarding hearing aid research.
- How can an application be formed to constitute this source identification process operating in real-time?

Notes: Vi skal have styr på hvad det er vi vil dem vores realtime implementering:

- Måle om der er støj, så vi kan skrue ned for den støj,jeg tror det var det Jan snakked om i første omgang.
- Kæde støjen sammen med locationerne for de aktive sources, måske det man gør i forhold til at retningsbestemme støj?
- Er første del blot at localisere sources?

Chapter 3

Sparse Signal Recovery

Through this chapter an introduction to the concept compressive sensing is given with associated theory which later on will be used in the development of the algorithm with used methods known from compressive sensing to estimate the mixing matrix $\mathbf{A}$ and the sparse source matrix $\mathbf{X}$.

3.1 Compressive Sensing

Compressive sensing is the theory of efficient recover/reconstruct a signal from minimal measurements. This recovery is often described as a linear model/system

$$\mathbf{y} = \mathbf{A}\mathbf{x},$$

which consist of observed data $\mathbf{y} \in \mathbb{R}^M$, a matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$ which models the linear measurements and signal $\mathbf{x} \in \mathbb{R}^N$. In compressive sensing terminology, $\mathbf{y}$ is the signal of interest that is wish recovered from minimal measurements meaning that the signal $\mathbf{x}$ must be sparse.
A signal is said to be k-sparse if the signal has at most k non-zeros coefficient

$$\|\mathbf{x}\|_0 = \text{card}(\text{supp}(\mathbf{x})) \leq k,$$

where the ℓ_0-norm is used. The function card is the cardinality of the support of $\mathbf{x}$. The support of $\mathbf{x}$ is giving as

$$\text{supp}(\mathbf{x}) = \{j \in [N] \ : \ x_j \neq 0\},$$

where $[N]$ a set [9, p. 41]. The set of all k-sparse signals is denoted as

$$\Sigma_k = \{\mathbf{x} \ : \ \|\mathbf{x}\|_0 \leq k\}.$$

It is of interest to recover the signal $\mathbf{y}$ when $M << N$ and $k < M$ [7, p. 8]. This lead to that the matrix $\mathbf{A}$ becomes rank-deficient and therefore have a non-empty

null-space [7, p. ix].

The linear model and finding the sparse signal $\mathbf{x}$ can be written as an optimisation problem

$$\min_{\mathbf{z}} \|\mathbf{z}\|_0 \quad \text{s.t} \quad \mathbf{y} = \mathbf{A}\mathbf{z},$$

where $\mathbf{z}$ is all possible candidates to an k-sparse signal $\mathbf{x}$.
Unfortunately, this optimisation problem is non-convex because of ℓ_0-norm and is therefore difficult to solve – it is a NP-hard problem. Instead by replacing the ℓ_0-norm with its convex approximation, the ℓ_1-norm, the optimisation problem become computational feasible [7, p. 27]

$$\min_{\mathbf{z}} \|\mathbf{z}\|_1 \quad \text{s.t} \quad \mathbf{y} = \mathbf{A}\mathbf{z}, \tag{3.1}$$

and instead we find the best k-term approximation of the signal $\mathbf{x}$.

3.1.1 Conditions on the Mixing Matrix

To ensure an exact or an approximate reconstruction of the sparse signal $\mathbf{x}$ some conditions associated on the matrix $\mathbf{A}$ must be satisfied.

Null Space Conditions

The null space property (NSP) is some necessary and sufficient condition for exact recovery. The null space of the matrix A is defined as

$$\mathcal{N}(A) = \{z \ : \ Az = 0\}.$$

Restricted Isometry Conditions

NSP do not take account for noise and we must therefore look at some stronger conditions which incoperate noise, the following restricted isometry property (RIP)

Definition 3.1 (Restricted Isometry Property)
A matrix A satisfies the RIP of order k if there exists a $\delta_k \in (0, 1)$ such that

$$(1 - \delta_k)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta_k)\|x\|_2^2,$$

holds for all $x \in \Sigma_k$

If a matrix A satisfy RIP then it will also satisfy the NSP as RIP is strictly stronger than NSP.

