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: 35

Date of Completion:
October 16, 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: 35

Afleveringsdato:
16. 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 16, 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.

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 [19]. 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 [19]. Lastly the transmission of the electric field through the biological tissue to the sensor has an unknown mixing effect on the signal, this process is called volume conduction[16,

p. 68][17].

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 measurements 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 [16, p. 11]. An example of EEG measurements within the four categories is illustrated by figure 1.2.

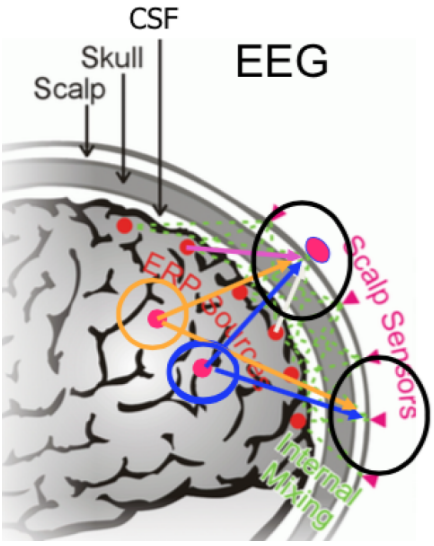

Figure 1.1: Illustration of volume conduction, source [4](we will make our own figure here instead)

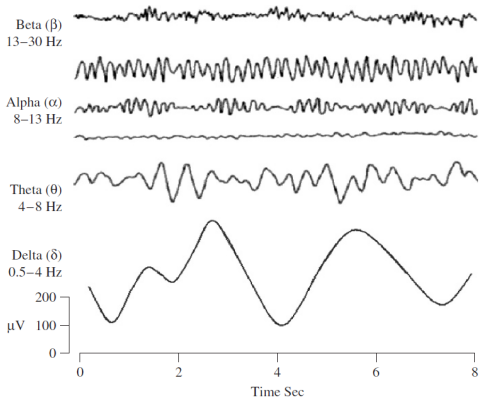

Figure 1.2: Example of time dependent EEG measurements within the four defined categories, source [16]

EEG is widely used within the medical field, especially research of the cognitive processes in the brain. Diagnosis and management of neurological disorders such as epilepsy is one example.

EEG capitalize on the procedure being non-invasive and fast. Neural activity can be measured within fractions of a second after a stimuli has been provided [19, 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[9]. Within neurobiology functional integration referrers to the study of the correlation among activities in different regions of the brain. In other words, how do different part of the brain work together to process information and conduct a response[10]. For this purpose separation and localisation of the single sources which contribute to the EEG measurement is of interest. An article from 2016 point out the importance of performing analysis regarding functional integration at source level rather than at EEG level. It is argued through experiments that analysis at EEG level do not allow interpretations about the interaction between sources[17].

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[18]. One main purpose of Eriksholm is to make it possible for a hearing aid to identify the attended sound source and hereby exclude noise from elsewhere [1], [5]. This is where EEG and occasionally so called in-ear EEG is interesting, especiallaly in conjunction with the technology of beamforing, which allows for receiving only signals from a specific direction. It is essentially the well known but unsolved cocktail problem which is sought improved by use of EEG. However the focus of this research do consider the correlation between EEG measurements and the sound source rather than localisation of the activated source from the EEG[1]. Hence a source localisation approach could potentially be of interest regarding hearing aids in order to improve the results. (Furthermore, a real-time application to provide feedback from EEG measurements would be essential.)?

1.1.1 Modelling

Considering the issue of localising activated sources from EEG measurements, a 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 [4]. This is in general referred to as a multiple measurement vector model. The aim in this case is to identify both $\mathbf{A}$ and $\mathbf{X}$ given the measurements $\mathbf{Y}$. For this specific set up the model is 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 [13], [12], 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 [7].
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[2].
This means that the EEG inverse problem can not be over-complete(er det correct i forhold til teorien om ICA?). 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 [4]. 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 adequate. To overcome this issue an extension of ICA was suggested, referred to as the ICA mixture model[2]. 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) which showed successful results regarding identification of more sources than available sensors [15]. 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 [20]. Here the hard orthonormal constraint in ICA is replaced with a soft reconstruction cost.
In 2015 O. Balkan et. al., [2], 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, this is an essential advantage when low-density EEG is considered.
Cov-DL was tested on found to outperform both AMICA and RICA[2], 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 [3] also by O. Balkan et. al., also targeting the case of more active sources than sensors [3]. Here the mixing matrix which is 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 recovery conditions was not fulfilled.

