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: 47

Date of Completion:
April 24, 2020

Abstract:

Here is the abstract

Preface

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

Aalborg University, April 24, 2020

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

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

Contents

Introduction

The problem addressed throughout this thesis arise from the increasing use of electroencephalographic measurements for a wide range of scientific purposes, especially within the medical field. An electroencephalography captures electric signals caused by activity within the brain. The signals from the brain is recorded over time by multiple sensors placed on the scalp. One essential issue concerning an electroencephalography is to extract the exact sources of the captured brain activity. This is of interest when studying correlation among activities in different parts of the brain, referred to as functional integration. The recorded signal from one sensor is basically a mixture of electric signals released from a various number of active neurons within the brain, forming one or several sources. Furthermore, this mixture is distorted as it travels through the scalp. The need for source extraction is confirmed by studies showing how analysis performed on electroencephalographic measurements differs significantly from similar analysis performed directly on the original source[13].

Considering this issue of source extraction from a mathematical perspective the electroencephalographic measurements can be modelled by a linear system of equations, from which it is possible to extract a limited number of sources under certain conditions. However, it is a general acknowledged issue that the true number of sources is unknown. The task complexity of extracting the sources from the linear system is increased in cases where the number of sources exceeds the number of sensors providing measurements.

This thesis explores a state of the art mathematical method for source extraction, embracing the case of more sources than sensors. Overall this method, published in 2015, consist of two steps, that is finding receptively the mixture the signals have undergone and then extracting the source signals. The two steps originates from two different approaches considering the mathematical orientation. The main goal of the thesis is to explore and unite the necessary theory into one algorithm. The practical aspect will include an implementation of the algorithm to be tested on new electroencephalographic measurements with the purpose of supporting(?) the current results. Furthermore the problem is connected to an current application within the hearing aid industry. Here the intention is to reduce the amount of energy spent by the hearing aid user. Basically, this is attempted by identifying the listening direction

intended by the user, from analysis of the active sources measured on the user. In this thesis the number of active sources are sought related to the amount of energy used by the hearing aid user. This includes considerations upon the issue of the true number of active sources being unknown.

skal opdateres

(følgende kan skrives bedre hvis ikke det skal være et andet sted)The thesis consist of a motivational part introducing electroencephalography and the potential use within research especially in the hearing aid industry. Furthermore, existing literature considering different mathematical approaches for source extraction are examined. The Motivational part is concluded by the problem statement specifying the objective of the thesis. Next is the theoretical part. The system model is specified and the solution approach are presented. The necessary theory are introduced leading to the state of the art algorithms for source extraction. The theoretical part is followed by implementation and test of the algorithm for verification. Next is.. Finally discussion and conclusion upon the achieved results are presented followed by a consideration upon further studies.

Chapter 1

Motivation

This chapter accounts for the motivation behind source extraction from an Electroencephalography (EEG). The concept of EEG is introduced along with current applications. The potential and importance of source extraction are considered and related to the hearing aid industry. The commonly applied mathematical model for EEG measurements is presented. Currently applied methods for source extraction are considered leading to a presentation of the current state of the art methods which succeeds to overcome the limitations of previous methods. Lastly the objective of this thesis is specified.

1.1 Introduction to EEG Measurements

EEG is an imaging technique used within the medical field. EEG is measuring electric signals on the scalp, caused by brain activity. The human central nerve system consist of various nerve cells connecting the neurons within the brain. Nerve cells respond to certain stimuli, for instance a physical stimuli, and transmit informations between neurons. Generally speaking these activities induce local currents that are transferred throughout the nerve system. Several nearby simultaneous activations result in local potential fields, referred to as one signal *source*[19]. EEG measurements are provided by a number of metal electrodes, referred to as sensors, carefully placed on the human scalp. Each sensor reads the present electrical signals over time. For the source signal to reach a sensor it has to penetrate the skull, skin and several other thin layers of biological tissue. This causes an unknown distortion and reduction of a signal. It is most likely that the measurement of one sensor is a sum of multiple signals from different sources. Nor is the range of a single sensor separated from the other sensors. Thus the same signal can easily be measured by two or more sensors. The process of distorsion and mixing of signals is called volume conduction [19, p. 68] [20]. From this it is clarified that EEG measurements is a mixture of fluctuating electrical signals originating from brain activities. Due to the mixing and the nature of the signals the

true number of sources is generally considered unknown[19]. Furthermore, EEG is a subject for interfering noise. Noise signals can occur in the measurements resulting from physical movement of e.g. eyes and jawbone [22]. The concept of volume conduction is sought illustrated on figure 1.1.

The source signals are classified within 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 wave for adults, associated with active thinking, active attention or solving concrete problems [19, p. 11]. An example of EEG measurements within the four categories is illustrated by figure 1.2.

does this comply to the non-gaussian assumption of ICA?

Generally, the distribution of EEG measurements of multiple sensors are considered multivariant Gaussian[19, p. 50]. Though the mean and covariance properties generally changes over time. Therefore EEG measurements are considered quasi-stationary i.e. stationary only within small intervals. This motivates the need for segmentation of the EEG measurements to achieve signals with similar characteristics.

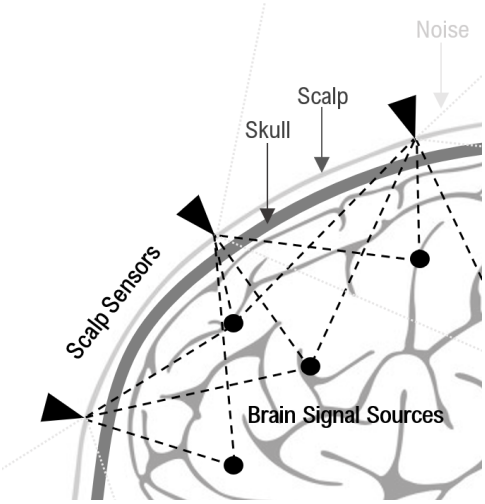

Figure 1.1: Illustration of volume conduction

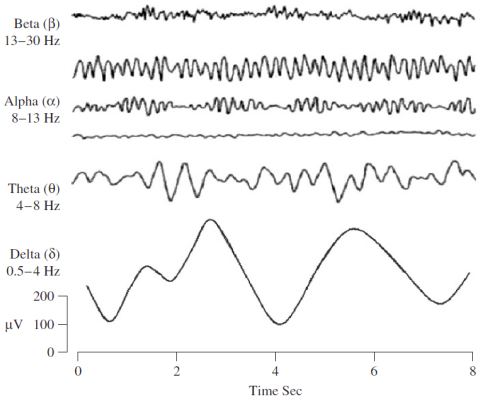

Figure 1.2: Example of time dependent EEG measurements within the four defined categories, source: [19]

1.1.1 Application

EEG performed on humans and animals have a great number of applications with both clinical and research purposes. Examples of clinical applications covers diagnosis and management of neurological disorders such as epilepsy and monitor alertness regarding coma or brain death. EEG capitalizes on the procedure being non-invasive

and fast. Neural activity can be measured within fractions of a second after a stimuli has been provided. These advantages contributes to the wide range of applications within research of the neural processes involved in or resulting from actions, emotions or cognition. Today such neural research are use in many different fields [22, p. 4]. The hearing aid industry is one example where this research is highly prioritized. At Eriksholm research center, which is a part of the hearing aid manufacturer Oticon, cognitive hearing science is a research area within fast development [21]. One main purpose at Eriksholm is to make it possible for a hearing aid to identify the user-intended sound source from real time EEG measurements and thereby exclude noise from elsewhere [2] [6]. It is essentially the well known but unsolved cocktail problem which is sought improved by use of EEG. This is where EEG and occasionally so called in-ear EEG is interesting. In conjunction with the technology of beamforming it is possible for a hearing aid to receive only signals from a specific direction.

Over the past two decades, functional integration has become an area of interest regarding EEG research [12]. Within neurobiology functional integration refers to the study of the correlation among activities in different regions of the brain. In other words, how do different parts of the brain work together to process information and conduct a response [13]. For this purpose separation and localization of the original sources which contribute to the EEG measurement is of interest. An article from 2016 [20] points 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 does not allow interpretations about the interaction between sources. This emphasize a potential for improving results within a wide range of EEG research, if the original active sources can be extracted from a specific EEG measurements.

1.1.2 Modelling

Consider the issue of extracting the activated sources from EEG measurements on the scalp. A known approach is to model the observed data by a linear system

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

The vector $\mathbf{y} \in \mathbb{R}^M$ is the EEG measurement of one time sample containing M sensor measurements. $\mathbf{x} \in \mathbb{R}^N$ is the corresponding N sources within the brain. The non-zero entries of $\mathbf{x}$ represent the active sources at the time of the measurement. $\mathbf{A} \in \mathbb{R}^{M \times N}$ is an unknown projection/transformation(?) matrix, also referred to as the mixing matrix resembling the volume conduction. The i-th column of $\mathbf{A}$ represents the relative projection weights from the i-th source to every sensor [5]. Representing one time sample the linear system is in general referred to as a single measurement vector model. It it only the measurement vector $\mathbf{y}$ that is known hence it is not possible to solve the linear system with respect to $\mathbf{x}$ using basic linear algebra.

The task in this case is to identify both $\mathbf{A}$ and then $\mathbf{x}$, given the measurement vector $\mathbf{y}$. This problem is referred to as the inverse problem of EEG. Finding $\mathbf{x}$ from the inverse problem is referred to as source separation and localization. Separation is to find the signal of each active source and localization is to place each active source signal at the right position within the source vector of dimension N, where N is the maximum number of sources to be active .

der er vel ikke N aktive source her?

Independent Component Analysis (ICA) is one commonly applied method to solve the inverse problem of EEG [16], [15]. ICA is a technique to find the matrix $\mathbf{A}$ such that the column wise elements of $\mathbf{X}$ is statistically independent. Thus statistical independence between the active sources is the essential assumption, which in the case of EEG are considered valid due to the volume conduction being effectively instantaneous [15, p. 3]. Application of ICA has shown great results regarding source separation of high-density EEG. However, a significant flaw to this method is that the EEG measurements are only separated into a number of sources that is equal to or less than the number of sensors [3]. Meaning that the EEG inverse problem can not be solved when it forms an under-determined system, which is the case when the maximum number of unknown sources N exceeds the number of sensors M. Such assumption undermines the reliability and usability of ICA, as the number of active sources easily exceed the number of sensors [5]. This is especially a drawback when low-density EEG are considered. Low-density EEG measurements are collected from equipment with less than 32 sensors, increasing the changes of M being less that N. However, improved capabilities of low-density EEG devices are desirable due to their relative low cost, mobility and ease to use.

This argues the importance of considering the inverse problem of EEG in the under-determined case where $N > M$. In the next section existing work considering the under-determined inverse problem of EEG is investigated further.

1.2 Related Work and Our Objective

As mentioned above ICA is a solid method for source separation in the case where separation into a number of sources equal to the number of sensors is adequate. The issue occurs in cases where the number of sources N exceeds the number of sensors M. To overcome this issue an extension of ICA was suggested, referred to as the ICA mixture model [3]. Instead of identifying one overcomplete mixing matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$ this approach learns N_{model} different mixing matrices $\mathbf{A}_i \in \mathbb{R}^{M \times M}$, to make computations more tractable. This method was further adapted into the Adaptive Mixture ICA (AMICA) which showed successful results regarding identification of more sources than sensors [18]. However, these successful results relies on the assumption that no more than M out of N possible sources is simultaneously active. That is explicit that the source vector of dimension N has at most M non-zero entries. This assumption is still an essential limitation to the frame work,

especially when considering low-density EEG. Other types of ICA algorithms for under-determined systems have been proposed, without overcoming the limitation of jointly active sources exceeding the number of sensors.

