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

Abstract:

Here is the abstract

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:
March 4, 2020

Preface

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

Aalborg University, March 4, 2020

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

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

Contents

Preface **v**

Introduction **3**

1 Motivation **5**
1.1 Introduction to EEG Measurements 5
1.2 Related Work and Our Objective 8

2 Problem Statement **11**

3 System Model **13**
3.1 System of Linear Equations 13
3.2 Multiple Measurement Vector Model of EEG 14
3.3 Solution Methods 15

4 Covariance-Domain Dictionary Learning **17**
4.1 Covariances Domain Representation 18
4.2 Determination of the Dictionary 19
4.3 Considerations and Remarks 22
4.4 Content to the chapter 25
4.5 Theory for appendix 26
4.6 Notes 26

5 Multiple Sparse Bayesian Learning **29**
5.1 Bayesian Interference 29
5.2 Empirical Bayesian Estimation 30

6 Baseline Algorithm **35**
6.1 Parameter Choice 35
6.2 Set-Up of the Baseline Algorithm 36
6.3 Tests 36
6.4 Conclusion 36

Bibliography **37**

A Extended ICA Algorithms **39**
A.1 Fixed-Point Algorithm - FastICA 39

B Supplementary theory for chapter 4 **43**
B.1 Introduction to Compressive Sensing 43
B.2 K-SVD Algorithm 45
B.3 Principal Component Analysis 47
B.4 General Optimation Theory...or more specific what? 47

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[12].

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*[18]. 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 [18, p. 68] [19]. 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[18]. 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 [21]. 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 [18, 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[18, 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.

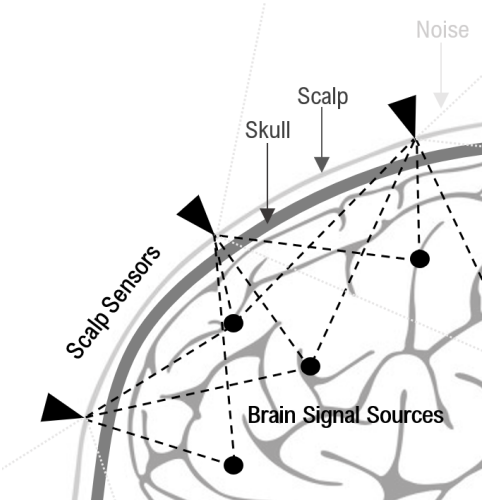

Figure 1.1: Illustration of volume conduction

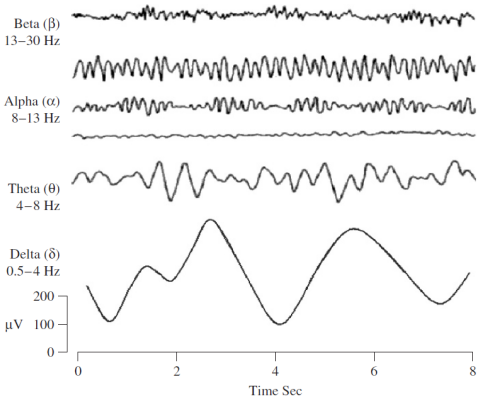

Figure 1.2: Example of time dependent EEG measurements within the four defined categories, source: [18]

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[21, 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 [20]. 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] [7]. 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 [11]. 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 [12]. For this purpose separation and localization of the original sources which contribute to the EEG measurement is of interest. An article from 2016 [19] 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. A known approach is to model the observed data by a linear system

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

$\mathbf{y} \in \mathbb{R}^M$ 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 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 [6]. Representing one time sample the linear system is in general referred to as a single measurement vector model. Only the measurement $\mathbf{y}$ 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 if 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.

Independent Component Analysis (ICA) is one commonly applied method to solve the inverse problem of EEG [15], [14]. ICA is a technique to find the matrix A such that the column wise elements of 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[14, 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 [4]. Meaning that the EEG inverse problem can not form 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 [6]. 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 [4]. 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 [17]. However, the 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 $\mathbf{Y}$[4] and a different method is proposed for finding $\mathbf{X}$ given $\mathbf{Y}$ and $\mathbf{A}$[5].

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, dictating a framework for solving an under-determined system when $\mathbf{X}$ is sufficiently sparse. The point is to transfer the EEG measurements into the covariance domain, for which higher dimensionality can be achieved compared to the original EEG sensor domain of size M. The transformation can be done when assuming the volume conduction is linear and that sources are uncorrelated. As a result the he theory of compressive sensing is found to apply, allowing to learn $\mathbf{A}$ through dictionary learning and an inverse transformation. 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 [4]. As mentioned, the Cov-DL algorithm only learns the mixing matrix $\mathbf{A}$, resembling the volume conduction, and not the explicit source activity $\mathbf{X}$.