The two state of the art methods for source identification makes the foundation of this thesis. This thesis propose an algorithm with the purpose of solving the EEG inverse problem using the presented methods on EEG measurement. To extent the existing results the algorithm is expanded into a real-time application, in order to provide feedback based on the source activity.
The intention of the feedback is to adjust the direction of the beam within the hearing aid depending on the source activity. For this, the application is tested within a simulation environment where the receiving direction of the test person can be adjusted in real-time. The quality of the final results is measured by the capability of improving the listener experience and the time used to proved useful feedback.

As such our contribution *(hopefully)* consists of tests of existing methods on new real-time measurement and furthermore include a feedback to control the microphone beam on a hearing aid.

note: Evt. kunne vi lave en figur der lidt ala mindmap sætte et system overblik op og så highlighte de "bokse" vi vælger at arbejde med.

Chapter 2

Problem Statement

From the motivation and related work described in chapter 1 it is stated that EEG measurement of the brain activity has great potential to contribute within the hearing aid industry, regarding the development of hearing aids with improved performance in situations as the cocktail party problem. By solving the overcomplete EEG inverse problem, in order to localise the sources of the brain activity, the results could be used to guide and adapt the hearing aids performance such as move the microphone beam 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 overcomplete case of less sensors than sources and how can such algorithm be extended to a real-time application providing feedback to improve the intentional listening experience?

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 overcomplete EEG inverse problem?
- How can M-SBL be used to estimate the source matrix $\mathbf{X}$ from the overcomplete EEG inverse problem?
- How can an application be formed to constitute this source identification process operating in real-time?
- How can the feedback of the system be used to control the microphone beam of a simulated hearing aid. Especially how to analyse the feedback versus the listening experience in order to improve this.

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 recovery or reconstruction of a signal from a minimal number of measurements. Assuming linear acquisition of the original information the relation between the measurements and the signal to be recovered is described by a linear model[8]

$$\mathbf{y} = \mathbf{A}\mathbf{x}, \tag{3.1}$$

where $\mathbf{y} \in \mathbb{R}^M$ is the observed data, $\mathbf{x} \in \mathbb{R}^N$ is the original signal and $\mathbf{A} \in \mathbb{R}^{M \times N}$ is a matrix which models the linear measurement process or in other word it maps from $\mathbb{R}^N$ to $\mathbb{R}^M$.
In compressive sensing terminology, $\mathbf{x}$ is the signal of interest which is sought recovered by solving the linear system (3.1). In the typical compressive sensing case where $M < N$ the system become underdetermined and there are infinitely many solutions, provided that a solution exist. Such system is also referred to as overcomplete(as the number of basis vectors is greater than the dimension of the input). However, by enforcing certian sparsity constraints it is possible to recover the wanted signal[8].
And another argument; If $M << N$, it leads to the matrix $\mathbf{A}$ being rank-deficient(but not necessarily?) which imply that $\mathbf{A}$ has a non-empty null space and this leads to infinitely many signals will yield the same solution $\mathbf{y} = \mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{x}'$ [6, p. ix]. Thus it is necessary to limit the solution space to a specific class of signals $\mathbf{x}$, for this certain constraints on sparsity is introduced.

A signal is said to be k-sparse if the signal has at most k non-zeros coefficient, for this purpose the ℓ_0-norm is defined

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

The function $\text{card}(\cdot)$ is the cardinality and the support of $\mathbf{x}$ is giving as

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

where $[N]$ a set of integers $\{1, 2, \cdots, N\}$ [8, p. 41]. The set of all k-sparse signals is denoted as

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

A signal is assumed to be k-sparse with $k < M$[6, p. 8](not found yet?)