In 2015 O. Balkan et. al. suggested a new approach also targeting the identification of more active sources than sensors regarding EEG measurements. One method is proposed for learning $\mathbf{A}$ from $\mathbf{y}$ [3] and a different method is proposed for finding $\mathbf{x}$ given $\mathbf{y}$ and $\mathbf{A}$ [4].

To learn $\mathbf{A}$ the suggested method, referred to as Cov-DL, is a covariance-domain based dictionary learning algorithm. The method is based upon theory of dictionary learning and compressive sensing. Which dictates a framework for solving an under-determined system when $\mathbf{x}$ contains a sufficiently amount of zeros. This is similar to the constraint of ICA. However, to overcome this the point is to transfer the EEG measurements into the covariance domain. In the covariance domain a higher dimensionality can be achieved compared to the original EEG sensor domain with dimension M. The transformation can be done when assuming a linear volume conduction and uncorrelated sources. As a result the theory of compressive sensing is found to apply to the covariance domain, allowing to learn $\mathbf{A}$ by dictionary learning – even in the case where the active sources exceeds the number of measurements.

The Cov-DL algorithm stands out from 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 and found to outperform AMICA [3]. As mentioned, the Cov-DL algorithm only learns the mixing matrix $\mathbf{A}$, resembling the volume conduction.

is it okay to mention sparsity here for the first time?

For the purpose of recovering $\mathbf{x}$, from $\mathbf{y}$ and $\mathbf{A}$, a multiple measurement sparse Bayesian learning (M-SBL) algorithm is proposed. This method is also targeting the case of more active sources than sensors.The method was proven to outperform the previously used algorithms, even when the defined recovery conditions regarding the found mixing matrix $\mathbf{A}$ was not fulfilled [4].

yderligere beskrivelse nødvedig? eller fjerne lidt fra Cov måske, da det samme lidt kommer i kap 3?

One drawback, which is not fully covered in the referred literature, is that the two methods rely on the number of active sources being known. In practise this is not the case. Hence an estimation of the number of active sources has to be considered for the algorithm to be useful in practice. To address this issue a simple approach is to optimise the result with respect to the number active source, provided that some prior assumption of the expected result can be made.

The two state of the art methods resulting in source separation and localization will make the foundation of this thesis. Our aim is to investigate and fully understand the two methods in order to implement and test a joint algorithm – recovering the original sources $\mathbf{x}$ from the measurements $\mathbf{y}$, when the number of active sources exceeds the number of measurements. Secondary it is of interest to consider the practical application of the algorithm, for instance within a hearing aid as described

in section 1.1. As mentioned, the number of active sources is in general unknown in practise thus it is first of all an estimation of the number of active sources which is of interest for practical use of the algorithm. For this we want to investigate whether it is possible to estimate the number of active sources, through optimization.

Chapter 2

Problem Statement

EEG scalp measurements, a mixture of fluctuating electrical signals originating from brain activities and noise, due to distorting elements such as scalp and biological tissues, can be described as a linear system

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

$\mathbf{Y}$ is the EEG scalp measurements measured from M sensors placed on the scalp, $\mathbf{A}$ is the mixing of the electrical signals denoted as the mixing matrix and $\mathbf{X}$ are the N original electrical signals, denoted as sources. Only the EEG measurements $\mathbf{Y}$ is known and it is of interest to identify the mixing matrix $\mathbf{A}$ and hereby the original sources $\mathbf{X}$. The original sources have been shown significant for practical use compared to the raw EEG scalp measurements. Especially, the under-determined case with more sources than sensors is of interest, resulting from low-density EEG devices which is beneficial due to low cost and easy application. In the linear algebraic sense an under-determined linear system have infinitely solution provided a solution exists and is therefore difficult to solve. Two state of the art methods are seen to solve the issue with success, the covariance-domain dictionary learning (Cov-DL) and multiple sparse Bayesian learning (M-SBL) algorithms. The Cov-DL algorithm recovers the mixing matrix from the given measurements $\mathbf{Y}$ while the M-SBL algorithm localised and identify the sources from the recovered mixing matrix and measurements. By combining the two state of art methods into one this could solve the inverse EEG problem – the identification of $\mathbf{A}$ and $\mathbf{X}$ given $\mathbf{Y}$. However, the algorithms used the knowledges of the number of activations within the sources as this is a unknown variable in practice. Hence a modification of the combined state of art methods is sought to increase the potential for practical use.

This motivates the following problem statement.

Based on state of the art method, how can we reproduce the recovering of original sources of brain activity from the EEG inverse problem, in the under-determined

case, and how can this be modified to increase the potential of practical use such as the unknown brain activity?

From the problem statement the following sub-questions is established for clarification.

- Can we reproduce the Cov-DL algorithm to estimate a mixing matrix $\mathbf{A}$ from a over-complete EEG inverse problem with synthetic and realistic EEG scalp measurements?
- Can we reproduce the M-SBL algorithm to estimate a source matrix $\mathbf{X}$ from a over-complete EEG inverse problem with synthetic and realistic EEG scalp measurements?
- How can the number of active sources be estimated, based only on the EEG scalp measurements?

Chapter 3

System Model

Through this chapter a model representing the EEG measurements is specified. Along with the model different terminologies is introduced and described for further use in this thesis. At last the solution approach for estimating the model variables(rather than parameters right?) is described, setting the outline of the remaining chapters of the thesis.

3.1 System of Linear Equations

Let $\mathbf{y} \in \mathbb{R}^M$ be some vector. By basic linear algebra $\mathbf{y}$ can be described as a linear combination of a coefficient matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$ and some scalar vector $\mathbf{x} \in \mathbb{R}^N$ such that

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

Let $\mathbf{y}$ and $\mathbf{A}$ be known, then 3.1 makes a system of M linear equations with N unknowns, referred to as a linear system.

To solve the linear system 3.1 with respect to $\mathbf{x}$ one must look at the three different cases that can occur, depending on the relation between the number of equations M and the number of unknowns N. For $M = N$, the system has one unique solution, provided that a solution exist. If the square coefficient matrix $\mathbf{A}$ has full rank the solution can be found by inverting $\mathbf{A}$.

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}.$$

For $M > N$ the system is over-determined, having more equations than unknown. In general there is no solution to an over-determined system. ~~An~~ /The exception occur when the system contains a sufficient amount of linearly dependent equations. For $M < N$ the system is under-determined, having fewer equations than unknowns. There exist infinitely many solutions to an under-determined system, provided that one solution exist[8, p. ix].

note: skal vi nævne løsnings metoder for underdetermined system her?

Consider now $\mathbf{y} \in \mathbb{R}^M$ as the observed measurements provided by M EEG sensors at time t. The linear system 3.1 is then considered as a single measurement vector (SMV) model. Modelling the EEG measurements by the SMV model embody the following interpretations, based on chapter 1. Remember from chapter 1 that EEG measurements basically is a mixture of original brain signals affected by volume conduction. $\mathbf{x}$ is seen as the original brain signal sources, each entry representing the signal of one source. Thus, $\mathbf{x} \in \mathbb{R}^N$ is referred to as the source vector. N is considered the maximum number of sources, however zero-entries may occur. Let k denote the number of non-zero entries in $\mathbf{x}$, referred to as the active sources at time t. The projection matrix $\mathbf{A}$, referred to as the mixing matrix, models the volume conduction by mapping the source vector from $\mathbb{R}^N$ to $\mathbb{R}^M$, where M is the number of sensors hence the dimension of the measurement vector $\mathbf{y}$.

3.2 Multiple Measurement Vector Model of EEG

In practise EEG measurements are sampled over time by a certain sample frequency. Thus multiple EEG measurement vectors are achieved. Let L be the total number of samples. Now the the SMV model is expanded to include L measurement vectors:

$$\mathbf{Y} = \mathbf{AX} + \mathbf{E}, \tag{3.2}$$

now $\mathbf{Y} \in \mathbb{R}^{M \times L}$ is the observed measurement matrix, $\mathbf{X} \in \mathbb{R}^{N \times L}$ is the source matrix, and $\mathbf{A} \in \mathbb{R}^{M \times N}$ is the mixing matrix. Furthermore $\mathbf{E} \in \mathbb{R}^{M \times L}$ is consider an additional noise matrix, to be expected from psychical measurements. The model is now referred to as the multiple measurement vector (MMV) model. As for (3.1) the solution set of the linear system (3.2) depends on the relation between N and M [8, p. 42].

In chapter 1 it is specified that the case of more sources than sensors, $N > M$, is the case of interest in this thesis.

3.2.1 Segmentation

In chapter 1 it is argued that EEG measurements are only stationary within small segments. Hence the following segmentation is considered.

Let f be the sample frequency of the observed EEG measurements $\mathbf{Y}$ and let t be a time interval in seconds determining the duration of one segment. Here s is the segment index. As such the observed EEG measurements can be divided into stationary segments $\mathbf{Y}_s \in \mathbb{R}^{M \times L_s}$, possibly overlapping, where $L_s = tf$ is the number of samples within one segment. For each segment the MMV model (3.2) holds and is rewritten into

$$\mathbf{Y}_s = \mathbf{AX}_s + \mathbf{E}_s, \quad \forall s. \tag{3.3}$$

Due to a segment being stationary it is assumed that each source remains either active or non-active throughout the segment. Thus, $\mathbf{X}_s$, consists of k non-zero rows – the active sources.

In order to characterise the source matrix with respect the amount of non-zero rows the term row sparseness is considered. By common definition the support of the segmented source matrix $\text{supp}(\mathbf{X}_s)$ denotes the index set of non-zero rows of $\mathbf{X}_s$. To count the non-zeros row of a matrix the ℓ_0-norm is defined:

$$\|\mathbf{X}\|_0 := \text{card}(\text{supp}(\mathbf{X})),$$

are you counting elements or rows?

where the function $\text{card}(\cdot)$ gives the cardinality of the input set. $\mathbf{X}_s$ is said to be k-sparse if it contains at most k non-zeros rows:

$$\|\mathbf{X}_s\|_0 \leq k$$

A model for the EEG measurements is now established. From the model the aim is to recover the source matrix $\mathbf{X}_s \forall s$, which gives us the separated original brain signals as intended by the problem statement. In the next section the solution method is presented and discussed – outlining the remaining chapters of the thesis.

3.3 Solution Method

Denne section er ikke blevet opdateret i forhold til den nye problemformulering. Altså vi mangler at inddrage 'reproducerbarhed'.

It is now justified that the EEG measurements can be modelled by the multiple measurement vector model defined by the system of linear equations (3.3), including an additional noise. By the problem statement cf. chapter 2 the aim is to recover the source vector $\mathbf{X}$, in the case where the number of sensors is less than the number of sources, $M < N$. That is recovering $\mathbf{X}$ from an under-determined linear system. Therefore, the solution must be found in the infinite solution space, provided that one solution exists, thus simple linear algebra can not be used. However, by considering numerical methods such as mathematical optimization it is possible restrict the solution by some constraint and then find the unique optimal solution which respect to a defined cost relative to the solution. The theory of compressive sensing dictates a framework for solving an under-determined system when $\mathbf{X}$ is known to have non-zeros entries. Specifically a unique solution $\mathbf{X}$ can be found when $\mathbf{X}$ is M-sparse, cf. theorem B.1.1 in appendix B.1. When $\mathbf{A}$ is unknown, as it is in the current case, the concept of dictionary learning can be used to determine $\mathbf{A}$, again under the assumption that $\mathbf{X}$ is M-sparse.

an approximation of X?

this is unclear should it not be when X is sparse?

