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: 39

Date of Completion:
January 16, 2020

Preface

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

Aalborg University, January 16, 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[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 electrodes providing measurements.

This thesis explores the 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.

(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 its application. 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 contribution provided in this thesis is presented.

1.1 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 brain consist of an enormous amounts of cells, called neurons, which are mutually connected in neural nets. When a neuron is activated, for instance by a physical stimuli, local current flows are produced among the neurons. This is considered as neural interaction across different parts of the human brain[21].

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. It takes a large amount of simultaneous active neurons to generate an electrical signal that is recordable on the scalp. Such signal is considered as one *source* of neural activity. The current has to penetrate the skull, skin and several other thin layers causing a reduced signal to reach the sensor. It is possible 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. Furthermore, interfering signals can occur in the measurements resulting from physical movement of e.g. eyes and jawbone [21]. Lastly, the transmission of the electric signal through the biological

tissue to the sensor has an unknown mixing effect on the signal. This process is called volume conduction [18, p. 68] [19]. This clarifies the mixture of fluctuating electrical signals with noise that forms the EEG measurements. The concept is sought illustrated on figure 1.1. The resulting EEG measurements 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.

tilføj beskrivelse af stationaitet

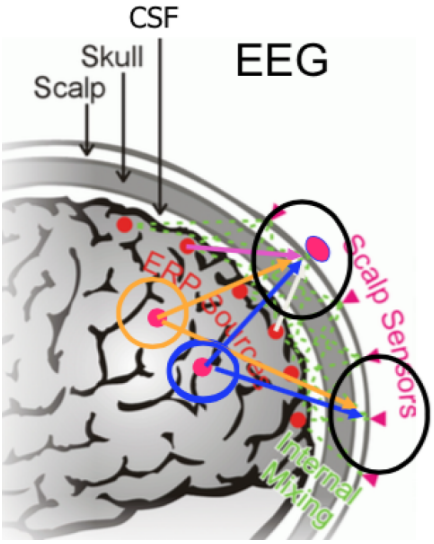

Figure 1.1: Illustration of volume conduction, source: [6](we will make our own figure here instead)

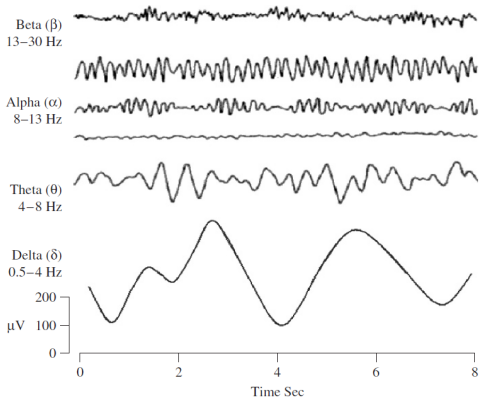

Figure 1.2: Example of time dependent EEG measurements within the four defined categories, source: [18]

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.

1.2 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

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

$\mathbf{Y} \in \mathbb{R}^{M \times L}$ is the EEG measurements where L is the number of samples over time with each sample consisting of M measurements, one for each sensor. $\mathbf{X} \in \mathbb{R}^{N \times L}$ makes L samples of N possible sources within the brain. The non-zero entries of a column in $\mathbf{X}$ represent the active sources at the time of the sample. $\mathbf{A} \in \mathbb{R}^{M \times N}$ is an unknown 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]. Due to the input $\mathbf{Y}$ being a matrix of multiple samples vectors of measurements the linear system is in general referred to as a multiple measurement vector model. Only the measurements $\mathbf{Y}$ are 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 measurements $\mathbf{Y}$. 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 from 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.

1.3 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 iden-

tification 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 underdetermined 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 considered in the referred literature is that the two methods rely on the the number of active sources being given. In practise this is not the case

test lige kode igen hvor forkert k angives

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.

Furthermore 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.

or

Vi får lidt meget ny information her til sidst måske: vi kender ikke antal aktive sources eller source overhoved i praksis, eller gør vi ?