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. ...(lidt nærmere tilkendegivelse af grundlæggende teori)... 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[5].

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 makes the foundation of this thesis. The aim of this thesis is to investigate and fully understand the two methods in order to implement and test a joint algorithm. Secondary it is of interest to consider the practical application of the results within a hearing aid. As described in section 1.1 it is of interest to reduce the amount of energy it takes to listen to a specific sound source surrounded by noise. For this purpose we want to relate the found number of active sources to the level of concentration that the test person is experiencing. As described this will include an estimation of the

number of active sources through

Chapter 2

Problem Statement

From the motivation and related work described in chapter 1 it is concluded that EEG measurement of the brain activity has great potential to contribute within the hearing aid industry, regarding the development of hearing aids with improved performance in situations as the cocktail party problem. By solving the overcomplete EEG inverse problem, in order to localize the sources of the brain activity, the results could be used to guide and adapt the performance of a hearing aid – e.g. by moving the microphone beam in the direction of interest. This leads to the following problem statement.

How can sources of activation within the brain be recovered from the EEG inverse problem, in the over-complete case of less sensors than sources, and how can is this recovery process be implemented as a real-time application providing feedback to improve the listening experience?

Update problem statement

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

- How can Cov-DL be used to estimate the mixing matrix $\mathbf{A}$ from the overcomplete EEG inverse problem?
- How can M-SBL be used to estimate the source matrix $\mathbf{X}$ from the overcomplete EEG inverse problem?
- How can the above methods be implemented as one application performing source recovery from EEG measurements in real-time.
- How can the real-time feedback of the system be analysed and used to control the microphone beam of a simulated hearing aid.

Chapter 3

System Model

Through this chapter a model representing the EEG measurements will introduced. Along the model different terminologies will be introduced and described for further use in this thesis. At last some solution methods will be described and will be the ones this thesis will investigate and use to estimate the parameters of for EEG measurements model.

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 in general 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. In general there is no solution to an over-detemined system. An/The exception occur when the system contains a sufficient amount of linearly dependent equations. For $M < N$ the system is under-determined. 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 under-determined system her?

Consider now $\mathbf{y} \in \mathbb{R}^M$ as the observed measurements from M EEG sensors at time t. The linear system 3.1 is then considered as a single measurement vector (SMV) model. Remember from chapter 1 that EEG measurements basically is mixture of original brain signals affected by volume conduction. Modelling the EEG measurements by the SMV model embody the following assumptions/interpretations. $\mathbf{x}$ 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 coefficient matrix $\mathbf{A}$ models the volume conduction by mapping the source vector from $\mathbb{R}^N$ to $\mathbb{R}^M$. $\mathbf{A}$ is referred to as the mixing matrix.

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

man kan ikke teoretisk modellere de non-stationære tilfælde med samme model?

Inchapter 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_s be the length of a 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 $L_s = t_s f$.

her er ts ikke konstant med afhængig af index s, ikke?

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:

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

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 our solution method is presented and discussed before the algorithms are established in the next chapters.

3.3 Solution Methods

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. From the problem statement cf. chapter 2 it is given that 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 system. Therefore, the solution must be found in the infinite solution space, provided that one solution exists – simple linear algebra can not be used. By considering numerical methods it is possible to restrict the solution by some constraints and then find the unique solution which is optimal with respect to some cost function.

One approach is the covariance-domain dictionary learning (Cov-DL) method which was proposed in 2015 by O. Balkan [4]. The background of this method take a part of compressive sensing which is method that recovers unknown signals and mixing matrices from under-determined linear systems. With this method a new term is introduce as sparsity. Sparsity can be view as the number of activation within the sources k – the number of non-zeros indices in $\mathbf{X}$. One thing which is necessary for compressive sensing is the number of activations k is limited by the number of sensors, $k \leq M$, to ensure uniquely recovery of $\mathbf{X}$. This is not always the case in the EEG measurements as more activations than sensors, $k \geq M$, occurs when using low-density equipment to measure the EEG measurements. To overcome this problem O. Balkan transform the compressive sensing problem to covariance domain to increase the dimensionality of the problem – instead the number of activations k is now limited by $k \leq \frac{M(M+1)}{2}$ which allows us to have $k \geq M$. A problem comes

with the transformation into the covariance-domain. With the recovering process taking place in the covariance-domain only the mixing matrix $\mathbf{A}$ is recoverable as it can transformed back to the time-domain without losing its characteristics. As the activations are increase in the covariance-domain, the source matrix $\mathbf{X}$ recovered are not the true source matrix as it cannot be transformed back to the time-domain without losing the sparsity and therefore not be a unique recovering.