As discussed in chapter 1 the aim of this thesis is to overcome the limitation of fewer sources than measurements, which is the limitation of compressive sensing.

a

A method to overcome this limitation, with respect to learning $\mathbf{A}$, is the Covariance-domain dictionary learning (Cov-DL) method[3], introduced in chapter 1. The

method manage to leverage the increased dimensionality of the covariance domain in order to the allow the theory of compressive sensing to apply to an under-determined system. However, this method does only apply to the process of learning $\mathbf{A}$, hence a different approach is necessary to recover $\mathbf{X}$.

as the relation between $\mathbf{X}$ and the corresponding δ is unknown(?) det kan vi ikke skrive her kan vi?

For recovering $\mathbf{X}$, given both $\mathbf{Y}$ and $\mathbf{A}$, the method Multiple sparse Bayesian learning (M-SBL), introduced in chapter 1, is considered. A method which ensure that the sparsity holds(?). O. Balkan [4] did also, in 2014, proposed a method which could identify the sources, in the time-domain, by creating a likelihood which ensure the wanted sparsity of the source matrix $\mathbf{X}$ and controlled by some variance. This method is called multiple sparse Bayesian learning (M-SBL) and takes advantage of a Bayesian approach. In [4] a variance dependent log-likelihood which has been induce by a empirical prior that ensure sparsity of the likelihood has been constructed to be minimised with respect to the variance. From the log-likelihood an estimate for the source matrix $\mathbf{X}$ is drawn with respect to the support set S which has been influence by variance used in the minimization.

Det er vel ikke metoden som gør dette?

?

evt. tilføj afslutning der siger hvad der kommer i kapitlerne?

Consider rewriting this sentence. It does not read well.

Chapter 4

Covariance-Domain Dictionary Learning

Through this chapter the method Covariance-domain dictionary learning (Cov-DL) is presented in details. Along the presentation of the general method, necessary computational details are derived for the practical solution. The purpose is to recover the mixing matrix $\mathbf{A}$ from the MMV model, derived in chapter 3, in the under-determined case. In the context of compressive sensing the matrix $\mathbf{A}$ in the MMV model is referred to as the dictionary matrix. That is the true mixing matrix is estimated as a dictionary matrix, which can take different forms. This will elaborated further in the section of dictionary learning.

obs! stemmer dette på tværs a kapitlerne, vi siger A bliver estimeret som en dictionary.

Cov-DL is an algorithm proposed by O. Balkan [3], leveraging the increased dimensionality of the covariance domain. The method have shown successful recovering of the mixing matrix $\mathbf{A}$, even in the non-sparse under-determined case with more active sources k than available measurements M, $k \geq M$. In short the algorithm consist of three steps. First the segmented MMV model of the EEG measurements is transformed into the covariance domain. Then, by the increased dimensionality of the covariance domain, it is possible to learn the mixing matrix of the covariance domain, denoted by $\mathbf{D}$, based on the theory of compressive sensing. Here two different cases will appear dependent on the relation between the number of sources N and the found dimension of the covariance domain, which of course depends on the number of measurements M. Lastly, an inverse transformation is performed on the found mixing matrix of the covariance domain $\mathbf{D}$, in order to obtain the wanted mixing matrix $\mathbf{A}$. An important aspect of this method is the prior assumption that the sources within one segment are uncorrelated, that is the rows of $\mathbf{X}_s$ being mutually uncorrelated.

The section is inspired by chapter 3 in [5] and the article [3]. Selected general theory supporting essential parts of the method is elaborated in appendix B.

4.1 Covariances Domain Representation

Consider a single sample vector $\mathbf{y}_i$, containing EEG measurements. The covariance of $\mathbf{y}_i$ is to be defined by

$$\mathbf{\Sigma}_{\mathbf{y}_i} = \mathbb{E}[(\mathbf{y}_i - \mathbb{E}[\mathbf{y}_i])(\mathbf{y}_i - \mathbb{E}[\mathbf{y}_i])^T],$$

where $\mathbb{E}[\cdot]$ is the expected value operator. Assume that all samples vectors $\mathbf{y}_i$ within one segment has zero mean and the same distribution. Then, the observed segmented EEG measurements matrix $\mathbf{Y}_s \in \mathbb{R}^{M \times L_s}$ is to be described in the covariance domain by the sample covariance $\widehat{\mathbf{\Sigma}}$ which is defined as the covariance among the M measurements across the L_s samples. That is a $M \times M$ matrix $\mathbf{\Sigma}_{\mathbf{Y}_s} = [\sigma_{jk}]$ with entries

$$\sigma_{jk} = \frac{1}{L_s} \sum_{i=1}^{L_s} y_{ji} y_{ki}.$$

Using matrix notation the sample covariance of $\mathbf{Y}_s$ can be written as

$$\widehat{\mathbf{\Sigma}}_{\mathbf{Y}_s} = \frac{1}{L_s} \mathbf{Y}_s \mathbf{Y}_s^T.$$

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

$$\widehat{\mathbf{\Sigma}}_{\mathbf{X}_s} = \frac{1}{L_s} \mathbf{X}_s \mathbf{X}_s^T = \mathbf{\Lambda}_s + \varepsilon$$

From the assumption of uncorrelated sources within $\mathbf{X}_s$ the sample covariance matrix is expected to be nearly diagonal, thus it can be written as $\mathbf{\Lambda}_s + \varepsilon$ where $\mathbf{\Lambda}_s$ is a diagonal matrix consisting of the diagonal entries of $\widehat{\mathbf{\Sigma}}_{\mathbf{X}_s}$ and ε is the estimation error[3]. Each segment is then modelled in the covariance domain as

$$\begin{aligned}
\widehat{\mathbf{\Sigma}}_{\mathbf{Y}_s} = \frac{1}{L_s} \mathbf{Y}_s \mathbf{Y}_s^T &= \frac{1}{L_s} (\mathbf{A}\mathbf{X}_s + \mathbf{E}_s)(\mathbf{A}\mathbf{X}_s + \mathbf{E}_s)^T \\
\mathbf{Y}_s \mathbf{Y}_s^T &= (\mathbf{A}\mathbf{X}_s)(\mathbf{A}\mathbf{X}_s)^T + \mathbf{E}_s \mathbf{E}_s^T + \mathbf{E}_s (\mathbf{A}\mathbf{X}_s)^T + \mathbf{A}\mathbf{X}_s \mathbf{E}_s^T \\
&= \mathbf{A}\mathbf{X}_s \mathbf{X}_s^T \mathbf{A}^T + \mathbf{E}_s \mathbf{E}_s^T + \mathbf{E}_s \mathbf{X}_s^T \mathbf{A}^T + \mathbf{A}\mathbf{X}_s \mathbf{E}_s^T \\
&= \mathbf{A}(\mathbf{\Lambda}_s + \varepsilon)\mathbf{A}^T + \mathbf{E}_s \mathbf{E}_s^T + \mathbf{E}_s \mathbf{X}_s^T \mathbf{A}^T + \mathbf{A}\mathbf{X}_s \mathbf{E}_s^T \\
&= \mathbf{A}\mathbf{\Lambda}_s \mathbf{A}^T + \mathbf{A}\varepsilon\mathbf{A}^T + \mathbf{E}_s \mathbf{E}_s^T + \mathbf{E}_s \mathbf{X}_s^T \mathbf{A}^T + \mathbf{A}\mathbf{X}_s \mathbf{E}_s^T \qquad (4.1) \\
&= \mathbf{A}\mathbf{\Lambda}_s \mathbf{A}^T + \widetilde{\mathbf{E}} \qquad (4.2)
\end{aligned}$$

From (4.1) to (4.2) all terms where noise is included are defined as a united noise term $\widetilde{\mathbf{E}}$. By vector notation (4.2) is rewritten to be vectorized. Because the covariance matrix $\widehat{\mathbf{\Sigma}}_{\mathbf{Y}_s}$ is symmetric it is sufficient to vectorize only the lower triangular parts,

including the diagonal. For this the function $\text{vec}(\cdot)$ is defined to map a symmetric $M \times M$ matrix into a vector of size $\frac{M(M+1)}{2} := \widetilde{M}$ making a row-wise vectorization of its upper triangular part. Furthermore, let $\text{vec}^{-1}(\cdot)$ be the inverse function for devectorisation. This results in the following model

$$\widehat{\mathbf{\Sigma}}_{\mathbf{Y}_s} = \sum_{i=1}^{N} \mathbf{a}_i \mathbf{\Lambda}_{s_{ii}} \mathbf{a}_i^T + \widetilde{\mathbf{E}}$$

$$\begin{aligned} \text{vec}(\widehat{\mathbf{\Sigma}}_{\mathbf{Y}_s}) &= \sum_{i=1}^{N} \text{vec}(\mathbf{a}_i \mathbf{a}_i^T) \mathbf{\Lambda}_{s_{ii}} + \text{vec}(\widetilde{\mathbf{E}}) \\ &= \sum_{i=1}^{N} \mathbf{d}_i \mathbf{\Lambda}_{s_{ii}} + \text{vec}(\widetilde{\mathbf{E}}) \\ &= \mathbf{D}\boldsymbol{\delta}_s + \text{vec}(\widetilde{\mathbf{E}}), \quad \forall s. \end{aligned} \tag{4.3}$$

Here $\boldsymbol{\delta}_s \in \mathbb{R}^N$ contains the diagonal entries of the source sample-covariance matrix $\mathbf{\Lambda}_s$ and the matrix $\mathbf{D} \in \mathbb{R}^{\widetilde{M} \times N}$ consists of the columns $\mathbf{d}_i = \text{vec}(\mathbf{a}_i \mathbf{a}_i^T)$. Note that $\mathbf{D}$ and $\boldsymbol{\delta}_s$ are unknown while $\text{vec}(\widehat{\mathbf{\Sigma}}_{\mathbf{Y}_s})$ is known from the observed data. By this transformation to the covariance domain one segments is now represented by the single measurement model with $\widetilde{M}$ "measurements". It has been shown that this transformed model allows for identification of $k \leq \widetilde{M}$ active sources [17], which is a much weaker sparsity constraint than the original sparsity constraint $k \leq M$. The purpose of the Cov-DL algorithm is to leverage this model to find the dictionary $\mathbf{A}$ from $\mathbf{D}$ and then still allow for $k \leq \widetilde{M}$ active sources to be identified. That is the number of active sources are allowed to exceed the number of observations as intended.

4.2 Determination of the Mixing Matrix

The goal is now to learn first $\mathbf{D}$ and then the associated mixing matrix $\mathbf{A}$. Two methods are considered relying on the relation of M and N. For now the noise vector is ignored.

4.2.1 Under-determined system

When $N > \widetilde{M}$ the transformed model (4.3) makes an under-determined system. This is similar to the original MMV model (3.2) being under-determined when $N > M$. Thus, it is from the theory of compressive sensing again possible to solve the under-determined system if a certain sparsity is withhold. Namely $\boldsymbol{\delta}_s$ being $\widetilde{M}$-sparse. Assuming the sufficient sparsity on $\boldsymbol{\delta}_s$ is withhold it is possible to learn the dictionary matrix of the covariance domain $\mathbf{D}$ by traditional dictionary learning methods applied

to the observations represented in the covariance domain $\text{vec}(\widehat{\mathbf{\Sigma}}_{\mathbf{Y}_s})$ for all segments s.