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?

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

New Chapter 3

This chapter gives an introduction to sparse signal recovery. Associated theory regarding compressive sensing is described along the common solution approaches and their limitations.

3.1 Single Measurement Vector Model

Some measurement vector $\mathbf{y} \in \mathbb{R}^M$ can be described as a linear combinations of a coefficient matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$ and some vector $\mathbf{x} \in \mathbb{R}^N$ such that

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

The measurement vector $\mathbf{y}$ consist of M measurements, $\mathbf{x}$ is an unknown vector of N elements, and the coefficient matrix $\mathbf{A}$ models the linear measurement process column-wise. 3.1 makes a system of linear equations with M equations and N unknowns and will be referred to as a linear system for the rest of the master thesis. In other literature the linear system is also denoted as a single measurement vector (SMV) model.

To solve the linear system one must look at the different cases that occurs for the number of equations M and unknowns N. In the case where the number of equations equal the number of unknowns, $M = N$, the coefficient matrix $\mathbf{A}$ becomes a square matrix. A solution for $\mathbf{x}$ can be found to the linear system, provided that a solution exist, if the square coefficient matrix $\mathbf{A}$ has full rank – $\mathbf{A}$ consist of linearly independent columns or rows – as the square matrix can be inverted and the solution is achieved by

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

Such linear system, when $M = N$, is called determined and there will exist a unique solution, provided that a solution exist. In the cases where $M > N$ and $M < N$ the linear system is called over-determined and under-determined, respectively. For

an over-determined system there does not exist a solution and for under-determined systems there exist infinitely many solutions, provided that one solution exist [8, p. ix] .

Her skal vi være opmærksom på at dette ikke altid gælder. At der godt kan være en løsning i over-determined tilfælde (I følge Rasmus) - Laura

From chapter 1 the linear system of interest consists of a observed EEG measurements vector $\mathbf{y} \in \mathbb{R}^M$ with M sensors and a unknown source vector $\mathbf{x} \in \mathbb{R}^N$ with N sources. The matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$ is the mixing matrix. As described in chapter 1, the case of interest is an under-determined linear system, $M < N$, as the EEG measurements have less sensors than sources – hence a solution has to be found within the infinite solution set. For the source vector $\mathbf{x}$ the activation of the sources is denoted by k and there are at most $k \leq N$ activations in the EEG measurements. To find the solution of the linear system (3.1) the number of activations k must also be included in the recovering.

The number of k active source in the source vector $\mathbf{x} \in \mathbb{R}^N$ can be view as non-zeros – the non-activations are described by zeros. To count non-zeros entries of a vector the ℓ_0-norm can be used:

$$\|\mathbf{x}\|_0 := \text{card}(\text{supp}(\mathbf{x})).$$

The function $\text{card}(\cdot)$ gives the cardinality of the input and the support of $\mathbf{x}$, consisting of the non-zeros entries, is given as

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

where $[N]$ is a set of integers $\{1, 2, \cdots, N\}$ [10, p. 41].

3.2 Multiple Measurement Vector Model

As the EEG measurements vary over time multiple measurement vectors are achieved. Therefore, to provided a practical use of the linear system the system is expanded to include the multiple measurement vectors, such that the linear system now becomes a multiple measurement vector (MMV) model:

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

where $\mathbf{Y} \in \mathbb{R}^{M \times L}$ 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. By expanding the system to include multiple measurement vectors some noise will occur. This is added to the linear system as a noise matrix $\mathbf{E} \in \mathbb{R}^{M \times L}$. L denotes the number of observed measurement vectors each consisting of M measurements, that is L samples given. For $L = 1$ the linear system (4.3) will just be the SMV model (3.1) .