To the recovering of $\mathbf{X}$ another method must be used and one which ensure that the sparsity holds. O. Balkan [5] 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 [5] 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.

Chapter 4

Covariance-Domain Dictionary Learning

Through this chapter the method "covariance-domain dictionary learning (Cov-DL)" is presented. 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 over-determined case.

Cov-DL is an algorithm proposed by O. Balkan [4], leveraging the increased dimensionality of the covariance domain. The method have shown successful recovering of the mixing matrix $\mathbf{A}$, even in the over-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 [6] and the article [4]. 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

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

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.

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

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 $\mathbf{\Lambda}_s + \boldsymbol{\varepsilon}$ where $\mathbf{\Lambda}_s$ is a diagonal matrix consisting of the diagonal entries of $\widehat{\mathbf{\Sigma}}_{\mathbf{X}_s}$ and $\boldsymbol{\varepsilon}$ is the estimation error[4]. Each segment is then modelled in the covariance domain as

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

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 $\frac{M(M+1)}{2}$ making a row-wise vectorization of its upper triangular part. Furthermore, let $\text{vec}^{-1}(\cdot)$ be the inverse function for de-vectorisation. This results in the following model

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

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

Here $\boldsymbol{\delta}_s \in \mathbb{R}^N$ contains the diagonal entries of the source sample-covariance matrix $\boldsymbol{\Lambda}_s$ and the matrix $\mathbf{D} \in \mathbb{R}^{(M(M+1))/2 \times N}$ 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 $\text{vec}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y}_s})$ is known from the observed data. By this transformation to the covariance domain one segments is now represented as the single measurement model with $M(M+1)/2$ "measurements". It has been shown that this model allow for identification of $k \leq M(M+1)/2$ active sources [16], 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 $k \leq M(M+1)/2$ 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 Dictionary

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.

4.2.1 Under-determined D

In the case of $N > \frac{M(M+1)}{2}$ the matrix $\mathbf{D}$ becomes under-determined. This is similar to the MMV model (3.2) being under-determined when $N > M$. Thus, it is again possible to solve the under-determined system if certain sparsity is withhold. Namely $\boldsymbol{\delta}_s$ being $\frac{M(M+1)}{2}$-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 $\text{vec}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y}_s})$ for all segments s. For this K-SVD algorithm, described in section 4.2.1 is used. Note here that the number of samples that are used to learn the

Trine: rettet frem hertil

dictionary is remarkable reduces as one segment effectively corresponds to one sample in the covariance domain. When $\mathbf{D}$ is learned it is possible to find an estimate of the mixing matrix $\mathbf{A}$ that generated $\mathbf{D}$ through the relation $\mathbf{d}_j = \text{vec}(\mathbf{a}_j\mathbf{a}_j^T)$. Here each column is found by 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 $\mathbf{a}_j^\star = \sqrt{\lambda_j}\mathbf{b}_j$. Here λ_j is the largest eigenvalue of $\text{vec}^{-1}(\mathbf{d}_j)$,

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

redegørelse for resultatet her skal laves

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

Dictionary Learning

As clarified from the proof of theorem B.1.1 the choice of the mixing matrix $\mathbf{A}$ is essential to achieve the best recovery of the sparse source vector $\mathbf{x}$ from the EEG measurement $\mathbf{y}$. For the estimation one could choose a dictionary as choice for the mixing matrix $\mathbf{A}$.

Skriv kort om hvad dictionary er og gør

Pre-constructed dictionaries do exist which in many cases results in simple and fast algorithms for reconstruction of $\mathbf{x}$ [9]. Pre-constructed dictionaries are 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 [9]. Hence the results of using such dictionaries depend on how well they fit the data of interest, which is creating a certain limitation. An alternative 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 different type of dictionary learning algorithms. One is the K-SVD which is to be elaborated in this section.