Dictionary Learning

Within the theory of compressive sensing the matrix $\mathbf{A}$ is referred to as a dictionary matrix, as it determines how a sparse vector $\mathbf{x}$ is transformed to the original non-sparse signal. When the dictionary is not known i prior it is essential how to choose the the dictionary matrix in order to achieve the best recovery, of the sparse vector $\mathbf{x}$ from the measurements $\mathbf{y}$. This is clarified from the proof of theorem B.1.1 in appendix B.1. One choice is a pre-constructed dictionary. In many cases the use of a pre-constructed dictionary results in simple and fast algorithms for reconstruction of $\mathbf{x}$ [10]. However, a pre-constructed dictionary is typically fitted to a specific kind of data. For instance the discrete Fourier transform or the discrete wavelet transform are used especially for sparse representation of images [10]. Hence the results of using such dictionaries depend on how well they fit the data of interest, which is creating a certain limitation.

The alternative option is to consider an adaptive dictionary based on a set of training data that resembles the data of interest. For this purpose learning methods are considered to empirically construct a fixed dictionary which can take part in the application. There exist several dictionary learning algorithms. One is the K-SVD algorithm which was presented in 2006 by Elad et al. and found to outperform pre-constructed dictionaries, when computational cost is of secondary interest [1]. The concept of the K-SVD algorithm is introduced, and the more detailed algorithm is to be found in appendix B.2. Consider the measurement matrix $\mathbf{Y} \in \mathbb{R}^M$ consisting of measurement vectors $\{\mathbf{y}_j\}_j^L$ making a set of L training examples forming a linear system

$$\mathbf{y}_j = \mathbf{A}\mathbf{x}_j.$$

from which one can learn a suitable dictionary $\hat{\mathbf{A}}$, and the sparse representation of the source matrix $\hat{\mathbf{X}} \in \mathbb{R}^N$ with the source vectors $\{\hat{\mathbf{x}}_j\}_j^L$. For a known sparsity constraint k the dictionary learning can be defined by the following optimisation problem.

$$\min_{\mathbf{A},\mathbf{X}} \sum_{j=1}^{L} \|\mathbf{y}_j - \mathbf{A}\mathbf{x}_j\|_2^2 \quad \text{subject to} \quad \|\mathbf{x}_j\|_1 \leq k,\ 1 \leq j \leq L. \tag{4.4}$$

where both $\mathbf{A}$ and $\mathbf{x}_j$ are variables to be determined. Learning the dictionary by the K-SVD algorithm constitute joint solving of the optimization problem with respect to $\mathbf{A}$ and $\mathbf{X}$ respectively. An initial $\mathbf{A}_0 = [\mathbf{a}_1, \cdots, \mathbf{a}_N]$ and the corresponding $\mathbf{X}_0$ is determined. Then, for each iteration an update rule is applied to each column of $\mathbf{A}_0$, that is updating first $\mathbf{a}_j$ and then the corresponding row $\mathbf{x}_i$. More details on

a_1?

which x_i is the corresponding row to the column a_j?

the K-SVD algorithm is found in appendix B.2. The uniqueness of $\mathbf{A}$ depends on the recovery sparsity condition. As clarified earlier in 3.3 the recovery of a unique solution $\mathbf{X}^*$ is only possible if $k < M$ [5].

Application of dictionary learning

By the establishments of an dictionary learning algorithm it is now used to learn the transformed dictionary matrix $\mathbf{D}$ in (4.3). Here the transformed and vectorised measurements $\left\{\text{vec}\left(\hat{\boldsymbol{\Sigma}}_{\mathbf{Y}}\right), \forall s\right\}$ makes the training dataset. By this note that each segment the original measurement sample constitute only one sample in the covariance domain. Thus the number of training samples depends on the length of a segment. When K-SVD is applied and $\mathbf{D}$ is found it is possible to estimate the mixing matrix $\mathbf{A}$ that generated found $\mathbf{D}$ through the relation

$$\mathbf{d}_j = \text{vec}(\mathbf{a}_j \mathbf{a}_j^T).$$

not entirely clear why you need to invert vec(d_j)?

Here each column is found from the optimisation problem

$$\min_{\mathbf{a}_j} \|\text{vec}^{-1}(\mathbf{d}_j) - \mathbf{a}_j \mathbf{a}_j^T\|_2^2,$$

for which the global minimizer is $\mathbf{a}_j^* = \sqrt{\lambda_j}\mathbf{b}_j$. Here λ_j is the largest eigenvalue of $\text{vec}^{-1}(\mathbf{d}_j)$,

redegørelse for resultatet her skal laves

$$\text{vec}^{-1}(\mathbf{d}_j) = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1N} \\ d_{21} & d_{22} & \cdots & d_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ d_{N1} & d_{N2} & \cdots & d_{NN} \end{bmatrix}, \quad j \in [N]$$

and $\mathbf{b}_j$ is the corresponding eigenvector.

By this each column of the mixing matrix $\mathbf{A}$ can be estimated hence it is possible to determine the mixing matrix in the case where the measurements transformed into the covariance domain makes an under-determined system, but the necessary sparsity constraint, $\boldsymbol{\delta}_s$ being $\widetilde{M}$-sparse (instead of M-sparse), is withhold.

good to restate here how $\widetilde{M}$ is defined.

4.2.2 Over-determined system

Consider again the measurements represented in the covariance domain (4.3). In the case of $N < \widetilde{M}$ an over-determined system is achieved where $\mathbf{D}$ is high and thin. In general such system is inconsistent. Thus it is not possible to find $\mathbf{D}$ by traditionally dictionary learning methods and different methods most be considered.

tall?

tjek, har vi tidligerer nævnt at der teoretisk godt kan være en løsning?

When $N < \widetilde{M}$ it is certain from the model (4.3) that the transformed measurements $\text{vec}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y}_s})$ will live on or near a subspace of dimension N. This subspace is spanned by the columns of $\mathbf{D}$, and is denoted as $\mathcal{R}(\mathbf{D})$. To learn $\mathcal{R}(\mathbf{D})$ without having to impose any sparsity constraint on $\boldsymbol{\delta}_s$ it is possible to use Principal Component

Analysis(PCA). When PCA is applied to the set $\{\text{vec}(\widehat{\mathbf{\Sigma}}_{\mathbf{Y}_s}), \forall s\}$ a set of N principal components are found. The principal components forms a set of basis vectors $\mathbf{U}$ such that $\mathcal{R}(\mathbf{U}) = \mathcal{R}(\mathbf{D})$. However, this do not imply that $\mathbf{D} = \mathbf{U}$. In the case of two sets of basis vectors spanning the same space, namely $\mathcal{R}(\mathbf{U}) = \mathcal{R}(\mathbf{D})$, the projection operator of the given subset must be unique(need prove here? or is there just one $P : V \to V$ where $V = \mathcal{R}(\mathbf{U}) = \mathcal{R}(\mathbf{D})$).The projection matrix $\mathrm{P} : \mathcal{R}(\mathbf{U}) \to \mathcal{R}(\mathbf{D})$(U?) can be find by considering the projection of an vector $\mathbf{b} \in \mathbb{R}^{\widetilde{M}}$ onto $\mathcal{R}(\mathbf{U})$, that is solving the least squares problem $\|\mathbf{Ax} - \mathbf{b}\|^2$ where $\mathbf{Ax} \subset \mathcal{R}(\mathbf{U})$ and P = $\mathbf{Ax}$(passer P=Ax her?). When $\mathbf{A}$ has full rank the solution is given by the normal equation

fact: any column in D must be a linear combination of columns in U and visa versa.
Any basis has the same dim(dimension theorem).

$$\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$$
$$\mathbf{x} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}$$

Ax ∈ R(U) ?

resulting in

$$\mathrm{P}\mathbf{b} = \mathbf{Ax} = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}$$
$$\mathrm{P} = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$$

is this something from existing work or your own work? Remember to cite!

Thus $\mathcal{R}(\mathbf{U})$ and $\mathcal{R}(\mathbf{D})$ having the same projection matrix is true if and only if $\mathbf{D}(\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T = \mathbf{U}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T$. Now, remember from the relation between $\mathbf{A}$ and $\mathbf{D}$ that $\mathbf{d}_i = \text{vec}(\mathbf{a}_i\mathbf{a}_i^T)$. From this it is possible to obtain $\mathbf{A}$ through the following optimisation problem

$$\min_{\{\mathbf{a}_i\}_{i=1}^N} \|\mathbf{D}(\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T - \mathbf{U}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\|_F^2$$
$$\text{s.t. } \mathbf{d}_i = \text{vec}(\mathbf{a}_i\mathbf{a}_i^T) \tag{4.5}$$

where $\mathbf{U}$ results from PCA performed on $\text{vec}(\widehat{\mathbf{\Sigma}}_{\mathbf{Y}_s})$. In the following section the optimization problem is analysed and processed in order to determine a suitable solution method.

4.2.3 Solution to optimization problem

The optimization problem (4.5) consist of an objective function forming a least-square problem with respect to the frobenius norm. That is a convex quadratic objective function. The constraints is a set of quadratic equality constraints. In general it is a thumb rule that non-linear equality constraint are not convex. Due to the constraints not being considered convex the optimization problem does not meet the requirements of a convex optimization problem. Hence the numerical solution methods for convex optimization problems, for which convergence is ensured, does not imply directly.

How do we see this?

Due to the nature of the constraints it should be possible to reformulate the objective function to include the constraints into the objective function. That is

Where did you get this from?

construction an unconstrained least-squares problem, which is a special subclass of convex optimization[7].

Let $\mathbf{D} = f(\mathbf{a}_0, \cdots, \mathbf{a}_N)$ then an optimization problem without constraints is achieved and it can be solved be use of for instance the gradient decent method.

description of the exact solution method is missing

4.3 Pseudo Code of the Cov-DL Algorithm

Algorithm 1 Cov-DL

procedure COV-DL($\mathbf{Y}_s$)
  for $s \leftarrow 1, \cdots,$ n_seg **do**
    compute sample covariance matrix $\widehat{\mathbf{\Sigma}}_{\mathbf{Y}_s}$
    $\mathbf{y}_{\text{cov}_s} = \text{vec}(\widehat{\mathbf{\Sigma}}_{\mathbf{Y}_s})$
  end for
  $\mathbf{Y}_{\text{cov}} = \{\mathbf{y}_{\text{cov}_s}\}_{s=1}^{\text{n\_seg}}$
  if $N \geq \widetilde{M}$ **then**
    procedure K-SVD($\mathbf{Y}_{\text{cov}}$)
      returns $\mathbf{D} \in \mathbb{R}^{\widetilde{M}} \times N$
    end procedure
    for $j \leftarrow 1, \cdots, N$ **do**
      $\mathbf{T} = \text{vec}^{-1}(d_j)$
      $\lambda_j \leftarrow \max\{\text{eigenvalue}(\mathbf{T})\}$
      $\mathbf{b}_j \leftarrow$ eigenvector(λ_j)
      $\mathbf{a}_j \leftarrow \sqrt{\lambda_j}\mathbf{b}_j$
    end for
    $\mathbf{A} = \{\mathbf{a}_j\}_{j=1}^{N}$
  end if

  if $N < \widetilde{M}$ **then**
    procedure PCA(vec($\mathbf{\Sigma}_{\mathbf{Y}_s}$))
      returns $\mathbf{U} \in \mathbb{R}^{\widetilde{M} \times N}$
    end procedure
    procedure MIN. $\mathbf{A}$ IN ($\|\mathbf{D}(\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T - \mathbf{U}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\|_F^2$)
      returns $\mathbf{A} = \{\mathbf{a}_j\}_{j=1}^{N}$
    end procedure
  end if
end procedure

4.4 Considerations and Remarks

Through this chapter different theory aspects haven been investigated to create a foundation to present one method to be use in the localisation of the sources from EEG measurements – the recovering of the mixing matrix $\mathbf{A}$ – yet one method is still to be presented. Before the method to recover the source matrix $\mathbf{X}$ from the found mixing matrix $\mathbf{A}$ and EEG measurements $\mathbf{Y}$ will be introduced some considerations and remarks regarding the Cov-DL algorithm must be taken – this will be used in the implementation of the algorithm which will be described in chapter **??**.