Inkludere segmentering her så ledes at vi kan sige at vi har at most k activation i rækkerne af X

Let f be the sample frequency of the observed data $\mathbf{Y}$ and let s denoted a segment index. As such the observed data can be divided into segments $\mathbf{Y}_s \in \mathbb{R}^{M \times t_s f}$,

f er samples pr sek., L er antal sampels i alt, Ls er antal samples pr segment og ts er længen pr segment i sekunder

possibly overlapping, where t_s is the length of the segments in seconds. For each segment the linear model still holds and is rewritten into

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

Cov-DL takes advantage of the covariance domain where the dimensionality is increased allowing for an enlarged number of sources to be active while the dictionary remains recoverable. 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. From the assumption of uncorrelated sources it can be assumed that the sample covariance of $\mathbf{X}_s$ becomes nearly diagonal. This is of importance when the system is transformed to the covariance domain.

The Cov-DL do only recover the mixing matrix $\mathbf{A}$ given the measurements $\mathbf{Y}$. Given $\mathbf{A}$ the source matrix $\mathbf{X}$ is to be recovered by use of the Multiple Sparse Bayesian Learning algorithm, this is described in section 5

The source matrix $\mathbf{X}$ consists of the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_L$ which have been stacked column-wise such that $\mathbf{X}$ consist of at most k non-zero rows – the activations of sources. As for the SMV model (3.1) the MMV model (4.3) is under-determined with $M \ll N$ and $k < M$ [8, p. 42].

The support of $\mathbf{X}$ denotes the index set of non-zero rows of $\mathbf{X}$ and $\mathbf{X}$ is said to be row-sparse. As the columns in $\mathbf{X}$ are k-sparse and as mentioned before, $\mathbf{X}$ has at most k non-zero rows, the non-zero values occur in common location for all columns. By using this joint information it is possible to recover $\mathbf{X}$ from fewer measurements [8, p. 43].

Chapter 4

Covariance-Domain Dictionary Learning

4.1 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 original measurements, then the relation between the measurements and the signal to be recovered can be modelled by the linear system (3.1) [10].
In compressive sensing terminology, $\mathbf{x} \in \mathbb{R}^N$ is the signal of interest which is sought recovered from the measurements $\mathbf{y} \in \mathbb{R}^M$ by solving the linear system (3.1). The coefficient matrix $\mathbf{A}$ is in the context of compressive sensing referred to as the mixing matrix or the dictionary matrix. In the typical compressive sensing case the system is under-determined, $M < N$, and there exist infinitely many solutions, provided that a solution exist. However, by enforcing certain sparsity constraints it is possible to recover the wanted signal, hence the term sparse signal recovery [10].

4.1.1 Optimisation Problem

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]

3.2: 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

$$\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{C}} \|\mathbf{x}\|_1 \quad \text{subject to} \quad \mathbf{A}\mathbf{x} = \mathbf{y}. \tag{4.1}$$

With this optimisation problem we find the best k-sparse solution $\mathbf{x}^*$. This method is referred to as Basis Pursuit.
The following theorem justifies that the ℓ_1 optimisation problem finds a sparse solution [10, p. 62-63].

Theorem 4.1.1
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

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

To prove this theorem one needs to realise that the set $\{\mathbf{a}_j, j \in S\} \leq M$, with $S = \text{supp}(\mathbf{x}^*)$, can not have more than M linearly independence columns. This will be done by a contradiction. So when $M \ll N$ a sparse signal is automatically achieved.

Proof
Assume that the set $\{\mathbf{a}_l, l \in S\}$ 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 $t \neq 0$,

$$\|\mathbf{x}^*\|_1 < \|\mathbf{x}^* + t\mathbf{v}\|_1 = \sum_{l\in S} |x_l^* + tv_l| = \sum_{l\in S} \text{sgn}(x_l^* + tv_l)(x_l^* + tv_l). \tag{4.2}$$

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 (4.2) and remembering $t \neq 0$:

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

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

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

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. ■

The Basis Pursuit algorithm makes the foundation of several algorithms solving alternative versions of (4.1) where noise is incorporated. An alternative solution method includes greedy algorithms such as the Orthogonal Matching Pursuit (OMP) [10, P. 65]. At each iteration of the OMP algorithm an index set S is updated by adding the index corresponding to a column in $\mathbf{A}$ that best describes the residual, hence greedy. That is the part of $\mathbf{y}$ that is not yet explained by $\mathbf{Ax}$ is included. Then $\mathbf{x}$ is updated as the vector, supported by S, which minimize the residual, that is also the orthogonal projection of $\mathbf{y}$ onto the span$\{\mathbf{a}_l \mid l \in S\}$. The algorithm for OMP can be found in the appendix **??**.