Theorem 3.1.1
If A satisfies the RIP of order $2k$ with the constant $\delta_{2k} < \sqrt{2} - 1$. Then

$$C = \frac{2}{1 - (1 + \sqrt{2})\delta_{2k}}$$

Coherence

The NSP provide a unique solution to the optimisation problem, (3.1), but is unfortunately complicated to investigate. Instead an alternative measure used for sparsity is presented.
Coherence is a measure of quality and determine if the matrix A is a good choice for the optimisation problem (3.1). A small coherence describe the performance of a recovery algorithm as good with that choice of $\mathbf{A}$.

Definition 3.2 (Coherence)
Coherence of the matrix $A \in \mathbb{R}^{M \times N}$, denoted as $\mu(A)$, with columns $\mathbf{a}_1, \ldots, \mathbf{a}_N$ for all $i \in [N]$ is given as

$$\mu(A) = \max_{1 \leq i < j \leq n} \frac{|\langle a_i, a_j \rangle|}{\|a_i\|_2 \|a_j\|_2}.$$

3.1.2 Multiple Measurement Vector Model

A multiple measurement vector (MMV) model consist of the source matrix $\mathbf{X} \in \mathbb{R}^{N \times L}$ which have $k < M$ rows that are non-zero (the activations of the sources), a observed mixed data matrix $\mathbf{Y} \in \mathbb{R}^{M \times L}$ and a dictionary matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$:

$$\mathbf{Y} = \mathbf{A}\mathbf{X},$$

where L stand for the time samples. From the MMV model the non-zero rows of the source matrix $\mathbf{X}$ are the one of interest that are wanted recovered [**PHD**].

Notes:

- DL recover more sources than sensors $N > M$ assumning the constraint is at any time we have $k < M$. This cause problems for the use on low density system where we have low M.

- Recovery is not possible if $k \geq M$ since any random dictionary is sufficient to represent data points Y using only M basis vectors.
- If the source signal is sparse it is enough just to find the non-zero rows of X denoted by the set S, because then the source signal can be obtained by the psudo-inverse solution $\hat{\mathbf{X}} = \mathbf{A}_S^{perp}\mathbf{Y}$ where $\mathbf{A}_S$ is derived from the dictionary matrix $\mathbf{A}$ by deleting the columns assoicated with the zero rows of X. S is called the support. (We identify the locations of sources)

3.2 Independent Component Analysis

Independent Component Analysis (ICA) is a method which assume statistical independent between the components. By statistical independent the component value do not give information of another component value. Furthermore, ICA also assume that the data of interest is nongaussian as in most practical cases the data do not follow the gaussian distribution [10, p. 3].
With this independence it is possible for ICA to separate the scalp measurements $\mathbf{Y}$ into the sources $\mathbf{X}$ and the mixing matrix $\mathbf{A}$.
Through this section the mathematical concepts of Independent Component Analysis (ICA) will be explained and defined.

Lets set up an situations. We have some measurements that has been affect by some surrounding noise or "**sideløbende**" measurements such as different conversations in a room. The measurements can be described by a vector $\mathbf{y}$ if we look at the one-dimensional case. $\mathbf{y}$ consist of the measurement from the original signal, a vector $\mathbf{x}$ and surrounding measurements, a matrix $\mathbf{A}$. This situation can be described as the linear model

$$\mathbf{y} = \mathbf{A}\mathbf{x} = \sum_{i=1}^{n} \mathbf{a}_i x_i$$

We know the measurements $\mathbf{y}$ but if also knew the mixing parameter in $\mathbf{A}$ then by inverting the linear model we could solve the system and find the original signal. But this is not the case as the mixing matrix also is unknown.
If we use the statistical properties of $\mathbf{x}$ then it would be possible to estimate both the mixing matrix and then the original signal. What ICA do is to assume statistical independence

Lets define the ICA model which is a generative model meaning that the observed data is generated by a process of mixing components which are latent component. Let n be the observed random variables such that $y_1, \ldots, y_n$ are modelled as a linear

combination of the random variables $x_1, \ldots, x_n$:

$$y_i = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n, \quad i = 1, \ldots, n$$

$$\mathbf{y} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \mathbf{x}$$

where $\mathbf{y} = \{y_i\}_{i\in[1,n]}$ and $\mathbf{x} = \{x_t\}_{t\in[1,n]}$. Furthermore, $\mathbf{x}$ is statistically mutually independent.