The length of each segment determined whenever the covariance of the source matrix $\mathbf{X}$ can be described as a diagonal matrix $\mathbf{\Lambda}$. That is a segment of L_s samples becomes stationary and therefore the sources within that segment becomes uncorrelated – the covariance of the source can be described by a diagonal matrix. The number of samples L_s used in one segment affect whenever the segment is stationary or not. This must be taken into account in the preprocessing part of the baseline algorithm when the EEG measurements are divided into segments.

For the Cov-DL algorithm when $\mathbf{D}$ is under-determined a dictionary learning algorithm K-SVD is used to learn the matrix $\mathbf{D}$ and by that an estimate for the mixing matrix $\hat{\mathbf{A}}$. Because of the segmentation the number of samples used in the dictionary learning are reduced remarkably and will affect the learning process. This is another point which must be taken into account in the preprocessing part of the code. To improved the dictionary learning the overlapping of the segments can be look into as each segment will have some similarity and therefore learn towards one direction.

For the Cov-DL algorithm when $\mathbf{D}$ is over-determined the solution tends to be unique when $M < N < \widetilde{M}$ from testing the solution. That is the cost function tends toward a local minima and therefore an unique solution occur in first run of one trial. For the baseline algorithm it would therefore be necessary to include several random initial points when finding the mixing matrix $\mathbf{A}$ for $\mathbf{D}$ being over-determined.

For a general perspective the sources within the source matrix $\mathbf{X}$ must not be constant over time when using the MMV model (3.2) ...

Find lige kilde på dette argument

Chapter 5

Multiple Sparse Bayesian Learning

In this chapter the multiple sparse Bayesian learning (M-SBL) method is described in details.As the method leverage a Bayesian framework the general concept of Bayesian inference is shortly introduced prior to the the M-SBL method, with respect to the model of interest (5.1). The chapter is inspired by [23] and the articles [24], [4].

citat: p. 9(pdf) sparse Bayesian learning (SBL) is an empirical Bayesian approaches, which use a parameterized prior to encourage sparsity through a process called evidence maximization

Consider again the multiple measurement vector (MMV) model for a non-segmented case of EEG measurements

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{E}, \tag{5.1}$$

with measurement matrix $\mathbf{Y} \in \mathbb{R}^{M \times L}$, source matrix $\mathbf{X} \in \mathbb{R}^{N \times L}$ mixing matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$ and noise matrix $\mathbf{E} \in \mathbb{R}^{M \times L}$. Note that $\mathbf{A}$ is known throughout the chapter, as it is found by Cov-DL in chapter 4.

The aim is to recover the source matrix $\mathbf{X}$ in the case of fewer measurements than active sources, $k > M$. In [4] it is proven that exact localization of the active sources can be achieved with M-SBL for $k > M$, when two sufficient conditions are satisfied. The basic approach of M-SBL is to find the support set S providing the non-zeros rows of the source matrix $\mathbf{X}$ which corresponds to localization of the active sources. Finally the value of the localized active sources are estimated.

5.1 Bayesian Inference

General Bayesian statistics builds upon the task of inferring what the model parameters must be, given the model and data. This is centred around Bayes' theorem, which is a posterior distribution of some unobserved variable given some observed variable.

Consider now the current non-segmented MMV model (5.1) within the Bayesian framework, the model parameter – the source matrix $\mathbf{X}$ – is wished estimated given

the measurement matrix $\mathbf{Y}$. Bayes' theorem becomes

$$p(\mathbf{X}|\mathbf{Y}) = \frac{p(\mathbf{Y}|\mathbf{X})p(\mathbf{X})}{p(\mathbf{Y})},$$

where $p(\mathbf{Y}|\mathbf{X})$ is the probability density function of $\mathbf{Y}$ given $\mathbf{X}$, also referred to as a likelihood function, $p(\mathbf{X})$ is a prior distribution of $\mathbf{X}$ and $p(\mathbf{Y})$ is the distribution of $\mathbf{Y}$ serving as a normalizing parameter. By maximizing the posterior distribution $p(\mathbf{X}|\mathbf{Y})$ with respect to $\mathbf{X}$, the maximum a posteriori (MAP) estimate, an estimate for the source matrix can be found as

$$\hat{\mathbf{X}}_{\text{MAP}} = \arg\max_{\mathbf{X}} \frac{p(\mathbf{Y}|\mathbf{X})p(\mathbf{X})}{p(\mathbf{Y})}.$$

hvordan er dette ikke korrekt? 'rasmus'

That is the estimate of $\mathbf{X}$, from the MMV model (5.1), with the highest probability of causing the given variable $\mathbf{Y}$. In the desired case where $M < N$ the MMV model (5.1) makes an under-determined system and potentially an infinitely number of solutions exist with equal likelihoods.

Let the source matrix $\mathbf{X}$ be seen as a variable which is drawn from some distribution $p(\mathbf{X})$, as such it is possible to narrow down the infinitely solution space. Assuming a prior belief that $\mathbf{Y}$ is generated from a sparse source matrix, gives a so-called sparsity inducing prior. That is $\mathbf{X}$ is drawn from some distribution which has a sharp, possibly infinite, spike at zero surrounded by fat tails. A specific example could be a prior distribution $p(\mathbf{X}) \propto \exp(-\|\mathbf{X}\|_0)$ [23, p. 14]. However, for simplicity a Gaussian prior is to prefer. The use of a Gaussian distribution can (almost) be justified if a mixture of two Gaussian distributions are considered such that the variable is drawn from one of the two with equal likelihood. One where the variance of the distribution is close to zero, resembling the narrow spike around the mean at zero. And, one with high variance resembling the fat tails.

why do you need a fat tail?

giver det mening med den mixture?

Different MAP estimation approaches exists separated by the choice of sparsity inducing prior and optimization method. However, regardless of the approach some problems have shown to occur when using a fixed and algorithm-dependent prior. One issue is the posterior not being sparse enough if a prior is not as sparse, leading to non-recovery. Another issue is that a combinatorial number of suboptimal local solutions can occur. By use of automatic relevance determination (ARD) the problems related to the fixed sparse prior can be avoided [23, p. 20]. The main asset of this alternative approach is the use of an empirical prior. That is an flexible prior distribution which depends on an unknown set of hyperparameters, which is to be learned from the data.

5.1.1 Empirical Bayesian Estimation

Assume the likelihood function $p(\mathbf{Y}|\mathbf{X})$ is Gaussian, with known noise variance σ^2. Due to $\mathbf{Y}$ containing samples from multiple sensors over time each entry of $\mathbf{Y}$ are

known covariance? Matrix

why is this so?

why independent?

independent and equally distributed with likelihood

$$\begin{aligned} p(y_{ij}|x_{ij}) &\sim \mathcal{N}(\mathbf{A}_{i\cdot}\mathbf{x}_{\cdot j}, \sigma^2) \\ &= \frac{1}{\sigma^2\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y_{ij} - \mathbf{A}_{i\cdot}\mathbf{x}_{\cdot j}}{\sigma}\right)^2\right) \end{aligned}$$

Tidligere definition a likelihood

$$\begin{aligned} p(\mathbf{y}_{\cdot j}|\mathbf{x}_{\cdot j}) &= \mathcal{N}(\mathbf{A}\mathbf{x}_{\cdot j}, \sigma^2\mathbf{I}) \\ &= (2\pi)^{-\frac{M}{2}} |\sigma^2\mathbf{I}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2\mathbf{I}} \|\mathbf{y}_{\cdot j} - \mathbf{A}\mathbf{x}_{\cdot j}\|_2^2\right) \end{aligned}$$

Kan det også skrives med eksponent $-M/2$ Gm ?

Now the empirical prior is defined by application of ARD. Similar to the entries of $\mathbf{Y}$ each parameter x_{ij} are independent and equally distributed by a Gaussian distribution with zero mean and a variance controlled by an unknown hyperparameter γ_i:

$$p(x_{ij}; \gamma_i) \sim \mathcal{N}(0, \gamma_i).$$

Note that every entry of row i is controlled by the same hyperparameter γ_i, that is one source signal over time is controlled by one hyperparameter. By combining the prior of each parameter, the prior of $\mathbf{X}$ is fully specified as follows

$$p(\mathbf{X}; \boldsymbol{\gamma}) = \prod_{i=1}^{N} p(\mathbf{x}_{i\cdot}; \gamma_i),$$

with the hyperparameter vector $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_N]^T$. Note that one column of the unknown sources $\mathbf{x}_{\cdot j}$ depends on the $\boldsymbol{\gamma}$. The prior can be composed as

$$p(\mathbf{x}_{\cdot j}; \boldsymbol{\gamma}) = \prod_{i=1}^{N} p(x_{ij}; \gamma_i).$$

giver det mening? hvis 5.2 skulle skrive om så skal det hele være produkter ikke?

Combining the prior and the likelihood $p(\mathbf{y}_{\cdot j}|\mathbf{x}_{\cdot j})$ the posterior of the j-th column of the source matrix $\mathbf{X}$ becomes

$$\begin{aligned} p(\mathbf{x}_{\cdot j}|\mathbf{y}_{\cdot j}; \boldsymbol{\gamma}) &= \frac{p(\mathbf{y}_{\cdot j}|\mathbf{x}_{\cdot j}; \boldsymbol{\gamma}) p(\mathbf{x}_{\cdot j}; \boldsymbol{\gamma})}{p(\mathbf{y}_{\cdot j}|\boldsymbol{\gamma})} \\ &= \frac{p(\mathbf{y}_{\cdot j}|\mathbf{x}_{\cdot j}; \boldsymbol{\gamma}) p(\mathbf{x}_{\cdot j}; \boldsymbol{\gamma})}{\int p(\mathbf{y}_{\cdot j}|\mathbf{x}_{\cdot j}) p(\mathbf{x}_{\cdot j}; \boldsymbol{\gamma})\, d\mathbf{x}_{\cdot j}} \\ &\propto p(\mathbf{y}_{\cdot j}|\mathbf{x}_{\cdot j}; \boldsymbol{\gamma}) p(\mathbf{x}_{\cdot j}; \boldsymbol{\gamma}) \qquad (5.2) \\ &\sim \mathcal{N}(\boldsymbol{\mu}_{\cdot j}, \boldsymbol{\Sigma}), \qquad (5.3) \end{aligned}$$

where the denominator is the marginal likelihood of $\mathbf{y}_{\cdot j}$ also referred to as the evidence. The marginalization is elaborated in the next section. The mean and covariance of (5.3) is given as

$$\boldsymbol{\Sigma} = \mathrm{Cov}(\mathbf{x}_{\cdot j}|\mathbf{y}_{\cdot j}; \boldsymbol{\gamma}) = \boldsymbol{\Gamma} - \boldsymbol{\Gamma}\mathbf{A}^T\boldsymbol{\Sigma}_y^{-1}\mathbf{A}\boldsymbol{\Gamma}, \quad \forall j = 1, \ldots, L \qquad (5.4)$$

$$\mathcal{M} = [\boldsymbol{\mu}_{\cdot 1}, \ldots, \boldsymbol{\mu}_{\cdot L}] = \mathbb{E}[\mathbf{X}|\mathbf{Y}; \boldsymbol{\gamma}] = \boldsymbol{\Gamma}\mathbf{A}^T\boldsymbol{\Sigma}_y^{-1}\mathbf{Y}, \qquad (5.5)$$

where $\mathbf{\Gamma} = \text{diag}(\boldsymbol{\gamma})$ and $\mathbf{\Sigma}_y = \sigma^2\mathbf{I} + \mathbf{A}\mathbf{\Gamma}\mathbf{A}^T$. The derivation of the posterior mean and covariance if found in appendix 5.2.