As clarified in section **??** the estimation of the dictionary matrix $\mathbf{A}$ is essential to achieve the best recovery of the sparse signal $\mathbf{x}$ from the measurements $\mathbf{y}$. 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. Different dictionary learning algorithms exist. One is the K-SVD which is to be elaborated in this chapter. 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]. Before the K-SVD algorithm can be investigated the linear model (3.1) must be expanded to be considering sets of training data.

4.2 Multiple Measurement Vector Model

The linear model (3.1) is also referred to as a single measurement vector (SMV) model. In order to adapt the model (3.1) to a practical use the model is expanded to include multiple measurement vectors and take noise into account. A multiple measurement vector (MMV) model consists of the observed measurement matrix $\mathbf{Y} \in \mathbb{R}^{M \times L}$, the source matrix $\mathbf{X} \in \mathbb{R}^{N \times L}$, the dictionary matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$ and the

noise vector $\mathbf{E} \in \mathbb{R}^{M \times L}$:

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{E}. \tag{4.3}$$

L denotes the number of observed measurement vectors each consisting of M measurements, that is L samples are given. For $L = 1$ the linear model will just be the SMV model (3.1). The matrix $\mathbf{X}$ consists of k-sparse vectors $\mathbf{x}_1, \ldots, \mathbf{x}_L$ which have been stacked column-wise such that $\mathbf{X}$ consist of at most k non-zero rows. As for the SMV model (3.1) the MMV model (4.3) is under-determined with $M \ll N$ and $k < M$ [8, p. 42].

The support of $\mathbf{X}$ denotes the index set of non-zero rows of $\mathbf{X}$ and $\mathbf{X}$ is said to be row-sparse. As the columns in $\mathbf{X}$ are k-sparse and as mentioned before, $\mathbf{X}$ has at most k non-zero rows, the non-zero values occur in common location for all columns. By using this joint information it is possible to recover $\mathbf{X}$ from fewer measurements [8, p. 43].

4.3 K-SVD

Consider $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_L]$, $\mathbf{y}_j \in \mathbb{R}^M$ as a training database, created by $\mathbf{y}_j = \mathbf{A}\mathbf{x}_j$ for which one want to learn the best suitable dictionary $\mathbf{A}$ and sparse representation $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_L]$, $\mathbf{x}_j \in \mathbb{R}^N$. For a known sparsity constraint k this can be defined by an optimisation problem similar to the general compressive sensing problem of multiple measurements [9]

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

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

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 recovery is only possible if $k < M$ [6].

4.3.1 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 $\mathbf{x}_{i.}$.
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 $x_{i.}$ er L lang

$$\begin{aligned}\|\mathbf{Y}-\mathbf{A}\mathbf{X}\|_F^2 &= \left\|\mathbf{Y}-\sum_{j=1}^{N}\mathbf{a}_j\mathbf{x}_{i\cdot}\right\|_F^2 \\ &= \left\|\left(\mathbf{Y}-\sum_{j\neq j_0}^{N}\mathbf{a}_j\mathbf{x}_{i\cdot}\right)-\mathbf{a}_{j_0}\mathbf{x}_{i_0\cdot}\right\|_F^2, \end{aligned} \tag{4.5}$$

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 (4.5) the term in the parenthesis is denoted by $\mathbf{E}_{j_0}$, an error matrix, and hence by minimising (4.5) with respect to $\mathbf{a}_{j_0}$ and $\mathbf{x}_{i_0\cdot}$ leads to the optimal contribution from j_0

$$\min_{\mathbf{a}_{j_0},\mathbf{x}_{i_0\cdot}} \| \mathbf{E}_{j_0}-\mathbf{a}_{j_0}\mathbf{x}_{i_0\cdot}\|_F^2 . \tag{4.6}$$