From the desire of finding a sparse solution $\mathbf{x}$ the following optimisation problem is considered

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

where $\mathbf{z}$ is all possible candidates to an k-sparse signal $\mathbf{x}$.
Unfortunately, this optimisation problem is non-convex due to the definition 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 [6, p. 27]

$$\min_{\mathbf{z}} \|\mathbf{z}\|_1 \quad \text{s.t} \quad \mathbf{y} = \mathbf{A}\mathbf{z}, \tag{3.2}$$

and instead we find the best k-term approximation of the signal $\mathbf{x}$. This method is referred to as Basis Pursuit and makes the foundation of several algorithms solving alternative versions of (3.2) where noise is incorporated. A different type of solution method includes greedy algorithm such as the Orthogonal Matching Pursuit.

3.1.1 Conditions on the Mixing Matrix

The construction of the matrix $\mathbf{A}$ is of course essential for the solution of the optimisation problem. So far no one has manage to construct a matrix which is proved optimal for some compressive sensing set up. However some certain constructions have shown sufficient recovery guarantee.

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.
(Next section are not finished, is it necessary with the details?)

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, (3.2), 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 (3.2). 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

Mangler at:
- fixed-point (fastICA)
- læs igennem

Independent Component Analysis (ICA) is a method which assume statistical independent between the components and nongaussiantity of the data. By those assumptions it is possible to recover the components and some mixing matrix $\mathbf{A}$ from some observed data [11, p. 3].
Through this section the mathematical concepts of ICA will be explained and defined in the noise-less case.

To understand what ICA is lets describe a situation where ICA could be used. Two people are having a conversation inside a room full of people which talk simultaneous which each other. The conversation between the two first mention people will the be affected by the surrounding conversations and noise. Such environment is often called the cocktail party problem and is a difficult environment to be in as a hearing impaired.
Let describe the situation with some mathematical notations. The observed conversation which affected by surrounding noise is denoted $\mathbf{y}$, the original individual conversations in the room is denoted by $\mathbf{x}$. The original conversations are all effected by the surrounding, lets denoted this mixture by a matrix $\mathbf{A}$. We omit the time indexes and the time dependent to view the problem with random vectors. The problem can then be described as a linear model

$$\mathbf{y} = \mathbf{A}\mathbf{x} = \sum_{i=1}^{n} \mathbf{a}_i x_i, \tag{3.3}$$

$$y_i = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n, \quad i = 1, \ldots, n. \tag{3.4}$$

The only known variables in this model are the observed conversation $\mathbf{y}$, the rest are unknown.

ICA is a method which can be use to recover the unknown mixture $\mathbf{A}$ and $\mathbf{x}$, which in the ICA aspect are the components, from the known $\mathbf{y}$. By use of the statistical properties of the components $\mathbf{x}$ it is then possible to estimate/recover the original conversations and then the mixture $\mathbf{A}$.
One of the properties is the components $\mathbf{x}$ are assumed statistical independent. By independence, it mean that the joint probability density function (pdf) can be factorised into the marginal pdfs of the components $\mathbf{x}$ such that

$$p(x_1, x_2, \ldots, x_n) = p_1(x_1)p_2(x_2)\cdots p_n(x_n),$$

for the random variables x_i.
Furthermore it is assumed that the independent components $\mathbf{x}$ do not have gaussian distributions as this will ruin the ICA method as when introducing the algorithms of ICA, cf. section XX, the used of higher-order cumulant is used. For gaussian distribution the cumulant will be zero which not wanted in the recovering process. Thus the distribution of the independent components are unknown.
The last thing which must be assumed to the ICA work is that the mixing matrix $\mathbf{A}$ must be a square matrix – there must be the same number of independent components and observation/mixing [11, p. 152-153].