Let the posterior mean $\mathcal{M}$ serve as the point estimate for the source matrix $\mathbf{X}$. It is clear that row sparsity is achieved whenever $\gamma_i = 0$. From this the posterior must satisfy the following

$$\mathbb{P}(\mathbf{x}_{i\cdot} = \mathbf{0}|\mathbf{Y}; \gamma_i = 0) = 1.$$

This ensures that the posterior mean $\mathcal{M}$ of the i-th row, $\boldsymbol{\mu}_{i\cdot}$, become zero, whenever $\gamma_i = 0$ as desired.

From this it is evident that for estimating the support set of $\mathbf{X}$ it is sufficient to estimate the hyperparameter $\boldsymbol{\gamma}$, from which the support set S can be extracted. Furthermore, the point estimate of $\mathbf{X}$, providing the source signal estimate, is given by $\mathcal{M}$ [23, p. 147]. This leads to the actual M-SBL algorithm for which the aim is to estimate $\boldsymbol{\gamma}$ and the corresponding $\mathcal{M}$.

5.2 Derivation of posterior mean and covariance(put in appendix)

The purpose is here to derive the mean and covariance of the posterior distribution

$$p\left(\mathbf{x}_{\cdot j}|\mathbf{y}_{\cdot j}; \boldsymbol{\gamma}\right) \sim \mathcal{N}(\boldsymbol{\mu}_{\cdot j}, \mathbf{\Sigma}).$$

from (5.3) in section 5.1.1. Let now $\mathbf{x}_{\cdot j} = \mathbf{x}$ and $\mathbf{y}_{\cdot j} = \mathbf{y}$

We have

$$p(\mathbf{x}; \boldsymbol{\gamma}) \sim \mathcal{N}(0, \boldsymbol{\gamma}\mathbf{I})$$

and

$$p(\mathbf{y}|\mathbf{x}) \sim \mathcal{N}(\mathbf{A}\mathbf{x}, \boldsymbol{\sigma}^2\mathbf{I}).$$

Now define one Gaussian random variable

$$\mathbf{z} = \begin{bmatrix} \mathbf{x} \\ \mathbf{y}|\mathbf{x} \end{bmatrix} \in \mathbb{R}^{N+M},$$

then the mean and covariance of $\mathbf{z}$ can be partitioned into

$$\mathbf{\Sigma}_\mathbf{z} = \begin{bmatrix} \mathbf{\Sigma}_{\mathbf{xx}} & \mathbf{\Sigma}_{\mathbf{xy}} \\ \mathbf{\Sigma}_{\mathbf{yx}} & \mathbf{\Sigma}_{\mathbf{yy}} \end{bmatrix} \in \mathbb{R}^{N+M \times N+M}$$

$$\boldsymbol{\mu}_\mathbf{z} = \begin{bmatrix} \boldsymbol{\mu}_\mathbf{x} \\ \boldsymbol{\mu}_{\mathbf{y}|\mathbf{x}} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{A}\mathbf{x} \end{bmatrix}.$$

Then, from [9], the posterior distribution of $\mathbf{x}$ given $\mathbf{y}$ is defined as

$$\mathbf{\Sigma} = \text{cov}(\mathbf{x}, \mathbf{x}|\mathbf{y}) = \mathbf{\Sigma}_{\mathbf{xx}} - \mathbf{\Sigma}_{\mathbf{xy}}\mathbf{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{\Sigma}_{\mathbf{yx}}$$
$$\boldsymbol{\mu} = \boldsymbol{\mu}_\mathbf{x} + \mathbf{\Sigma}_{\mathbf{xy}}\mathbf{\Sigma}_{\mathbf{yy}}^{-1}(\mathbf{y} - \boldsymbol{\mu}_\mathbf{y})$$

Consider first the conditional covariance. Each covariance within the expression is now found:
The covariance of **x** comes directly from the distribution

$$\boldsymbol{\Sigma}_{\mathbf{xx}} = \gamma\mathbf{I}$$

The covariance between **x** and **y**|**x** is found by .....

how to treat **y** being conditional here? can't get the right result

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathbf{yx}} &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[XY] - 0\mathbb{E}[Y] \\ &= \mathbb{E}[XY] \\ &\dots \\ &= \gamma\mathbf{I}\mathbf{A}^T ?\end{aligned}$$

$Y = Ax + E$? $\mathbb{E}[YX] = \mathbb{E}[(Ax+E)x^T] = A\Sigma_{xx} = \gamma A$

Lastly the covariance of **y**|**x** is similarly found by the definition of conditional covariance as follows

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathbf{yy}} = \operatorname{cov}(\mathbf{y},\mathbf{y}|\mathbf{x}) = \boldsymbol{\Sigma}_{\mathbf{yy}|\mathbf{x}} &= \boldsymbol{\Sigma}_{\mathbf{yy}} - \boldsymbol{\Sigma}_{\mathbf{yx}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\boldsymbol{\Sigma}_{\mathbf{xy}} \\ &= \boldsymbol{\sigma}^2\mathbf{I} - \gamma\mathbf{I}\mathbf{A}^T(\gamma\mathbf{I})^{-1}\mathbf{A}\gamma\mathbf{I} \\ &= \boldsymbol{\sigma}^2\mathbf{I} - \mathbf{A}\gamma\mathbf{I}\mathbf{A}^T\end{aligned}$$

Σ_{yy} $\Sigma_{yy|x}$

The resulting conditional covariance becomes

$$\boldsymbol{\Sigma} = \gamma\mathbf{I} - \gamma\mathbf{I}\mathbf{A}^T(\boldsymbol{\Sigma}_{\mathbf{yy}|\mathbf{x}})^{-1}\mathbf{A}\gamma\mathbf{I}.$$

Consider now the mean of **x** conditional on **y**

$$\begin{aligned}\boldsymbol{\mu} &= \boldsymbol{\mu}_{\mathbf{x}} + \boldsymbol{\Sigma}_{\mathbf{xy}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\left(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}}\right) \\ &= 0 + \gamma\mathbf{I}\mathbf{A}^T\left(\boldsymbol{\Sigma}_{\mathbf{yy}}\right)^{-1}\left(\mathbf{y} - \mathbf{A}\mathbf{x}\right)\end{aligned}$$

in the result we have used, the mean of y is subtracted.?

5.3 M-SBL for estimation of X

The M-SBL algorithm is now specified in order to estimate the hyperparameter γ and then the corresponding unknown sources **X**. Due to the empirical Bayesian strategy the unknown variables, making the source matrix **X** are integrated out, also referred to as marginalized. By integrating the posterior with respect to the unknown sources **X** the marginal likelihood of the observed mixed data **Y** is achieved [23, p. 146]

$$\begin{aligned}\mathcal{L}(\gamma;\mathbf{Y}) &= \int p(\mathbf{Y}|\mathbf{X})p(\mathbf{X};\gamma)\, d\mathbf{X} \\ &= p(\mathbf{Y}|\gamma)\end{aligned}$$

when to use ; inst

The resulting marginal likelihood of $\boldsymbol{\gamma}$ is to be maximised with respect to $\boldsymbol{\gamma}$, that is the maximum likelihood estimate (MLE) which is considered the resulting ARD based M-SBL cost function. The $-2\log(\cdot)$ transformation is applied in order for the cost function to be minimized, and factors not depending on $\mathbf{Y}$ is removed. Resulting in the following log likelihood.

ops på L faktor i sidste linje..

$$\begin{aligned}\ell(\boldsymbol{\gamma};\mathbf{Y}) &= -2\log(p(\mathbf{Y};\boldsymbol{\gamma})) \\ &= -2\log\left(2\pi^{\frac{M}{2}}|\boldsymbol{\Sigma}_y|^{\frac{1}{2}}\exp\left(-\frac{1}{2}\sum_{j=1}^{L}\mathbf{y}_{\cdot j}^T\boldsymbol{\Sigma}_y^{-1}\mathbf{y}_{\cdot j}\right)\right) \\ &= L\log(|\boldsymbol{\Sigma}_y|)+\sum_{j=1}^{L}\mathbf{y}_{\cdot j}^T\boldsymbol{\Sigma}_y^{-1}\mathbf{y}_{\cdot j}. \end{aligned} \tag{5.6}$$

It is not expected that an explicit solution to the minimization problem can be found by differentiating and letting the expression equal to zero, hence it has to be solved iteratively based on a initial parameter guess $\boldsymbol{\gamma}^{(0)}$. One iterative method is the expectation maximisation (EM) algorithm. In general each iteration consist of an expectation (E) step, where a function determines the expectation of the likelihood function given the currently estimated parameters. The E-step is followed by an maximization (M) step which computes the parameters by maximizing the expected likelihood found in the E-step. In this case the E-step is to compute the posterior moments using (5.4) and (5.5) while the M-step is the following update rule of γ_i [23, p.147]

remember to cite

$$\gamma_i^{(k+1)} = \frac{1}{L}\|\boldsymbol{\mu}_{i\cdot}\|_2^2 + \boldsymbol{\Sigma}_{ii}, \quad \forall i = 1,\ldots,N.$$

The M-step is in general very slow on large data. An alternative is to use a fixed point update rule to fasten convergence on large data, however convergence is no longer ensured. The fixed point updating step is achieved by taking the derivative of the marginal log likelihood $\ell(\boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$ and equating it with zero. This lead to the following update rule which can replace the above M-step in the EM-algorithm

$$\gamma_i^{(k+1)} = \frac{\frac{1}{L}\|\boldsymbol{\mu}_{i\cdot}\|_2^2}{1-\gamma_i^{-1(k)}\boldsymbol{\Sigma}_{ii}}, \quad \forall i = 1,\ldots,N.$$

mangler at udlede denne regl

Empirically this alternative update rule have shown use full in highly under-determined large scale cases by driving many hyper parameters toward zero allowing for the corresponding weight in the source matrix to be discarded. For simultaneous sparse approximation problems this is the process referred to as multiple sparse Bayesian learning, M-SBL.