The K-SVD algorithm was presented in 2006 by Elad et al. and found to outperform pre-constructed dictionaries when computational cost is of secondary interest [1]. For the K-SVD algorithm several measurements vectors $\mathbf{y}$ must used to learn the dictionary $\mathbf{A}$. Therefore, the K-SVD algorithm will be used on the MMV model as described in (3.2). Consider the EEG measurements matrix $\mathbf{Y} \in \mathbb{R}^M$ consisting of the measurement vectors $\{\mathbf{y}_j\}_j^L$ constructed from the linear system

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

By divided $\mathbf{Y}$ into segment of samples summing up to L samples total a training database is created 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 learing can be defined by an optimisation problem similar to the ℓ_1 optimisation problem defined in (B.1) instead of multiple measurements [9]

gør opmærksom på at i forhold til det tildligere defineret P1 problem har vi nu støj derfor er de byttet rund

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

The learning consists of jointly solving the optimization problem on $\mathbf{X}$ and $\mathbf{A}$. 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 $k < M$ [6].

4.2.2 Over-determined D

In the case of $N < \frac{M(M+1)}{2}$ an over-determined system is achieved and it is not possible to find $\mathbf{D}$ by dictionary learning methods. By assuming that $\frac{M(M+1)}{2}$ will be close to N, because $N > M$ is given,then the measurements in the covariance domain $\text{vec}(\widehat{\mathbf{\Sigma}}_{\mathbf{Y}_s})$ will live on or near a subspace of dimension N. This subspace is spanned by the columns of $\mathbf{D}$, and is denoted as $(\mathbf{D})$. 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). By use of PCA a set of basis vectors $\mathbf{U}$ is achieved such that $\mathcal{R}(\mathbf{U}) = \mathcal{R}(\mathbf{D})$. This however do not imply that $\mathbf{D} = \mathbf{U}$. In the case of two sets of basis vectors span the same space, namely $\mathcal{R}(\mathbf{U}) = \mathcal{R}(\mathbf{D})$, the projection operator of the given subset must be unique. Which is true if and only if $\mathbf{D}(\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T = \mathbf{U}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T$. Remember from the above derivation the condition that $\mathbf{d}_i = \text{vec}(\mathbf{a}_i\mathbf{a}_i^T)$. From this it is possible to obtain $\mathbf{A}$ through the optimisation problem

how else can they live on the same space??

evt. teoretisk beskrivelse af PCA i appendix?

kilde foruden phd p. 51?

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

where $\mathbf{U}$ is learned by use of PCA performed on $\text{vec}(\widehat{\mathbf{\Sigma}}_{\mathbf{Y}_s})$. To solve this optimization problem the cost function is minimized by use of quasi-Newton optimization methods. Several specific quasi-Newton methods exist but the basic principal will be presented here. The Newton optimization method is a multidimensional gradient method. The method is based on a quadratic approximation of the optimization problem by use of the Taylor series, which is elaborated in [3, p. 29]. Let $f(\mathbf{x})$ be the cost function and $\boldsymbol{\delta}$ be the change in $\mathbf{X}$. By differentiating the Taylor approximation of $f(\mathbf{x}+\boldsymbol{\delta})$ and setting it equal to zero, the optimal change in $\mathbf{x}$ is found to be $\boldsymbol{\delta} = -\mathbf{H}^{-1}\mathbf{g}$. Where $\mathbf{g}$ is the gradient and $\mathbf{H}$ is the Hessian. The quasi-Newton methods deviate from the

basic Newton method by letting the direction search be based on a positive semi-definite matrix $\mathbf{S}$ which is generated from available data in order to approximate $\mathbf{H}^{-1}$. Details of the method is found in [3, p. 175]

4.2.3 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 $N > \frac{M(M+1)}{2}$ **then**
    procedure K-SVD($\mathbf{Y}_{\text{cov}}$)
      returns $\mathbf{D} \in \mathbb{R}^{(M(M+1))/2 \times N}$
    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 $N < \frac{M(M+1)}{2}$ **then**
    procedure PCA(vec($\mathbf{\Sigma}_{\mathbf{Y}_s}$))
      returns $\mathbf{U} \in \mathbb{R}^{(M(M+1))/2 \times N}$
    end procedure
    procedure QUASI-NEWTON(problem (4.6))
      returns $\mathbf{A} = \{\mathbf{a}_j\}_{j=1}^{N}$
    end procedure
  end if
end procedure

4.3 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 6.