3.2.1 Estimation of Independent Components

Notes: Estimation with maximization of nongaussianity (see section 7.5 for nonguassianity)

Kurtosis

When estimation ICA with maximization of nongaussianity a measure of the nongaussianity is needed. Kurtosis is a quantitative measure used for nongaussianity of random variables. Kurtosis of a random variable y is defined as

$$\text{kurt}(y) = \mathbb{E}[y^4] - 3(\mathbb{E}[y^2])^2,$$

which is the fourth-order cumulant of the random variable y. By assuming that the random variable y have been normalised such that its variance $\mathbb{E}[y^2] = 1$, the kurtosis is rewritten as

$$\text{kurt}(y) = \mathbb{E}[y^4] - 3.$$

Because of this definition the kurtosis of nongaussian random variables the kurtosis will almost always be non-zero. For gaussian random variables the fourth moment equals $3(\mathbb{E}[y^2])^2$ thus the kurtosis will then be zero [10, p. 171].
By using the absolute value of the kurtosis gaussian random variables are still zero but the nongaussian random variables will be greater than zero. In this case the random variables are called supergaussian.

One complication with the use of kurtosis as measure is the used of measured samples as the kurtosis is sensitive to outliers in the measured data set [10, p. 182].

Gradient Algorithm

For ICA the wish is to maximise the nongaussianity and therefore maximise the absolute value of kurtosis. One way to do this is to use a gradient algorithm.

With a gradient algorithm you start from an initial vector $\mathbf{w}$ and then compute the direction. The direction is computed from the absolute kurtosis of $y = \mathbf{w}^T\mathbf{z}$ giving some samples of the mixture vector $\mathbf{z}$. The direction which give us the highest kurtosis is the direction where $\mathbf{w}$ is moved.
The gradient of the absolute value of kurtosis is computed as

$$\begin{aligned}\frac{\partial|\text{kurt}(\mathbf{w}^T\mathbf{x})|}{\partial\mathbf{w}} &= 4\text{sign}(\text{kurt}(\mathbf{w}^T\mathbf{x}))(\mathbb{E}[\mathbf{z}(\mathbf{w}^T\mathbf{z})^3] - 3\mathbf{w}\mathbb{E}[(\mathbf{w}^T\mathbf{z})^2] \\ &= 4\text{sign}(\text{kurt}(\mathbf{w}^T\mathbf{x}))(\mathbb{E}[\mathbf{z}(\mathbf{w}^T\mathbf{z})^3] - 3\mathbf{w}\|\mathbf{w}\|^2). \qquad (3.2)\end{aligned}$$

The absolute value of kurtosis is optimised onto the unit sphere, $\|\mathbf{w}\|^2 = 1$, the algorithm must project onto the unit sphere in every step. This can easly be done by dividing $\mathbf{w}$ with its norm.
As it is the direction of $\mathbf{w}$ of interest the last part of (3.2) can be omitted and instead the gradient of the absolute value of kurtosis is computed as

$$\frac{\partial|\text{kurt}(\mathbf{w}^T\mathbf{x})|}{\partial\mathbf{w}} = 4\text{sign}(\text{kurt}(\mathbf{w}^T\mathbf{x}))(\mathbb{E}[\mathbf{z}(\mathbf{w}^T\mathbf{z})^3]$$

The expectation operator from the kurtosis definition can not be omitted and must therefore be estimated. This can be done by a time-average, denoted as γ:

$$\gamma = ((\mathbf{w}^T\mathbf{z})^4 - 3) - \gamma$$

From the all above the following algorithm can be stated.