By the unknowing of the right-hand side of (3.3) some statistical can not be computed such as the variance of $\mathbf{x}$. Instead ICA assume that $\mathbf{x}$ has a unit variance (=1). This of course must also applied to the mixing matrix $\mathbf{A}$ which will be restricted in

the recovering method which will be described in section XX. By this addition we simplify the algorithm for ICA. Further to simplify even more, we can with out loss of generality assume that $\mathbf{y}$ and $\mathbf{x}$ have zero mean. If the observed data do not have zero mean already then by preprocessing the data by centering, zero mean can be achieved. This is done by finding the sample mean of $\mathbf{y}$ and subtract it from the data before ICA is applied:

$$\mathbf{y} = \mathbf{y} - \mathbb{E}[\mathbf{y}].$$

This addition do not affect the recover/estimation of $\mathbf{A}$ [11, p. 154].

By preprocessing the data before applying the ICA algorithm with whiting have shown to reduce the complexity and make the recovering process more simple. Furthermore, when whiting the data $\mathbf{y}$ the independent components becomes uncorrelated and have variance equal to one.
The data $\mathbf{y}$ is whited by finding the eigenvalue decomposition (EVD) of the covariance matrix of $\mathbf{y}$

$$\mathbb{E}[\mathbf{y}\mathbf{y}^T] = \mathbf{E}\mathbf{D}\mathbf{E}^T,$$

where $\mathbf{D}$ is the diagonal matrix with eigenvalues of the covariance matrix in the diagonal and $\mathbf{E}$ is the associated eigenvectors. By multiplying the data $\mathbf{y}$ with a whiting matrix $\mathbf{V}$ constructed from the EVD, the data becomes whited:

$$\begin{aligned} \mathbf{y}_{\text{white}} &= \mathbf{V}\mathbf{y}, \quad \mathbf{V} = \mathbf{E}\mathbf{D}^{-1/2}\mathbf{E}^T, \\ \mathbf{y}_{\text{white}} &= \mathbf{V}\mathbf{A}\mathbf{x} = \mathbf{A}_{\text{white}}\mathbf{x} \end{aligned} \tag{3.5}$$

A by product of whiting $\mathbf{y}$ is that the new mixing matrix $\mathbf{A}_{\text{white}}$ become orthogonal and therefore ICA can restrict its search for $\mathbf{A}_{\text{white}}$ to the orthogonal matrix space. By computation mean this implies instead of having n^2 computations the orthogonal space only have $n(n-1)/2$ computations [11, p. 159].
Furthermore, to support the assumption of nongaussiantity the whiting preprocessing can be used. As an orthogonal mixing matrix $\mathbf{A}_{\text{white}}$ is achieved it is not possible to distinct the pdfs of the $\mathbf{y}_{\text{white}}$ and $\mathbf{x}$ as $\mathbf{A}_{\text{white}}$ is no longer included in the pdf of $\mathbf{y}_{\text{white}}$ and therefore are the two pdfs equal [11, p. 161-163].

3.2.1 Estimation of Independent Components

A way to estimate/recover the independent components could be to take advantage of the assumption of nongaussiantity. By finding the estimate which maximise the nongaussiantity the real independent component can be found based on it must have a nongaussian distribution as mention in section XX. But before the estimation a measure of nongaussiantity must be introduce – this could be the kurtosis.

Kurtosis

Kurtosis is a quantitative measure used for nongaussianity of random variables. Kurtosis of a random variable y is defined as

Tjek lige op denne definition

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

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 [11, 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 kurtosis as a is that kurtosis is sensitive to outliers [11, p. 182].

The 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 mixture vector $\mathbf{z}$ is the whited observed mixture vector $\mathbf{y}_{\text{white}}$. 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

Regn lige efter

$$\begin{aligned}\frac{\partial|\text{kurt}(\mathbf{w}^T\mathbf{x})|}{\partial\mathbf{w}} &= 4\text{sign}(\text{kurt}(\mathbf{w}^T\mathbf{x}))(\mathbb{E}[\mathbf{z}(\mathbf{w}^T\mathbf{z})^3] - 3\mathbf{w}\mathbb{E}[(\mathbf{w}^T\mathbf{z})^2] \\ &= 4\text{sign}(\text{kurt}(\mathbf{w}^T\mathbf{x}))(\mathbb{E}[\mathbf{z}(\mathbf{w}^T\mathbf{z})^3] - 3\mathbf{w}\|\mathbf{w}\|^2). \qquad (3.6)\end{aligned}$$