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 $M < N < (M(M+1))/2$ 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

Optimization Worksheet

4.4 Content to the chapter

4.4.1 Over-determined D

In the case of $N < \frac{M(M+1)}{2}$ an over-determined system is achieved and it is not possible to find $\mathbf{D}$ by dictionary learning methods. By assuming that $\frac{M(M+1)}{2}$ will be close to N, because $N > M$ is given,then the measurements in the covariance domain vec($\widehat{\mathbf{\Sigma}}_{\mathbf{Y}_s}$) 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). By use of PCA a set of basis vectors $\mathbf{U}$ is achieved such that $\mathcal{R}(\mathbf{U}) = \mathcal{R}(\mathbf{D})$. This however do not imply that $\mathbf{D} = \mathbf{U}$. In the case of two sets of basis vectors span the same space, namely $\mathcal{R}(\mathbf{U}) = \mathcal{R}(\mathbf{D})$, the projection operator of the given subset must be unique. Which is true if and only if $\mathbf{D}(\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T = \mathbf{U}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T$. Remember from the above derivation the condition that $\mathbf{d}_i = \text{vec}(\mathbf{a}_i\mathbf{a}_i^T)$. From this it is possible to obtain $\mathbf{A}$ through the optimisation problem

how else can they live on the same space??

evt. teoretisk beskrivelse af PCA i appendix?

kilde foruden phd p. 51?

$$\min_{\mathbf{a}_i} \|\mathbf{D}(\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T - \mathbf{U}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\|_F^2$$
$$\text{s.t. } \mathbf{d}_i = \text{vec}(\mathbf{a}_i\mathbf{a}_i^T) \tag{4.6}$$

where $\mathbf{U}$ is learned by use of PCA performed on vec($\widehat{\mathbf{\Sigma}}_{\mathbf{Y}_s}$). In the following section the optimization problem is analysed and processed in order to determine a suitable method to solve the problem. Additional optimization theory to support the analysis is found in appendix **??**

4.4.2 Solution to optimization problem

The optimization problem (4.6) 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... To

make the constraint convex and linear the constraints are rewritten with respect to the assumption that $\mathbf{a}_i\mathbf{a}_i^T = \mathbf{A}_i$. This results in the following constraints

$$\mathbf{d}_i = \text{vec}(\mathbf{A}_i) \tag{4.7}$$

$$\mathbf{A}_i \geq 0 \tag{4.8}$$

$$\text{rank}(\mathbf{A}_i) = 1 \text{ altid rank 1 når ydre produkt} \tag{4.9}$$

by this a set of (hopefully)convex and linear constraints are achieved, both equality and inequality constraints. Now a classic quadratic programming problem is achieved for which effective solution methods exist.

4.5 Theory for appendix

- notes for driving optimization problem
  We have the matrix of measurements $\text{vec}(\widehat{\mathbf{\Sigma}}_{\mathbf{Y}_s}) \in \mathbb{R}^{\frac{m(m+1)}{2}} \forall s$. from the model we know that the matrix $\mathbf{D}$ makes a basis of the measurements vectors, denoted $\mathcal{R}(\mathbf{D})$, that is a basis of dimension N, provided that $N < \frac{M(M+1)}{2}$ which is the case where the system is over-determined(more equation than variables, which has no solution. hence no ordinary dictionary learning). we want to learn D, but first we can estimate $\mathcal{R}(\mathbf{D})$. we know we can do that by PCA of the set of all $\text{vec}(\widehat{\mathbf{\Sigma}}_{\mathbf{Y}_s})$. But why:
  PCA is a method for dimensionality reduction.
  we represent the measurements by its first N principal components. each found component is linear a combination of the measurements, and has the highest possible variance constraint by being ortogonal to the previously found components. the found components makes an uncorrelated orthogonal basis set $\mathbf{U}$. thus $\mathcal{R}(\mathbf{D}) = \mathcal{R}(\mathbf{U})$ must be true. however this does not imply that $\mathbf{D} = \mathbf{U}$. j principal components is the first j eigenvectors(can be shown) of the covariance matrix, from the L samples it is possible to determined the sample-mean(to be subtracted from the samples) and the covariance matrix(from which to determine the eigenvalues).[p. 125, ICA book] note that it is important to perform the PCA on the found segments where the covariance matrix is stationary. on-line/real time PCA is discussed on page 132[ICA book].
- convex quadratic objective function
- convex solution set/ feasible set

4.6 Notes

noter fra møde:

- augmented Lagrangian
- semidefinite relaxzation

noter:

- any matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$ is called **over complete** if one or more columns is a linear combination of the others – this must be the case when $N > M$ which is also an **under-determined** system (så det der står i rapporten er rigtig dog skal vi nok ikke kalde D-underdetermined, men systemet med D.)

noter til optimeringsproblemet:

- we are dealing with an constraint optimization problem
- when constraints are included it means that we not necessarily can search for the solution along the direction of the negative gradient, but must use methods to determine the feasible search directions given the constraints.
- an important concept within constrained optimization is lagrange multipliers.
- furthermore, we have the 1. order necessary condition (KKT) for $x*$ to be a solution to a constraint problem. and the 2. order conditions.[PO]
- we have a set of equality constraints. for a vector to be a regular point(possible solution?)to the vector has to be a solution to the constraint and the jacobian of the constraints. Jacobian matrisen er $p \times n$ hvor $p < n$ skal gælde.[PO]
- our equality constrints are quadratic and non-convex - according to Jan.
- why non-convex?
  a solution set is convex when you can draw a line between any solution and remain inside the solution set.
- why is our quadratic objective convex?
  is the objective has a positive semi definite Hessian matrix then is can be viewed as a convex programming problem.