Algorithm 1 Gradient Algorithm with Kurtosis

1. $\gamma = ((\mathbf{w}^T\mathbf{z})^4 - 3) - \gamma$
2. $\mathbf{w} = \gamma\mathbf{z}))\mathbb{E}[\mathbf{z}(\mathbf{w}^T\mathbf{z})^3]$
3. $\mathbf{w} = \frac{\mathbf{w}}{\|\mathbf{w}\|}$

Notes: A measure of nongaussianity for the vector b which estimate 1 IC Have some outliners so we introduce negentropy

Negentropy

Another measure of nongaussianity is the negentropy which based of on the differential entropy. The differential entropy H of a random variable $\mathbf{y}$ with density $p_y(\boldsymbol{\theta})$ is defined as

$$H(\mathbf{y}) = -\int p_y(\boldsymbol{\theta})\log(p_y(\boldsymbol{\theta})\ d\boldsymbol{\theta}.$$

The entropy describe the information of a random variable and for variables that becomes more random the entropy becomes larger, e.g. Gaussian random variable has a high entropy, in fact Gaussian random variable has the highest entropy among the random variables of the same variance. Furthermore, the entropy is small for clustered random variables [10, p. 182].

To use the negentropy to define the nongaussianity within random variables, we normalised the differential entropy to obtain a entropy value equal to zero when the random variable is gaussian and non-negative otherwise. The negentropy J is defined as

$$J(\mathbf{y}) = H(\mathbf{y}_{\text{gaus}}) - H(\mathbf{y}),$$

with $\mathbf{y}_{\text{gaus}}$ been a gaussian random variable of same covariance and correlation as $\mathbf{y}$ [10, p. 182].

As the kurtosis is sensitive for outliers the negentropy is difficult to compute computationally as the negentropy require a estimate of the pdf. Instead it could be an idea to use an approximation of the negentropy.

Approximation of Kurtosis and Negentropy

Gradient Algorithm with Negentropy

As described in section **??** the gradient algorithm is used to maximising negentropy. The gradient of the approximated negentropy is given as

$$\mathbf{w} = \gamma \mathbb{E}[\mathbf{z} g(\mathbf{w}^T \mathbf{z})]$$

with respect to $\mathbf{w}$ and where $\gamma = \mathbb{E}[G(\mathbf{w}^T \mathbf{z})] - \mathbb{E}[G(\nu)]$ with ν being the standardised gaussian random variable. To omitted the expectation γ as we did with the sign of kurtosis, γ is estimated as

$$\gamma = (G(\mathbf{w}^T \mathbf{z}) - \mathbb{E}[G(\nu)]) - \gamma.$$

Algorithm 2 Gradient Algorithm

1. Center the observed data to achieve zero mean
2. Whiten the centered data
3. Create the initial random vector $\mathbf{w}$ and the initial value for γ
4. Update
$$\mathbf{w} = \gamma \mathbf{z} g(\mathbf{w}^T \mathbf{z})$$
5. Normalise $\mathbf{w}$
$$\mathbf{w} = \frac{\mathbf{w}}{\|\mathbf{w}\|}$$
6. Check sign of γ, if not a known prior, update
$$\gamma = (G(\mathbf{w}^T \mathbf{z}) - \mathbb{E}[G(\nu)]) - \gamma$$
7. Repeat until convergence

Notes: ICA can be used on Gaussian variables as little is done in addition to decorrelate for Gaussian variable
Whiting is useful to be done beore ICA
A drawback of ICA is the system must be $N \leq M$ meaning that there must more sensors than sources which is not the case in this project where we look at low density EEG system, $M \leq N$. Furthermore, ICA need that the sources are stationary which is not the nature of EEG that are very much nonstationary [**PHD**].
Instead a mixture model of ICA model where we assume that the amount of activation k in N sources are equal to M (sensor). We can used the short time frame of the sources to make them stationary

3.3 Covariance-Domain Dictionary Learning

Covariance-domain dictionary learning (Cov-DL) is an algorithm which can identify more sources N than sensors M for the linear model of observed EEG data

$$\mathbf{Y} = \mathbf{AX}.$$

Cov-DL takes advantage of dictionary framework and transformation into another domain – covariance domain – to recover the mixing matrix $\mathbf{A}$ from the observed

data $\mathbf{Y}$. Cov-DL work together with another algorithm to find the sparse source matrix $\mathbf{X}$, in this thesis M-SBL is used for the source recovery and is described in section 3.4.
In the following section we assume that $\mathbf{X}$ is known but in practice a random sparse matrix will be used to represent the sources.
This section is inspired by chapter 3 in [5] and the article [2].