The optimal solution to (4.6) 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

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

with $\mathbf{U} \in \mathbb{R}^{M\times M}$ and $\mathbf{V} \in \mathbb{R}^{N\times N}$ being unitary matrices[1] and $\mathbf{\Sigma} = \text{diag}[\boldsymbol{\sigma}_1,\ldots,\boldsymbol{\sigma}_M] \in \mathbb{R}^{M\times N}$ 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 (4.6). 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

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

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 1.

Algorithm 1 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

The dictionary learning algorithm K-SVD is a generalisation of the well known K-means clustering also referred to as vector quantization. In K-means clustering a set of K vectors is learned referred to as mean vectors. Each signal sample is then represented by its nearest mean vector. That corresponds to the case with sparsity constraint $k = 1$ and the representation reduced to a binary scalar $x = \{1, 0\}$. Further instead of computing the mean of K subsets the K-SVD algorithm computes the SVD factorisation of the K different sub-matrices that correspond to the K columns of $\mathbf{A}$. The section is inspired by chapter 3 in [6] and the article [4]. INTRO..

4.4 Introduction

Covariance-domain dictionary learning (Cov-DL) is an algorithm proposed by O. Balkan [4], claiming to successfully identify more active sources k than available measurements M from the multiple measurement vector model

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

where $\mathbf{Y} \in \mathbb{R}^{M \times L}$ is the observed measurement matrix, $\mathbf{X} \in \mathbb{R}^{N \times L}$ the the source matrix and $\mathbf{E} \in \mathbb{R}^{M \times L}$ is the additional noise matrix .
Let f be the sample frequency of the observed data $\mathbf{Y}$ and let s denoted a segment index. As such the observed data can be divided into segments $\mathbf{Y}_s \in \mathbb{R}^{M \times t_s f}$, possibly overlapping, where t_s is the length of the segments in seconds. For each segment the linear model still holds and is rewritten into

f er samples pr sek., L er antal sampels i alt, Ls er antal samples pr segment og ts er længen pr segment i sekunder

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

Cov-DL takes advantage of the covariance domain where the dimensionality is increased allowing for an enlarged number of sources to be active while the dictionary remains recoverable. 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. From the assumption of uncorrelated sources it can be assumed that the sample covariance of $\mathbf{X}_s$ becomes nearly diagonal. This is of importance when the system is transformed to the covariance domain.
The Cov-DL do only recover the mixing matrix $\mathbf{A}$ given the measurements $\mathbf{Y}$. Given $\mathbf{A}$ the source matrix $\mathbf{X}$ is to be recovered by use of the Multiple Sparse Bayesian Learning algorithm, this is described in section 5

4.5 Covariances domain representation

Consider the covariance of a vector $\mathbf{x}_i$

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

Assume that all samples has zero mean and the same distribution within one segment. The observed measurements $\mathbf{Y}_s \in \mathbb{R}^{M \times L}$ can be described in the covariance domain by the sample covariance $\hat{\mathbf{\Sigma}}$ 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

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

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

$$\hat{\mathbf{\Sigma}}_{\mathbf{Y}_s} = \frac{1}{L}\mathbf{Y}_s\mathbf{Y}_s^T.$$