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 (3.6) 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. 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. Compute $\mathbf{w} = \gamma\mathbf{z}\mathbb{E}[\mathbf{z}(\mathbf{w}^T\mathbf{z})^3]$
5. Normalise $\mathbf{w} = \frac{\mathbf{w}}{\|\mathbf{w}\|}$
6. Update $\mathbf{w}$
7. Update $\gamma = ((\mathbf{w}^T\mathbf{z})^4 - 3) - \gamma$
8. Repeat until convergence

Fixed-Point Algorithm - FastICA A fixed-point algorithm to maximise the nongaussity is more efficient than the gradient algorithm as the gradient algorithm converge slow as depending on the choice of γ which could be chosen such that convergence never gonna be reach. The fixed-point algorithm is an alternative that could be used. By looking at the gradient of kurtosis given in (3.6) and then set the equation equal with $\mathbf{w}$:

$$\mathbf{w} \propto (\mathbb{E}[\mathbf{z}(\mathbf{w}^T\mathbf{z})^3] - 3\|\mathbf{w}\|^2\mathbf{w})$$

By this new equation, the algorithm find $\mathbf{w}$ by simply calculating the right-hand side:

$$\mathbf{w} = \mathbb{E}[\mathbf{z}(\mathbf{w}^T\mathbf{z})^3] - 3\mathbf{w}$$

As with the gradient algorithm the fixed-point algorithm do also divide the found $\mathbf{w}$ by its norm. Therefore is $\|\mathbf{w}\|$ omitted from the equation.
Instead of γ the fixed-point algorithm compute $\mathbf{w}$ directly from previous $\mathbf{w}$.

The fixed-point algorithm have been summed in the following algorithm.

Algorithm 2 Fixed-Point Algorithm with Kurtosis

1. Center the observed data to achieve zero mean
2. Whiten the centered data
3. Create the initial random vector $\mathbf{w}$
4. Compute $\mathbf{w} = \mathbb{E}[\mathbf{z}(\mathbf{w}^T\mathbf{z})^3] - 3\mathbf{w}$
5. Normalise $\mathbf{w} = \frac{\mathbf{w}}{\|\mathbf{w}\|}$
6. Update $\mathbf{w}$
7. Repeat until convergence

The fixed-point algorithm is also called for FastICA as the algorithm has shown to converge fast and reliably, then the current and previous $\mathbf{w}$ laid in the same direction [11, p. 179].

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 [11, 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 $\mathbf{y}$ [11, 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

The way to approximate the negentropy is to look at the high-order cumulants using polynomial density expansions such that the approximation could be given as

$$J(y) \approx \frac{1}{12}\mathbb{E}[y^3]^2 + \frac{1}{48}\text{kurt}(y)^2. \tag{3.7}$$

This is in the scalar case as in practice the approximation can be in the scalar case. The random variable y has zero mean and unit variance and the kurtosis is introduced in the approximation. The approximation suffers from nonrobustness with the kurtosis and therefore a more generalised approximation is presented to avoid the nonrobustness.

For the generalised approximation the use of expectations of nonquadratic functions is introduced. The polynomial functions y^3 and y^4 from (3.7) are replaced by G^i with i been an index and G been some function. The approximation in (3.7) then becomes

$$J(y) \approx (\mathbb{E}[G(y)] - \mathbb{E}[G(\nu)])^2.$$

The choice of G can lead to a better approximation than (3.7) and by choosing one with do not grow to fast more robust estimators can be obtained. The choice of G could be the two following functions

$$G_1(y) = \frac{1}{a_1}\log(\cosh(a_1 y)), \quad 2 \le a_1 \le 2$$
$$G_2(y) = -\exp\left(\frac{-y^2}{2}\right)$$

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. g is the derivative of the nonquadratic function G. To omitted the expectation γ as we did with the sign of kurtosis, γ is estimated as

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

For the choice of g the derivative of the functions presented in (**??**) could be use to achieve a robust result. Alternative a derivative which correspond to the fourth-power as seen in the kurtosis could be used. The functions g could be

$$g_1(y) = \tanh(a_1 y), \quad 1 \leq a_1 \leq 2$$
$$g_2(y) = y \exp\left(\frac{-y^2}{2}\right)$$
$$g_3(y) = y^3$$

Algorithm 3 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

Fixed-Point Algorithm with Negentropy

As described in the section with kurtosis, the fixed-point algorithm removed the learning parameter and compute $\mathbf{w}$ directly:

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

.