(?)
In dictionary learning framework the inverse problem is defined as

$$\min_{A,X} = \frac{1}{2}\sum_{s=1}^{N_d} \|\mathbf{Y} - \mathbf{A}\mathbf{X}\|_F^2 + \gamma \sum_{s=1}^{N_d} g(\mathbf{x}_s),$$

where the function $g(\cdot)$ promotes sparsity of the source vector at time t The true dictionary $\mathbf{A}$ is recovered if the sources $\mathbf{x}_s$ are sparse ($k_s < M$).

Introduction to Our Covariances: Let s be the time segments that $\mathbf{Y}$ is divided into and let it be sampled with the frequency S_f such that our observed data is known overlapping segments $\mathbf{Y}_s \in \mathbb{R}^{M \times t_s S_f}$ where t_s is the length of the segments in seconds. With the segments the linear model still holds and is rewritten into

$$\mathbf{Y}_s = \mathbf{A}\mathbf{X}_s, \quad \forall s.$$

The sources are assumed uncorrelated in time segments of the whole time scheme of the observed data.
In the covariance-domain, the observed segmented data $\mathbf{Y}_s$ is described by its covariance:

$$\begin{aligned}
\mathbf{\Sigma}_{\mathbf{Y}_s} &= \mathbf{A}\mathbf{\Lambda}\mathbf{A}^T + \mathbf{E}_s \\
\text{vech}(\mathbf{\Sigma}_{\mathbf{Y}_s}) &= \sum_{i=1}^{N} \mathbf{\Lambda}_{s_{ii}} \text{vech}(\mathbf{a}_i \mathbf{a}_i^T) + \text{vech}(\mathbf{E}_s) \\
\text{vech}(\mathbf{\Sigma}_{\mathbf{Y}_s}) &= \mathbf{D}\boldsymbol{\delta}_s + \text{vech}(\mathbf{E}_s), \quad \forall s.
\end{aligned}$$

The vector $\boldsymbol{\delta}_s$ contains the diagonal entries of the source sample-covariance matrix

$$\mathbf{\Sigma}_{\mathbf{X}_s} = \frac{1}{L_s}\mathbf{X}_s\mathbf{X}_s^T,$$

and the matrix $\mathbf{D} \in \mathbb{R}^{M(M+1)/2 \times N}$ consists of the columns $\mathbf{d}_i = \text{vech}(\mathbf{a}_i\mathbf{a}_i^T)$. D and δ_s are unknown.
Our goal is to learn $\mathbf{D}$ and the find the associated matrix $\mathbf{A}$. When we have the dictionary matrix we can find the mixing matrix by

$$\min_{\mathbf{a}_i} \|\mathbf{d}_i - \text{vech}(\mathbf{a}_i\mathbf{a}_i^T)\|_2^2.$$

Introduction to Covariance Domain Transformation: By the assumption of uncorrelated sources, the sample covariance source matrix is given as

$$\begin{aligned}\mathbf{\Sigma}_{\mathbf{X}_s} &= \frac{1}{L_s}\mathbf{X}_s\mathbf{X}_s^T \\ &= \mathbf{\Lambda} + \mathbf{E},\end{aligned}$$

where $\mathbf{\Lambda}$ is the diagonal matrix of $\mathbf{\Sigma}_{\mathbf{X}_s}$. With this mindset, the linear model given in (**??**) can then be modelled as

$$\begin{aligned}\mathbf{Y}_s\mathbf{Y}_s^T &= \mathbf{A}\mathbf{X}_s\mathbf{X}_s^T\mathbf{A}^T \\ \mathbf{\Sigma}_{\mathbf{Y}_s} &= \mathbf{A}\mathbf{\Sigma}_{\mathbf{X}_s}\mathbf{A}^T \\ \mathbf{\Sigma}_{\mathbf{Y}_s} &= \mathbf{A}\mathbf{\Lambda}\mathbf{A}^T + \mathbf{E} \\ &= \sum_{i=1}^{N}\mathbf{\Lambda}_{ii}\mathbf{a}_i\mathbf{a}_i^T + \mathbf{E}.\end{aligned}$$