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

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

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

$$\begin{aligned}
\hat{\mathbf{\Sigma}}_{\mathbf{Y}_s} = \frac{1}{L_s}\mathbf{Y}_s\mathbf{Y}_s^T &= \frac{1}{L_s}\left(\mathbf{A}\mathbf{X}_s + \mathbf{E}_s\right)\left(\mathbf{A}\mathbf{X}_s + \mathbf{E}_s\right)^T \\
\mathbf{Y}_s\mathbf{Y}_s^T &= (\mathbf{A}\mathbf{X}_s)(\mathbf{A}\mathbf{X}_s)^T + \mathbf{E}_s\mathbf{E}_s^T + \mathbf{E}_s(\mathbf{A}\mathbf{X}_s)^T + \mathbf{A}\mathbf{X}_s\mathbf{E}_s^T \\
&= \mathbf{A}\mathbf{X}_s\mathbf{X}_s^T\mathbf{A}^T + \mathbf{E}_s\mathbf{E}_s^T + \mathbf{E}_s\mathbf{X}_s^T\mathbf{A}^T + \mathbf{A}\mathbf{X}_s\mathbf{E}_s^T \\
&= \mathbf{A}(\mathbf{\Delta}_s + \boldsymbol{\varepsilon})\mathbf{A}^T + \mathbf{E}_s\mathbf{E}_s^T + \mathbf{E}_s\mathbf{X}_s^T\mathbf{A}^T + \mathbf{A}\mathbf{X}_s\mathbf{E}_s^T \\
&= \mathbf{A}\mathbf{\Delta}_s\mathbf{A}^T + \mathbf{A}\boldsymbol{\varepsilon}\mathbf{A}^T + \mathbf{E}_s\mathbf{E}_s^T + \mathbf{E}_s\mathbf{X}_s^T\mathbf{A}^T + \mathbf{A}\mathbf{X}_s\mathbf{E}_s^T \qquad (4.7) \\
&= \mathbf{A}\mathbf{\Delta}_s\mathbf{A}^T + \widetilde{\mathbf{E}} \qquad (4.8) \\
& \qquad (4.9)
\end{aligned}$$

er yderligere argumentation nødvendig her?

From (4.7) to (4.8) all terms where noise is included are defined as a united noise term $\widetilde{\mathbf{E}}$. . By vector notation (4.8) is rewritten and then vectorised. Because the covariance matrix $\hat{\mathbf{\Sigma}}_{\mathbf{Y}_s}$ is symmetric it is sufficient to vectorize only the lower triangular parts, including the diagonal. For this the function $\mathrm{vec}(\cdot)$ 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 $\mathrm{vec}^{-1}(\cdot)$ be the inverse function for de-vectorisation. This results in the following model

$$\begin{aligned}
\mathbf{\Sigma}_{\mathbf{Y}_s} &= \sum_{i=1}^{N} \mathbf{\Delta}_{s_{ii}}\mathbf{a}_i\mathbf{a}_i^T + \widetilde{\mathbf{E}} \\
\mathrm{vec}(\mathbf{\Sigma}_{\mathbf{Y}_s}) &= \sum_{i=1}^{N} \mathbf{\Lambda}_{s_{ii}}\mathrm{vec}(\mathbf{a}_i\mathbf{a}_i^T) + \mathrm{vec}(\widetilde{\mathbf{E}}) \\
&= \sum_{i=1}^{N} \mathbf{d}_i\mathbf{\Lambda}_{s_{ii}} + \mathrm{vec}(\widetilde{\mathbf{E}}) \\
&= \mathbf{D}\boldsymbol{\delta}_s + \mathrm{vec}(\widetilde{\mathbf{E}}), \quad \forall s. \qquad (4.10)
\end{aligned}$$

Here $\boldsymbol{\delta}_s \in \mathbb{R}^N$ contains the diagonal entries of the source sample-covariance matrix $\mathbf{\Lambda}_s$ and the matrix $\mathbf{D} \in \mathbb{R}^{M(M+1)/2 \times N}$ consists of the columns $\mathbf{d}_i = \mathrm{vec}(\mathbf{a}_i\mathbf{a}_i^T)$. Note

that $\mathbf{D}$ and $\boldsymbol{\delta}_s$ are unknown while $\text{vec}(\boldsymbol{\Sigma}_{\mathbf{Y}_s})$ 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.

4.6 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.