Chapter 5

Multiple Sparse Bayesian Learning

As described in section 3.3 the covariance-domain dictionary learning (Cov-DL) is not a fitted method to use in the recovering of the source matrix $\mathbf{X}$ as the found source matrix is not the true recovering. Instead a different method, multiple sparse Bayesian Learning (M-SBL), is used for the recovering process of the last element of the multiple measurement model (MMV) 3.2.

This chapter will include a introduction to M-SBL and some theory of its Bayesian background. The chapter is inspired by [22] and the articles [24], [5].

Let first consider the MMV model for a non-segmented case

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

with the mixing matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$ which is in this chapter a known matrix and source matrix $\mathbf{X} \in \mathbb{R}^{N \times L}$ which is wished recovered from the known EEG measurement matrix $\mathbf{Y} \in \mathbb{R}^{M \times L}$ in the case of $M < N$. The approach is to find the support set S of $\mathbf{X}$ providing the non-zeros rows of $\mathbf{X}$ which corresponds to localization of the active sources. The support set S is constructed from a minimisation of a log-likelihood function defined from Bayesian framework.

5.1 Bayesian Interference

From Bayesian interference theory a posterior distribution is constructed to predict a new distribution of a unknown data sample such that a distribution of the unknown samples are return. The optimum sample estimate of a parameter – for the posterior distribution – can be found by e.g. maximum a posterior (MAP) estimation .

With the knowledge of $\mathbf{A}$ and $\mathbf{Y}$ it is possible to find maximum likelihood estimate of $\mathbf{X}$. By maximising the likelihood $p(\mathbf{Y}|\mathbf{X})$ an estimate of $\mathbf{X}$ can be achieved in the case of more sensors than sources, $M > N$. But in the desired

Kilde (WIKI): This has the disadvantage that it does not account for any uncertainty in the value of the parameter, and hence will underestimate the variance of the predictive distribution

case where $M < N$ the estimation becomes complicate as the MMV model becomes under-determined and potentially an infinitely number of solutions exist with equal likelihoods. As the optimisation problem of the MMV model is NP-hard another estimation method must be used.

Within this Bayesian framework the source matrix $\mathbf{X}$ can be seen as a variable which is drawn from some distribution $p(\mathbf{X})$ such that it is possible to narrow down the infinitely solution space. Assuming a prior belief that $\mathbf{Y}$ is generated from a sparse coefficient matrix, that is the distribution from where $\mathbf{X}$ is drawn has a sharp, possibly infinite, spike at zero surrounded by fat tails.

The MAP estimation of the MMV can then be view as

$$\begin{aligned}\hat{\mathbf{X}} &= \arg\max_{\mathbf{X}} p(\mathbf{Y}|\mathbf{X})p(\mathbf{X}) \\ &= \arg\max_{\mathbf{X}} \frac{p(\mathbf{Y}|\mathbf{X})p(\mathbf{X})}{p(\mathbf{Y})} \quad \text{(Bayers Formular)} \\ &= \arg\max_{\mathbf{X}} p(\mathbf{X}|\mathbf{Y}).\end{aligned}$$

with a prior distribution $p(\mathbf{X}) \propto \|\mathbf{X}\|_0$. From this it it possible to view the optimization problem as a MAP estimation challenge.

5.2 Empirical Bayesian Estimation

Different MAP estimation approaches exists separated by the choice of sparsity inducing prior and optimization method. Some problems have shown to occur when using a fixed and algorithm-dependent prior as the posterior is not sparse enough if a prior is not as sparse leading to a 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 sparse prior can be avoided. ARD is a method where a prior is introduced to determine the relevance of a parameter. This prior is modulated by a vector of hyperparameters affecting the prior variance of each row in $\mathbf{X}$. This can also be view as a regularisation of the solution space which is narrowed to consist only of relevant information [23]. An empirical prior can be used with ARD as the empirical prior is flexible and depends on the unknown hyperparameter $\boldsymbol{\gamma}$ and therefore more data-dependent – such prior can be controlled to induce sparsity.