As the covariance matrices are symmetric the lower triangular part can be vectorised:

$$\begin{aligned}\text{vech}(\mathbf{\Sigma}_{\mathbf{Y}_s}) &= \sum_{i=1}^{N}\mathbf{\Lambda}_{ii}\text{vector}(\mathbf{a}_i\mathbf{a}_i^T) + \text{vech}(\mathbf{E}) \\ \text{vech}(\mathbf{\Sigma}_{\mathbf{Y}_s}) &= \sum_{i=1}^{N}\mathbf{\Lambda}_{ii}\mathbf{d}_i + \text{vech}(\mathbf{E}) \\ \text{vech}(\mathbf{\Sigma}_{\mathbf{Y}_s}) &= \mathbf{D}\delta + \text{vech}(\mathbf{E}),\end{aligned}$$

where $\mathbf{d}_i = \text{vech}(\mathbf{a}_i\mathbf{a}_i^T)$. The size of the vectorised covariance matrices is $\frac{M(M+1)}{2}$. By use of the covariance domain it is possible to identify $\mathcal{O}(M^2)$ sources given the true dictionary matrix $\mathbf{A}$.

Notes: In the case of EEG, this allows at most k = O(M) EEG sources to be simultaneously active which limits direct applicability of dictionary learning to low-density EEG systems.
We wish to handled cases where we have $\binom{N}{k}$ sources, where $1 \leq k \leq N$ can be jointly active.
section 3.3.1 in phd.

3.4 MSB

See chapter 2 in PHD.

As described in earlier sections, the support set of sources is wish recovered. A way to recover the set is to use sparse bayesian learning (M-SBL) on the MMV model.

To insure full recovery some sufficient condition must be applied on the dictionary matrix $\mathbf{A}$ and the sources.

Lets first sketch the case. For the MMV model we assume that we have more sources than sensors $M \leq N$ and the activations inside the sources k are less than the sensors $k \leq N$. At last we assume that the mixing happen instantaneous meaning that no time delay occur – we will work in the time domain.
We will look at two sufficient conditions for exact recovery of the support set S: orthogonality/uncorrelated of the active sources k and constraint on the dictionary matrix $\mathbf{A}$.

3.4.1 M-SBL Algorithm

The i-th row of the sources matrix $\mathbf{X}$, $\mathbf{x}_{i.}$, has an L-dimensional independent gaussian prior with zero mean and a variance controlled by γ_i which is unknown:

$$\begin{aligned} p(\mathbf{x}_{i.}; \gamma) &= \mathcal{N}(0, \gamma_i \mathbf{I}) \\ p(\mathbf{y}_{.j} \mid \mathbf{x}_{.j}) &= \mathcal{N}(\mathbf{A}\mathbf{x}_{.j}, \sigma^2 \mathbf{I}) \\ p(\mathbf{Y} \mid \mathbf{X}) &= \prod_{j=1}^{L} p(\mathbf{y}_{.j} \mid \mathbf{x}_{.j}) \end{aligned}$$

By integrating the unknown sources $\mathbf{X}$ the marginal likelihood of the observed mixed data $\mathbf{Y}$, $p(\mathbf{Y}; \gamma)$ is achieved. By applying $-2\log(\cdot)$ the marginal likelihood function is transformed to the cost function

$$\begin{aligned} \mathcal{L}(\gamma) &= -2\log(p(\mathbf{Y}; \gamma)) = -2\log\left(\int p(\mathbf{Y} \mid \mathbf{X}) p(\mathbf{X}; \gamma)\, d\mathbf{X}\right) \\ &= \log(|\mathbf{\Sigma}|) + \frac{1}{L}\sum_{t=1}^{L} \mathbf{y}_{.t}^T \mathbf{\Sigma}^{-1} \mathbf{y}_{.t} \end{aligned}$$

with

$$\mathbf{\Sigma} = (\mathbf{A}\mathbf{\Gamma}\mathbf{A}^T + \sigma^2 \mathbf{I}), \quad \mathbf{\Gamma} = \operatorname{diag}(\gamma).$$

To reach local minimum of the cost function we use a fixed point update that is fast and decrease the likelihood function at every step,

$$\gamma_i^{(k+1)} = \frac{\gamma_i^{(k)}}{\sqrt{\mathbf{a}_i^T (\mathbf{\Sigma}^{(k)})^{-1} \mathbf{a}_i}} \frac{\|\mathbf{Y}^T (\mathbf{\Sigma}^{(k)})^{-1} \mathbf{a}_i\|_2}{\sqrt{n}}$$

After convergence the support set $\hat{S}$ is extracted from the solution $\hat{\gamma}$ by $\hat{S} = \{i, \hat{\gamma}_i \neq 0\}$.