Write the expression from this equation to the on in the algorithm

Algorithm 4 Fixed-Point Algorithm with Negentropy (FastICA)

1. Center the observed data to achieve zero mean
2. Whiten the centered data
3. Create the initial random vector $\mathbf{w}$
4. Update
$$\mathbf{w} = \mathbb{E}[\mathbf{z}g(\mathbf{w}^T\mathbf{z})] - \mathbb{E}[g'(\mathbf{w}^T\mathbf{z})]\mathbf{w}$$
5. Normalise $\mathbf{w}$
$$\mathbf{w} = \frac{\mathbf{w}}{\|\mathbf{w}\|}$$
6. Repeat from 4. until convergence

Notes: 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. 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 this 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 [4] and the article [2].

(hvor henne indgår dette afsnit?)

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 ($k_s < M$).

Notes, maybe prior to this section:

- *in general for $k > M$ it is not possible to recover $\boldsymbol{x}$ as the system is underdetermined/overcomplete, furthermore we can not find the true dictionary $\boldsymbol{A}$ by dictionary learning methods because when $k > M$, where M is the dimension of $\boldsymbol{y}$, any random dictionary can be used create $\boldsymbol{y}$ from $\geq M$ basis vectors. that is generally the accuracy of recovery a $\boldsymbol{A}$ increases as $k << M$.*

- *To avoid this Cov-DL was proposed by O. Balkan. Though the theory of getting from M sources to M^2 was established in [14]. saying that the design of the measurement/dictionary matrix is essential to overcome this issue, new conditions will be developed for this to be a success [14],*

- *about the approach: "This will in turn lead to the development of new sampling schemes and justify the need for the use of nested and coprime sampling. Another noteworthy point is that we distinguish the recovery of the sparse vector from that of its support."[14].*

- *In [14] it is in fact the support of x which if found by use of the covariance domain, and this is what Balkan uses on the dictionary matrix i suppose.*

- *the assumed prior(which has not been exploited before [14]): a prior on the correlation of the received signal is to assume a (pseudo)diagonal covariance matrix for the unknown signal, that is $\frac{1}{L_s}\boldsymbol{X}_s\boldsymbol{X}_s^T$[14], which is then under the assumption of uncorrelated sources in x[4]. This is what lead to the correlation constraint (3.3) in phd.*

- *the conditions mentioned on the measurement/dictionary matrix in [14] is: "Sparse sampling schemes such as nested and coprime sampling will be shown to satisfy these conditions". For $\boldsymbol{A}$ we need "$O(M^2)$ Kruskal Rank for their Khatri-Rao products". This is not necessarily what Balkan uses.*

- *The prior mentioned by Balkan is the uncorrelated sources.*

We know:

- *Covariance matrix:(when we have several observations)* $\Sigma_{\boldsymbol{x}} = Cov(\boldsymbol{x}) = Cov(\boldsymbol{x}, \boldsymbol{x}) = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T] - \mu_x \mu_{x^T}$ *, symmetric and positive simidefinite.*
- *Sample covariance matrix:* $'\Sigma_{\boldsymbol{x}} = \frac{1}{L}\sum_{i=1}^{L}(\boldsymbol{x}_i - \mathbb{E}[\boldsymbol{x}])(\boldsymbol{x}_i - \mathbb{E}[\boldsymbol{x}])^T$ *using unbiased estimate since we assume* $\mathbb{E}[\boldsymbol{x}] = 0$
- *this becomes* $\frac{1}{L}\boldsymbol{X}\boldsymbol{X}^T$*, where L is the length of the sample segment.*
- *the covariance of two vectors* $cor(\boldsymbol{x}, \boldsymbol{y})$ *is also referred to as the cross-correlation.*