Let the likelihood $p(\mathbf{Y}|\mathbf{X})$ be Gaussian, with known noise variance σ^2. Then for each column in $\mathbf{Y}$ and $\mathbf{X}$ the likelihood is written as

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

With the use of ARD the i-th row of the source matrix $\mathbf{X}$, $\mathbf{x}_{i\cdot}$, is assigned an L-dimensional independent Gaussian prior with zero mean and a variance controlled by γ_i which is unknown:

$$p(\mathbf{x}_{i\cdot};\gamma_i) = \mathcal{N}(0,\gamma_i\mathbf{I}).$$

By combing the row priors

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

a full prior of $\mathbf{X}$ is achieved modulated by the hyperparameter vector $\boldsymbol{\gamma} = [\gamma_1,\ldots,\gamma_M]^T$. By combining the full prior and the likelihood $p(\mathbf{y}_j|\mathbf{x}_j)$ the posterior of the j-th column of the source matrix $\mathbf{X}$ becomes

Er ved at finde ud af hvorfor vi kan sætte dem sammen på denne måde

$$p(\mathbf{x}_j|\mathbf{y}_j;\boldsymbol{\gamma}) = \frac{p(\mathbf{x}_j,\mathbf{y}_j;\boldsymbol{\gamma})}{\int p(\mathbf{x}_j,\mathbf{y}_j;\boldsymbol{\gamma})\ d\mathbf{x}_j} = \mathcal{N}(\boldsymbol{\mu}_j,\boldsymbol{\Sigma}),$$

with mean and covariance given as

$$\begin{aligned}\boldsymbol{\Sigma} &= \mathrm{Cov}(\mathbf{x}_j|\mathbf{y}_j;\boldsymbol{\gamma}) = \boldsymbol{\Gamma} - \boldsymbol{\Gamma}\mathbf{A}^T\boldsymbol{\Sigma}_y^{-1}\mathbf{A}\boldsymbol{\Gamma}, \quad \forall j = 1,\ldots,L \\ \mathcal{M} &= [\boldsymbol{\mu}_{1\cdot},\ldots,\boldsymbol{\mu}_{L\cdot}] = \mathbb{E}[\mathbf{X}|\mathbf{Y};\boldsymbol{\gamma}] = \boldsymbol{\Gamma}\mathbf{A}^T\boldsymbol{\Sigma}_y^{-1}\mathbf{Y}, \end{aligned} \tag{5.1}$$

where $\boldsymbol{\Gamma} = \mathrm{diag}(\boldsymbol{\gamma})$ and $\boldsymbol{\Sigma}_y = \sigma^2\mathbf{I} + \mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^T$. Let the posterior mean $\mathcal{M}$ serve as the point estimate for $\mathbf{X}$ without involving the support set S. It is clear that row sparsity is achieved whenever $\gamma_i = 0$. From this the posterior must satisfy the following

Er det nemt at se at de to sigmaer vi bruger ikke er den samme? Den ene er en kovarianse mens den anden er en variable

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

which ensure that the posterior mean $\mathcal{M}$ of the i-th row, $\boldsymbol{\mu}_{i\cdot}$, will be zero. Now, instead of estimating the sparsity profile of our source matrix $\mathbf{X}$ it is sufficient to estimate the hyperparameters γ_i[22, p. 147]. Each different hyperparameter $\boldsymbol{\gamma}$ correspond to different hypothesis for the prior distribution of the underlying generation of $\mathbf{Y}$. Therefore the determination of $\boldsymbol{\gamma}$ is seen as a model selection for which an empirical Bayesian strategy can be used. By the empirical Bayesian strategy the unknown weights, making the source matrix $\mathbf{X}$, are treated as nuisance parameters and are integrated out. By integrating the likelihood of $\mathbf{Y}$ with respect to the unknown sources $\mathbf{X}$ the marginal likelihood of the observed mixed data $\mathbf{Y}$, $p(\mathbf{Y};\boldsymbol{\gamma})$ is achieved [22, p. 146]. By applying the $-2\log(\cdot)$ transformation the marginal likelihood function is transformed to a cost function

$$\begin{aligned}\ell(\boldsymbol{\gamma}) &= -2\log\left(\int p(\mathbf{Y}|\mathbf{X})p(\mathbf{X};\boldsymbol{\gamma})\ d\mathbf{X}\right) \\ &= -2\log(p(\mathbf{Y};\boldsymbol{\gamma})) \\ &= \log(|\boldsymbol{\Sigma}_y|) + \frac{1}{L}\sum_{j=1}^{L}\mathbf{y}_j^T\boldsymbol{\Sigma}_y^{-1}\mathbf{y}_j\end{aligned}$$