Notes:

- With M-SBL the **support set** of source can be recover for $k \geq M$, with some sufficient condition on the dictionary and sources. $M \leq k \leq N$ the support set can be recovered in the noiseless case.
- We assume that mixing at the sensors is instantaneous (no time delay between sources and sensors) and the environment is anechoic. (M-SBL)
- The sufficient conditions for exact support recovery for M-SBL in the regime k ≥ M are twofold: 1) orthogonality (uncorrelated) of the active sources, 2) The second condition imposes a constraint on the sensing dictionary A
- Bayesian replace the troublesome prior with a distribution that, while still encouraging sparsity, is somehow more computationally convenient. Bayesian approaches to the sparse approximation problem that follow this route have typically been divided into two categories: (i) maximum a posteriori (MAP) estimation using a fixed, computationally tractable family of priors and, (ii) empirical Bayesian approaches that employ a flexible, parameterized prior that is 'learned' from the data

Bibliography

[1] Alickovic, Emina et al. “A Tutorial on Auditory Attention Identification Methods”. In: *Front. Neurosci* 13:153 (2019).

[2] Balkan, O. Y., Kreutz-Delgado, K., and Makeig, S. “Covariance-Domain Dictionary Learning for Overcomplete EEG Source Identification”. In: *CoRR* abs/1512.00156 (2015).

[3] Balkan, Ozgur, Kreutz-Delgado, Kenneth, and Makeig, Scott. “Covariance-Domain Dictionary Learning for Overcomplete EEG Source Identification”. In: *ArXiv* (2015).

[4] Balkan, Ozgur, Kreutz-Delgado, Kenneth, and Makeig, Scott. “Localization of More Sources Than Sensors via Jointly-Sparse Bayesian Learning”. In: *IEEE Signal Processing Letters* (2014).

[5] Balkan, Ozgur Yigit. “Support Recovery and Dictionary Learning for Uncorrelated EEG Sources”. Master thesis. University of California, San Diego, 2015.

[6] Bech Christensen, Christian et al. “Toward EEG-Assisted Hearing Aids: Objective Threshold Estimation Based on Ear-EEG in Subjects With Sensorineural Hearing Loss”. In: *Trends Hear, SAGE* 22 (2018).

[7] C. Eldar, Yonina and Kutyniok, Gitta. *Compressed Sensing: Theory and Application*. Cambridge University Presse, New York, 2012.

[8] Delorme, Arnaud et al. “Blind separation of auditory event-related brain responses intoindependent components”. In: *PLoS ONE* 7(2) (2012).

[9] Foucart, Simon and Rauhut, Hoger. *A Mathematical Introduction to Compressive Sensing*. Springer Science+Business Media New York, 2013.

[10] Hyvarinen, A., Karhunen, J., and Oja, E. *Independent Component Analysis*. Ed. by Haykin, Simon. John Wiley and Sons, Inc., 2001.

[11] Makeig, Scott et al. “Blind separation of auditory event-related brain responses into independent components”. In: *Proc. Natl. Acad. Sci. USA* 94 (1997).

[12] Makeig, Scott et al. “Independent Component Analysis of Electroencephalographic Data”. In: *Advances in neural information processing systems* 8 (1996).

[13] Palmer, J. A. et al. “Newton Method for the ICA Mixture Model”. In: *ICASSP 2008* (2008).

[14] Sanei, Saeid and Chambers, J.A. *EEG Signal Processing*. Johm Wiley and sons, Ltd, 2007.

[15] Steen, Frederik Van de et al. “Critical Comments on EEG Sensor Space Dynamical Connectivity Analysis”. In: *Brain Topography* 32 p. 643-654 (2019).

[16] Teplan, M. “Fundamentals of EEG”. In: *Measurement science review* 2 (2002).

[17] V. Le, Quoc et al. “ICA with Reconstruction Cost for Efficient Overcomplete Feature Learning”. In: *NIPS’11 International Conference on Neural Information Processing Systems P. 1017-1025* (2011).

Appendix A

Appendix A