From the resulting $\boldsymbol{\gamma}^*$ the support set S of the source matrix $\mathbf{X}$ can be extracted,

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

concluding the localization of active sources within $\mathbf{X}$. In practise some arbitrary small threshold can be used such that that any sufficiently small hyperparameter is discarded. For identification of the active sources the estimate of the source matrix $\mathbf{X}$ is given as $\mathbf{X}^* = \mathcal{M}^*$, with $\mathcal{M}^* = \mathbb{E}[\mathbf{X}|\mathbf{Y};\boldsymbol{\gamma}^*]$. This leads to the following estimate

$$\mathbf{X}^* = \begin{cases} \mathbf{x}_{i\cdot} = \boldsymbol{\mu}^*_{i\cdot}, & i \in S \\ \mathbf{x}_{i\cdot} = \mathbf{0}, & i \notin S \end{cases}$$

5.3.1 Pseudo Code for the M-SBL Algorithm

Algorithm 2 M-SBL

procedure M-SBL($\mathbf{Y}, \mathbf{A}$, iterations)
  $\boldsymbol{\gamma} = \mathbf{1} \in \mathbb{R}^{\text{iterations}+2\times N\times 1}$
  iter = 0
  while $\boldsymbol{\gamma} \geq 10^{-16}$ **do**
    $\boldsymbol{\Gamma} = \text{diag}(\boldsymbol{\gamma}^{\text{iter}})$
    for $i = 1, \ldots, N$ **do**
      $\boldsymbol{\Sigma} = \boldsymbol{\Gamma} - \boldsymbol{\Gamma}\mathbf{A}^T\boldsymbol{\Sigma}_y^{-1}\mathbf{A}\boldsymbol{\Gamma}$
      $\mathcal{M} = \boldsymbol{\Gamma}\mathbf{A}^T\boldsymbol{\Sigma}_y^{-1}\mathbf{Y}$
      $\gamma_i^{(\text{iter}+1)} = \dfrac{\frac{1}{L}\|\boldsymbol{\mu}_{i\cdot}\|_2^2}{1 - \gamma_i^{-1(\text{iter})}\boldsymbol{\Sigma}_{ii}}$
    end for
    if iter = iterations **then**
      Break
    end if
    iter + = 1
  end while
  Return $\mathcal{M}^*, \boldsymbol{\gamma}^*$
end procedure
procedure SUPPORT($\mathcal{M}^*, \boldsymbol{\gamma}^*$, k)
  Support = $\mathbf{0} \in \mathbb{R}^k$
  $\boldsymbol{\gamma}_{\text{value}} = \boldsymbol{\gamma}^*(-2)$
  for j in range(k) **do**
    if $\boldsymbol{\gamma}_{\text{value}}(\arg\max(\boldsymbol{\gamma}_{\text{value}})) \mathrel{!}= 0$ **then**
      Support(j) = $\arg\max(\boldsymbol{\gamma}_{\text{value}})$
      $\boldsymbol{\gamma}_{\text{value}}(\arg\max(\boldsymbol{\gamma}_{\text{value}})) = 0$
    end if
  end for
  $\mathbf{X} = \mathbf{0} \in \mathbb{R}^{N\times L-2}$
  for i in Support **do**
    $\mathbf{X}(i) = \mathcal{M}^*(-1)(i)$
  end for
  Return $\mathbf{X}$
end procedure

5.3.2 Sufficient Conditions for Exact Source Localization

In [4] it is proven that exact source localization is guaranteed in the under-determined case, $k > M$ when the following conditions on $\mathbf{A}$ is fulfilled. The theorem is based on

a theoretical analysis of the minima where noise-free conditions are considered, that is letting $\sigma^2 \to 0$. Thus the following theorem applies to the noise less case. First, defined a function $f : \mathbb{R}^{M \times N} \to \mathbb{R}^{\frac{M(M+1)}{2} \times N}$, such that for $B = f(\mathbf{A})$ the j-th column is given as $\mathbf{b}_{\cdot j} = \text{vec}(\mathbf{a}_{\cdot j}\mathbf{a}_{\cdot j}^T)$. Here the function vec$(\cdot)$ corresponds to the function defined in section 4.1, being a vectorization of the lower triangular part of a matrix.

upper ?

Theorem 5.3.1
Given a dictionary matrix $\mathbf{A}$ and a set of observed measurement $\mathbf{Y}$, M-SBL recovers the support set of any size k exactly in the noise-free case, if the following conditions are satisfied.

1. The active sources $\mathbf{X}_S$ are orthogonal. That is, $\mathbf{X}_S\mathbf{X}_S^T = \mathbf{\Lambda}$, where $\mathbf{\Lambda}$ is a diagonal matrix and S the support set.
2. Rank$(f(\mathbf{A})) = N$.

The proof can be found in [4, p. 16].

Bibliography

[1] Aharon, M., Elad, M., and Bruckstein, A. "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation". In: *IEEE Transactions on signal processing* Vol. 54, No. 11 (2006).

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

[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] Boyd, S. and Vandenberhe, L. *Convex Optimization*. Cambridge University Press, 2004.

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

[9] Eaton, Morris L. *Multivariate Statistics: a Vector Space Approach*. John Wiley and Sons, 1983.

[10] Elad, M. *Sparse and Redundant Representations*. Springer, 2010.

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

[12] Friston, Karl J. "Functional and Effective Connectivity: A Review". In: *Brain Connectivity* 1 (2011).

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

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

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

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

[17] Pal, Piya and Vaidyanathan, P. P. “Pushing the Limits of Sparse Support Recovery Using Correlation Information”. In: *IEEE Transactions on Signal Processing* VOL. 63, NO. 3, Feb. (2015).

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

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

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

[21] *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.

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

[23] Wipf, D. P. “Bayesian Methods for Finding Sparse Representations”. PhD thesis. University of California, San Diego, 2006.

[24] Wipf, D. P. and Rao, B. D. “An Empirical Bayesian Strategy for Solving the Simultaneous Sparse Approximation Problem”. In: *IEEE Transactions on Signal Processing* Vol. 55.No. 7 (2007).

Appendix A

Extended ICA Algorithms

This appendix provide an extension to the basic algorithm for ICA regarding the measure of non-Gaussianity and the computation method. This extended algorithm is referred to as fast ICA and is more commonly used for source separation. This is the algorithm used to apply ICA on EEG measurements for comparison within the thesis.

A.1 Fixed-Point Algorithm - FastICA

An advantage of gradient algorithms is the possibility of fast adoption in non-stationary environments due the use of all input, $\mathbf{y}$, at once. A disadvantage of the gradient algorithm is the resulting slow convergence, depending on the choice of γ for which a bad choice in practise can disable convergence. A fixed-point iteration algorithm to maximise the non-Gaussianity is an alternative that could be used.
Consider the gradient step derived in section **??**. In the fixed point iteration the sequence of γ is omitted and replaced by a constant. This builds upon the fact that for a stable pointof the gradient algorithm the gradient must point in the direction of $\mathbf{b}_j$, hence be equal to $\mathbf{b}_j$. In this case adding the gradient to $\mathbf{b}_j$ does not change the direction and convergence is achieved.
Letting the gradient given in (**??**) be equal to $\mathbf{w}$ and considering the same simplifications again suggests the new update step as [14, p. 179]

wiki: The fixed point is stable if the absolute value of the derivative of $\mathbf{w}$ at the point is strictly less than 1"

$$\mathbf{b}_j \leftarrow \mathbb{E}[\mathbf{y}(\mathbf{b}_j^T\mathbf{y})^3] - 3\mathbf{b}_j.$$

After the fixed point iteration $\mathbf{b}_j$ is again divided by its norm to withhold the constraint $\|\mathbf{b}_j\| = 1$. Instead of γ the fixed-point algorithm compute $\mathbf{b}_j$ directly from previous $\mathbf{b}_j$.
The fixed-point algorithm is referred to as FastICA. The algorithm has shown to converge fast and reliably, then the current and previous $\mathbf{w}$ laid in the same direction [14, p. 179].

A.1.1 Negentropy

An alternative measure of non-Gaussianity is the negentropy, which is based on the differential entropy. The differential entropy H of a random vector $\mathbf{y}$ with density $p_y(\boldsymbol{\eta})$ is defined as

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

The entropy describes the information that a random variable gives. The more unpredictable and unstructured a random variable is higher is the entropy, e.g. Gaussian random variables have a high entropy, in fact the highest entropy among the random variables of the same variance [14, p. 182].

Negentropy is a normalised version of the differential entropy such that the measure of non-Gaussianity is zero when the random variable is Gaussian and non-negative otherwise. The negentropy J of a random vector $\mathbf{y}$ is defined as

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

with $\mathbf{y}_{\text{gaus}}$ being a Gaussian random variable of the same covariance and correlation as $\mathbf{y}$ [14, p. 182].

As the kurtosis is sensitive for outliers the negentropy is instead difficult to compute computationally as the negentropy require a estimate of the pdf. As such an approximation of the negentropy is needed.

To approximate the negentropy it is common to use the higher order comulants including the kurtosis. The following approximation is stated without further elaboration, the derivation can be found in [14, p. 182].

A.1.2 Fixed-Point Algorithm with Negentropy

Maximization of negentropy by use of the fixed-point algorithm is now presented, for derivation of the fixed point iteration see [14, p. 188]. Algorithm 3 show Fast ICA using negentropy, this is the algorithm which is implemented for comparison with the source separation methods which are tested in this thesis.

Algorithm 3 Fast ICA – with negentropy

procedure PRE-PROCESSING($\mathbf{y}$)
    Center measurements $\mathbf{y} \leftarrow \mathbf{y} - \bar{\mathbf{y}}$
    Whitening $\mathbf{y} \leftarrow \mathbf{y}_{white}$
end procedure

procedure FASTICA($\mathbf{y}$)
    $k = 0$
    Initialise random vector $\mathbf{b}_{j(k)}$ ▷ unit norm
    for $j \leftarrow 1, 2, \cdots, N$ **do**
        while convergance critia not meet **do**
            $k = k + 1$
            $\mathbf{b}_{j(k)} \leftarrow \mathbb{E}[\mathbf{y}(\mathbf{b}_j^T\mathbf{y})] - \mathbb{E}[g'(\mathbf{b}_j^T\mathbf{y})]\mathbf{b}_j$ ▷ g defined in [14, p. 190]
            $\mathbf{b}_{j(k)} \leftarrow \mathbf{b}_j / \|\mathbf{b}_j\|$
        end while
        $x_j = \mathbf{b}_j^T\mathbf{y}$
    end for
end procedure

Appendix B

Supplementary theory for chapter 4

description of content of the chapter.

B.1 Introduction to Compressive Sensing

Compressive sensing is the theory of efficient recovery of a signal from a minimal number of observed measurements. It is build upon empirical observations assuring that many signals can be approximated by remarkably sparser signals. Assume linear acquisition of the observed measurements, then the relation between the measurements and the signal to be recovered can be modelled by the multiple measurement vector (MMV) model (3.2) [11].

Through this section the introduction of the theory behind compressive sensing will be presented for one measurement vector of (3.2), $\mathbf{y}$, such that the theory is based on the linear system (3.1). This will be done for simplicity but the theory will still apply for the extend linear system (3.2).

In compressive sensing terminology, $\mathbf{x} \in \mathbb{R}^N$ is the signal of interest which is sought recovered from the EEG measurement $\mathbf{y} \in \mathbb{R}^M$ by solving the linear system (3.1). In the typical compressive sensing case the system is under-determined, $M < N$, and there will therefore exist infinitely many solutions, provided that one solution exist. However, by enforcing certain sparsity constraints it is possible to recover the wanted signal, hence the term sparse signal recovery [11]. The sparsity constraints are the ones presented in 3.1 where the ℓ_0 is introduced to count the non-zeros of the signal of interest, the source vector $\mathbf{x}$. The number of non-zeros (active sources) k describe how sparse the source vector is.

To find a k-sparse solution to the linear system (3.1) it can be viewed as the

following optimisation problem.

$$\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{C}} \|\mathbf{x}\|_0 \quad \text{subject to} \quad \mathbf{A}\mathbf{x} = \mathbf{y}.$$

Unfortunately, this optimisation problem is non-convex due to the definition of the ℓ_0-norm and is therefore difficult to solve – it is an NP-hard problem. Instead, by replacing the ℓ_0-norm with the ℓ_1-norm, the optimisation problem can be approximated and hence becomes computationally feasible [8, p. 27]

4.1: skal vi indfør z som en approximation til x. og så et nyt omega eller? eller kan vi lade x* være løslingen til både P0 og P1

$$\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{C}} \|\mathbf{x}\|_1 \quad \text{subject to} \quad \mathbf{A}\mathbf{x} = \mathbf{y}. \tag{B.1}$$

With this optimisation problem the best k-sparse solution $\mathbf{x}^*$ can be found. The optimisation problem is referred to as ℓ_1 optimisation problem or Basis Pursuit. The following theorem justifies that the ℓ_1 optimisation problem finds a sparse solution [11, p. 62-63].