Consider the multiple measurement vector model as above

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

Let s be the index of time segments that the observed data $\mathbf{Y} \in \mathbb{R}^{M \times L}$ is divided into and let S_f be the sample frequency. As such the observed data is divided into segments $\mathbf{Y}_s \in \mathbb{R}^{M \times t_s S_f}$, possibly overlapping, where t_s is the length of the segments in seconds. For each segment the linear model still holds and is rewritten into

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

An important aspect of this method is the prior assumption that the sources are statistical independent within the defined time segments. This implies the entries in $\mathbf{X}_s$ to be uncorrelated.

3.3.1 Covariances domain representation

The observed data $\mathbf{Y}_s$ can be described in the covariance domain by the sample covariance matrix. Considering the observations $\mathbf{Y}_s \in \mathbb{R}^{M \times L}$ the sample covariance is defined to find the covariance among the M variables across the L observations, that is essentially the covariance matrix averaged over all observations, resulting in a $M \times M$ matrix $\Sigma_{\mathbf{Y}_s} = [\sigma_{jk}]$. Each entry is defined by[wiki?]

$$\sigma_{jk} = \frac{1}{L}\sum_{i=1}^{L}(y_{ji} - \mathbb{E}[y_j])(y_{ki} - \mathbb{E}[y_k])$$

Let the observations(argument for at vi antager zero mean på vores observationer, pre-normalisering?) be normalised resulting in zeros mean $\mathbb{E}[\mathbf{Y}_s] = 0$.
Using matrix notation the sample covariance of $\mathbf{Y}_s$ can be written as

$$\Sigma_{\mathbf{Y}_s} = \frac{1}{L}\mathbf{Y}_s\mathbf{Y}_s^T$$

Similar the sources $\mathbf{X}_s$ can be described in the covariance domain by the sample covariance matrix

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

From the assumption of uncorrelated sources the sample covariance matrix is expected to be nearly diagonal, thus it can be expressed as

$$\mathbf{\Sigma}_{\mathbf{X}_s} = \Lambda + \mathbf{E}.$$

where Λ is a diagonal matrix consisting of the diagonal entries of $\Sigma_{\mathbf{X}_s}$ and $\mathbf{E}$ contains the remaining entries which is expected to be zeros or nearly zeros[2].

Each segment of observations can be modelled as

skal vi have $\mathbb{E}[]$ omkring her eller ej?, wiki covariance.
giver det mening at udelade $\frac{1}{L}$ på begge sider her?

$$\begin{aligned}\mathbf{Y}_s\mathbf{Y}_s^T &= (\mathbf{A}\mathbf{X}_s)(\mathbf{A}\mathbf{X}_s)^T \\ &= \mathbf{A}\mathbf{X}_s\mathbf{X}_s^T\mathbf{A}^T \\ &= \mathbf{A}\Sigma_{X_s}\mathbf{A}^T \\ &= \mathbf{A}\Lambda\mathbf{A}^T + \mathbf{E} = \sum_{i=1}^{N}\Lambda_{ii}\mathbf{a}_i\mathbf{a}_i^T + \mathbf{E}\end{aligned} \tag{3.8}$$

Remember that N is the dimension of X hence the number of possible sources and k is the number of active sources. It has been shown that by this model it is possible to identify $O(M^2)$ sources given the true dictionary[14](dog er vektoriseringen også includeret i resultatet - så måske flyttes udtalelsen?). The purpose of the Cov-DL algorithm is instead to find the dictionary $\mathbf{A}$ from this expression and then still allow for $O(M^2)$ sources to be identified.