To minimise the marginal log likelihood $\ell(\boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$ the evidence maximisation (EM) algorithm can be used. The E-step of the EM algorithm is to compute the posterior moments using (5.1) while the M-step is the following update rule of γ_i:

Tjek lige om det dence eller expect

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

The M-step is very slow on large data. Instead one could use a fixed point update to fasten the 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 equation which can replace the above M-step in the EM-algorithm:

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

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 localisation 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 $\mathbf{X}^* = \mathcal{M}^* \approx \mathbf{X}$, with $\mathcal{M}^* = \mathbb{E}[\mathbf{X}|\mathbf{Y}; \boldsymbol{\gamma}^*]$. This leads to the following estimate

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

Algorithm 2 M-SBL

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

Chapter 6

Baseline Algorithm

Through the first 6 chapters a motivation behind investigation the identification of sources from EEG measurement with more sources than sensors as been exploit. From the linear multiple measurement vector (MMV) model

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

the theory and methods behind recovering the mixing matrix $\mathbf{A}$ and the source matrix $\mathbf{X}$ have being investigated leading to two algorithms – covariance-domain dictionary learning (Cov-DL) algorithm and multiple sparse Bayesian learning (M-SBL) algorithm – which recover the mixing matrix and source matrix from the EEG measurements with more sources than sensors.

Through this chapter the two algorithm will be implemented into one algorithm – the baseline algorithm – with the purpose to be tested to investigate how good the recovering process is. For the testing different data sets will be used. Some data sets will be synthetic to investigate whenever a solution occur and then the baseline algorithm will be tested on realistic data sets.

First and all, a discussing whenever to use k or N as a parameter for the number of active sources within the brain, as this number is unknown the realistic case, will be starting this chapter.

6.1 Parameter Choice

- **Første del:** Diskussion af hvad vi gerne vil se ud fra vores kode (god fejl, nem rekonstruktion, ...). Derefter introduktion til k = N, med k > M.
  - Vi ved ikke hvor man N der er per hjerne (ændres sig på mange måde)
  - Diskussion af resultater med N og k
  - Figurer/tests af forskellige tilfælde (Jo højere N jo dårligere placering, ikke samme placering)

  - Lav test med forskellige værdier af k og den sande k – ved N = k og k forskellige fra den sande k kan man læse det på figuren samt værdierne.
  - Ud fra resultaterne konkludere vi bruge k fremad rettet.

6.2 Set-Up of the Baseline Algorithm

- Indsæt flow diagram
- Beskriv opbygningen af koden (hvilke funktioner (cov, m-sbl) har vi inddelt koden i og hvad gør de
- Kig på kompleksiteten af kode (evt. optimering)

6.3 Tests

- Introduktion til test afsnittet. Hvad fokusere vi på, hvad anvender vi, osv.
- Beskriv fejl (hvad for et fejlmål bruger vi). Varians? MSE (vi behøvede hvis ikke at normalisere)? LSE?
- Simulering af data
  - Beskriv data vi anvender
  - Beskriv hvordan vi har lavet det (det syntetiske)
- Test af hver funktion (metode)
- Test af baseline algoritme.
- (Test lige for os selv med ICA, evt. gemme resultater til appendix)

6.4 Conclusion

Opsumering/konklusion af kapitel – henvisning/led hen til test af rigtig data

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] Antoniou, A. and Lu, W-S. *Practical Optimization, Algorithms and Engineering Applications*. Springer, 2007.

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

[5] 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).

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

[7] 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).

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

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

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

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

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

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

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

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

[16] 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).

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

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

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

[20] *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.

[21] Teplan, M. "Fundamentals of EEG Measurement". In: *Measurement science review* 2 (2002).

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

[23] Wipf, D. P. and Nagarajan, S. S. "A New View of Automatic Relevance Determination". In: *Advances in Neural Information Processing Systems 20, MIT Press* (2008).

[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.

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

Letting the gradient given in (**??**) be equal to $\mathbf{w}$ and considering the same simplifications again suggests the new update step as [13, p. 179]

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

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 [13, 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 [13, 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 $\mathbf{y}$ [13, 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 [13, 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 [13, 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 [13, 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) [10].

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 [10]. 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 [10, 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 $t \neq 0$:

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

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) [10, 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?