Theorem B.1.1
A mixing matrix $\mathbf{A} \in \mathbb{R}^{M\times N}$ is defined with columns $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_N]$. By assuming uniqueness of a solution $\mathbf{x}^*$ to

$$\min_{\mathbf{x}\in\mathbb{R}^N} \|\mathbf{x}\|_1 \quad \text{subject to} \quad \mathbf{A}\mathbf{x} = \mathbf{y},$$

the system $\{\mathbf{a}_j, j \in \text{supp}(\mathbf{x}^*)\}$ is linearly independent, and in particular

$$\|\mathbf{x}^*\|_0 = \text{card}(\text{supp}(\mathbf{x}^*)) \leq M.$$

Proof
Assume that the set $\{\mathbf{a}_l, l \in S\}$ of l columns from matrix $\mathbf{A} \in \mathbb{R}^{M\times N}$ is linearly dependent with the support $S = \text{supp}(\mathbf{x}^*)$. Thus a non-zero vector $\mathbf{v} \in \mathbb{R}^N$ supported on S exists such that $\mathbf{A}\mathbf{v} = \mathbf{0}$ – the system is linear dependent. The unique solution $\mathbf{x}^*$ can then be written as, for any $t \neq 0$,

$$\|\mathbf{x}^*\|_1 < \|\mathbf{x}^* + t\mathbf{v}\|_1 = \sum_{l\in S} |x_l^* + tv_l| = \sum_{l\in S} \text{sgn}(x_l^* + tv_l)(x_l^* + tv_l). \tag{B.2}$$

For a small $|t|$

$$|t| < \min_{l\in S} \frac{|x_l^*|}{\|\mathbf{v}\|_\infty},$$

then the sign function become

$$\text{sgn}(x_l^* + tv_l) = \text{sgn}(x_l^*), \quad \forall l \in S.$$

By including this result in (B.2) and remembering $t \neq 0$:

$$\|\mathbf{x}^*\|_1 < \sum_{l \in S} \operatorname{sgn}(x_l^*)(x_l^* + t v_l) = \sum_{l \in S} \operatorname{sgn}(x_l^*) x_l^* + t \sum_{l \in S} \operatorname{sgn}(x_l^*) v_l = \|\mathbf{x}^*\|_1 + t \sum_{l \in S} \operatorname{sgn}(x_l^*) v_l.$$

From this it can be seen that it is always possible to choose $t \neq 0$ small enough such that

$$t \sum_{l \in S} \operatorname{sgn}(x_l^*) v_l \leq 0,$$

which contradicts that $\mathbf{v}$ make the columns of $\mathbf{A}$ linear dependent. Therefore, the set $\{\mathbf{a}_l, l \in S\}$ must be linearly independent. ■

From the theorem is must be conclude that the choice of the mixing matrix $\mathbf{A}$ has a significant impact on whenever a unique solution $\mathbf{x}^*$ exist for the ℓ_1 optimisation problem (B.1). Therefore, when recovering $\mathbf{A}$, some considerations regarding the recovering process of $\mathbf{A}$ must be taken into account. A method for the recovering of $\mathbf{A}$ could be to use a dictionary. This will be explain in the following section 4.2.1.

An alternative solution method to the ℓ_1 optimisation includes greedy algorithms such as the Orthogonal Matching Pursuit (OMP) [11, P. 65]. The OMP algorithm is an iteration process where an index set S is updated – at each iteration – by adding indices corresponding to the columns of $\mathbf{A}$ which describe the residual best possible, hence greedy. The vector $\mathbf{x}$ is then updated by a vector supported on S which minimise the residual, that is the orthogonal projection of $\mathbf{y}$ onto the $\operatorname{span}\{\mathbf{a}_l \mid l \in S\}$.

B.2 K-SVD Algorithm

The dictionary learning algorithm K-SVD provides an updating rule which is applied to each column of $\mathbf{A}_0 = [\mathbf{a}_0, \ldots, \mathbf{a}_N]$ where $\mathbf{A}_0$ being a random initial dictionary matrix. Updating first $\mathbf{a}_j$ and then the corresponding coefficients in $\mathbf{X}$ which it is multiplied with the i-th row in $\mathbf{X}$ denoted by $\mathbf{x}_{i\cdot}$. Let $\mathbf{a}_{j_0}$ be the column to be updated and let the remaining columns be fixed. By rewriting the objective function in (4.4) using matrix notation it is possible to isolate the contribution from $\mathbf{a}_{j_0}$.

Tjek nedenstående udledning. a og x er ikke lige lange da a_j er M lang mens $x_{i\cdot}$ er L lang

$$\begin{aligned} \|\mathbf{Y} - \mathbf{A}\mathbf{X}\|_F^2 &= \left\| \mathbf{Y} - \sum_{j=1}^{N} \mathbf{a}_j \mathbf{x}_{i\cdot} \right\|_F^2 \\ &= \left\| \left(\mathbf{Y} - \sum_{j \neq j_0}^{N} \mathbf{a}_j \mathbf{x}_{i\cdot} \right) - \mathbf{a}_{j_0} \mathbf{x}_{i_0\cdot} \right\|_F^2, \end{aligned} \tag{B.3}$$

where $i = j$, $i_0 = j_0$ and where F is the Frobenius norm that works on matrices

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{M} \sum_{j=1}^{N} |a_{ij}|^2}.$$

In (B.3) the term in the parenthesis is denoted by $\mathbf{E}_{j_0}$, an error matrix, and hence by minimising (B.3) with respect to $\mathbf{a}_{j_0}$ and $\mathbf{x}_{i_0\cdot}$ leads to the optimal contribution from j_0

$$\min_{\mathbf{a}_{j_0},\mathbf{x}_{i_0\cdot}} \| \mathbf{E}_{j_0} - \mathbf{a}_{j_0}\mathbf{x}_{i_0\cdot}\|_F^2 . \tag{B.4}$$

The optimal solution to (B.4) is known to be the rank-1 approximation of $\mathbf{E}_{j_0}$. This comes from the Eckart–Young–Mirsky theorem [**?**] saying that a partial single value decomposition (SVD) makes the best low-rank approximation of a matrix such as $\mathbf{E}_{j_0}$. The SVD is given as

$$\mathbf{E}_{j_0} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T \in \mathbb{R}^{M\times N},$$

with $\mathbf{U} \in \mathbb{R}^{M\times M}$ and $\mathbf{V} \in \mathbb{R}^{N\times N}$ being unitary matrices[1] and $\boldsymbol{\Sigma} = \text{diag}[\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_M] \in \mathbb{R}^{M\times N}$ a diagonal matrix. $\boldsymbol{\sigma}_j$ are the non-negative singular values of $\mathbf{E}_{j_0}$. The best k-rank approximation to $\mathbf{E}_{j_0}$, with $k < \text{rank}(\mathbf{E}_{j_0})$ is then given by :

kilde

$$\mathbf{E}_{j_0}^{(k)} = \sum_{j=1}^{k} \boldsymbol{\sigma}_j \mathbf{u}_j \mathbf{v}_j^T .$$

Since the outer product always have rank-1 letting $\mathbf{a}_{j_0} = \mathbf{u}_1$ and $\mathbf{x}_{i_0\cdot} = \boldsymbol{\sigma}_j \mathbf{v}_1^T$ solves the optimisation problem (B.4). However in order to preserve the sparsity in $\mathbf{X}$ while optimising, only the non-zero entries in $\mathbf{x}_{i_0\cdot}$ are allowed to vary. For this purpose only a subset of columns in $\mathbf{E}_{j_0}$ is considered, those which correspond to the non-zero entries of $\mathbf{x}_{i_0\cdot}$. A matrix $\mathbf{P}_{i_0}$ is defined to restrict $\mathbf{x}_{i_0\cdot}$ to only contain the non-zero-rows corresponding to N_{j_0} non-zero rows:

$$\mathbf{x}_{i_0\cdot}^{(R)} = \mathbf{x}_{i_0\cdot}\mathbf{P}_{i_0}$$

where R denoted the restriction. By applying the SVD to the error matrix which has been restricted $\mathbf{E}_{j_0}^{(R)} = \mathbf{E}_{j_0}\mathbf{P}_{i_0}$ and updating $\mathbf{a}_{j_0}$ and $\mathbf{x}_{i_0\cdot}^{(R)}$ the rank-1 approximation is found and the original representation vector is updated as $\mathbf{x}_{i_0\cdot} = \mathbf{x}_{i_0\cdot}^{(R)}\mathbf{P}_{i_0}^T$.

The main steps of K-SVD is described in algorithm 4.

[1] Unitary matrix: $\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{I}$

Algorithm 4 K-SVD

$k = 0$
Initialize random $\mathbf{A}_{(0)}$
Initialize $\mathbf{X}_{(0)} = \mathbf{0}$

procedure K-SVD($\mathbf{A}_{(0)}$)
    Normalize columns of $\mathbf{A}_{(0)}$
    while error $\geq$ limit **do**
        $j = j + 1$
        for $j \leftarrow 1, 2, \ldots, L$ **do** ▷ updating each col. in $\mathbf{X}_{(k)}$
            $\hat{\mathbf{x}}_j = \min_{\mathbf{x}} \|\mathbf{y}_j - \mathbf{A}_{(k-1)}\mathbf{x}_j\|$ subject to $\|\mathbf{x}_j\| \leq k$ ▷ use Basis Pursuit
        end for
        $\mathbf{X}_{(k)} = \{\hat{\mathbf{x}}_j\}_{j=1}^{L}$
        for $j_0 \leftarrow 1, 2, \cdots, N$ **do**
            $\Omega_{j_0} = \{j \mid 1 \leq j \leq L, \mathbf{X}_{(k)}[j_0, j] \neq 0\}$
            From Ω_{j_0} define $\mathbf{P}_{i_0}$
            $\mathbf{E}_{j_0} = \mathbf{Y} - \sum_{j \neq j_0}^{N} \mathbf{a}_j \mathbf{x}_{i\cdot}$
            $\mathbf{E}_{j_0}^{(R)} = \mathbf{E}_{j_0}\mathbf{P}_{i_0}$
            $\mathbf{E}_{j_0}^{(R)} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$ ▷ perform SVD
            $\mathbf{a}_{j_0} \leftarrow \mathbf{u}_1$ ▷ update the j_0 col. in $\mathbf{A}_{(k)}$
            $(\mathbf{x}_{i_0\cdot})^{(R)} \leftarrow \boldsymbol{\sigma}_1 \mathbf{v}_1$
            $\mathbf{x}_{i_0\cdot} \leftarrow (\mathbf{x}_{i_0\cdot})^{(R)} \mathbf{P}_{i_0}^T$ ▷ update the i_0 row in $\mathbf{X}_{(k)}$
        end for
        error = $\|\mathbf{Y} - \mathbf{A}_{(k)}\mathbf{X}_{(k)}\|_F^2$
    end while
end procedure

B.3 Principal Component Analysis

B.4 General Optimation Theory...or more specific what?

Appendix C

List of Scripts

.. remember to setup