3.3.2 Determination of dictionary

In order enable the possibilities of identifying $O(M^2)$ sources and learning the corresponding dictionary A the model in (3.8) is rewritten. At first the both sides of the expression is vectorized. Because the covariance matrix is symmetric it is sufficient to vectorize only the lower triangular parts, including the diagonal. For this the function vec$(\cdot)$ is defined to map a symmetric $M \times M$ matrix into a vector of size $\frac{M(M+1)}{2}$ making a vectorisation of its lower triangular part.

er det muligt at vektoriceringen kun er til for at gøre dictionary learning problemet simplere? og ikke har noget med $O(M^2)$ at gøre, læs Pal2015

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

The vector $\boldsymbol{\delta}_s \in \mathbb{R}^N$ contains the diagonal entries of the source sample-covariance matrix $\boldsymbol{\Sigma}_{\mathbf{X}_s}$ and the matrix $\mathbf{D} \in \mathbb{R}^{M(M+1)/2 \times N}$ consists of the columns $\mathbf{d}_i = \text{vech}(\mathbf{a}_i \mathbf{a}_i^T)$. Note that $\mathbf{D}$ and δ_s are unknown while the left hand side is known from the observed data.

include noise in the original problem

From this expression it is now possible to learn $\mathbf{D}$ and then find the associated matrix $\mathbf{A}$. Comparing this expression to the original compressive sensing problem (3.1) it is clear that higher dimensionality is achieved by representing the problem in the covariance domain. Which allows for the number of active sources to exeed the number of observartions(måske for tidligt at konkludere her).

The Cov-DL method consists of two different algorithms for recovery of $\mathbf{D}$ and $\mathbf{A}$ depending on the number of total sources N relative to the number of observations M.

Cov-DL1 – overcomplete D

The case where $N > \frac{M(M+1)}{2}$ result in a overcomplete system similar to the original system being overcomplete when $N > M$....

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.
- i phd så forstås source localisation som at finde support for x og source identification forstås som at finde værdierne i de non-zero indgange(?)

Trine indhold:

- til CS afsnit: kilde på at nok sparse giver den rigtige løsning, se [39] fra phd
- Dictionary learning:
  - jointly solveing optimisation problem where g is introduced
  - her siges at k<M= er nødvendigt for recovery(af x går jeg ud fra) Fordi at enhver random dictionary kan bruges til at reprensentere y i dim M ved brug af M basis vektore i A.

  - altså når k > M så udgør de berørte søjler i A ikke længere en basis og x kan ikke bestemmes entydigt ? derfor er k<M et krav i standard algorithmer
  - der sættes så et nyt frame work hvor vi kan bruge standard algoritmer i vores tilfælde med k>M (forhåbenligt)

- EEG is non-stationary(afsnit 1.3.1 phd) source dynamics ændres med tiden afhængig af opave, fejlkilder kan være non stationære. så ICA dur ikke mixture model ICA var bedre til non stationær med statigvæk limited til k = M.

- the contribution er så at: support for X i MMV problem kan findes for k>M, altså altal non-zero entry k i x, med sufficient conditions ved sparse bayesian learning. herefter ved brug at uncorrelation af sources i et dictionary learning frame work så kan vi recover x når k>M.

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

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

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 $\hat{S} = \{i, \hat{\gamma}_i \neq 0\}$.

Algorithm 5 M-SBL

1. Given some data $\mathbf{y}$ and a dictionary $\mathbf{\Phi}$.
2. Initialise γ
3. Compute $\mathbf{\Sigma}$ and μ
4. Update γ using EM or Fixed-point
5. Repeat step 3 and 4 until convergence

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

[3] 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).

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

[5] 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).

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

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

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

[9] Friston, Karl J. “Functional and Effective Connectivity: A Review”. In: *BRAIN CONNECTIVITY* 1 (2011).

[10] Friston, Karl J. “Functional integration and inference in the brain”. In: *Progress in Neurobiology 590* 1-31 (2002).

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

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

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

[14] Pal, Piya and Vaidyanathan, P. P. “Pushing the Limits of Sparse Support Recovery Using Correlation Information”. In: ().

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

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

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

[18] *Studies within Steering of hearing devices using EEG and Ear-EEG*. `https://www.eriksholm.com/research/cognitive-hearing-science/eeg-steering`. Accessed: 2019-10-03.

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

[20] 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