Under-determined D

In the case of $N > \frac{M(M+1)}{2}$ $\mathbf{D}$ becomes under-determined. This is similar to the original system 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 $\text{vec}(\boldsymbol{\Sigma}_{\mathbf{Y}_s})$ for all s. For this K-SVD algorithm, described in section 4.3 is used. Note here that the number of samples that are used to learn the 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 the original 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 $\mathbf{a}_j^* = \sqrt{\lambda_j}\mathbf{b}_j$. Here λ_j is the largest eigenvalue of $\text{vec}^{-1}(\mathbf{d}_j)$,

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

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

redegørelse for resultatet her skal laves

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.

how else can they live on the same space??

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 $\text{vec}(\mathbf{\Sigma}_{\mathbf{Y}_s})$ 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 $\mathbf{D} = \mathbf{U}$.

evt. teoretisk beskrivelse af PCA i appendix?

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

kilde foruden phd p. 51?

$$\min_{\mathbf{a}_i} \|\mathbf{D}(\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T - \mathbf{U}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\|_F^2$$
$$\text{s.t. } \mathbf{d}_i = \text{vec}(\mathbf{a}_i\mathbf{a}_i^T) \tag{4.11}$$

where $\mathbf{U}$ is learned by use of PCA performed on $\text{vec}(\mathbf{\Sigma}_{\mathbf{Y}_s})$.

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]

4.6.1 Pseudo Code of the Cov-DL Algorithm

Algorithm 2 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**
      $\mathbf{T} = \text{vec}^{-1}(d_j)$
      $\lambda_j \leftarrow \max\{eigenvalue(\mathbf{T})\}$
      $\mathbf{b}_j \leftarrow eigenvector(\lambda_j)$
      $\mathbf{a}_j \leftarrow \sqrt{\lambda_j}\mathbf{b}_j$
    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 (4.11))
      returns $\mathbf{A} = \{\mathbf{a}_j\}_{j=1}^{N}$
    end procedure
  end if
end procedure

4.6.2 Remarks to be considered and investigated in he code

- the number of samples for dictionary learning is reduced remarkably when Cov-DL is used, this is mentioned above. increasing overlap will improved this.
- the effect of the length of segments L_s. it will effect the diagonality of $cov(X)$
- the values of the individual sources must not be constant over time, that is the power of the sourced, it will not be at problem for EEG data.

- for Cov-DL2 the solution tends to be unique when $M < N < M(M+1)/2$, that is the found cost function may have a local minima. for this it is recommended to use several random initial point.

Chapter 5

Multiple Sparse Bayesian Learning

....
INTRODUCTION
....

Consider the multiple measurement vector model (MMV)

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

where the unknown matrices $\mathbf{A} \in \mathbb{R}^{M \times N}$ and $\mathbf{X} \in \mathbb{R}^{N \times L}$ are wished recovered from the known measurement matrix $\mathbf{Y} \in \mathbb{R}^{M \times L}$ in the case of $M < N$. In chapter 4 the mixing matrix $\mathbf{A}$ was found. In this chapter, a method to recover the source matrix $\mathbf{X}$ is sought. The approach is to find the support set S of $\mathbf{X}$ providing the non-zeros rows of $\mathbf{X}$ which corresponds to localisation of the active sources. By ... the active sources are identified in order to fully recover $\mathbf{X}$.
The chapter is inspired by [22] and the articles [24], [5].

5.1 Maximum a Posterior 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 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.

Bayesian probability ...

indsæt her hvad tanken bag Bayesian egenligt er

Within this Bayesian framework the source matrix $\mathbf{X}$ is 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.

By applying an exponential function $\exp(-(\cdot))$ transformation onto our optimisation problem a Gaussian likelihood function $p(\mathbf{Y}|\mathbf{X})$ with a λ-dependent variance is achieved:

$$p(\mathbf{Y}|\mathbf{X}) \propto \exp\left(-\frac{1}{\gamma}\|\mathbf{Y}-\mathbf{A}\mathbf{X}\|_2^2\right),$$

with a prior distribution $p(\mathbf{X}) \propto \exp(-\|\mathbf{X}\|_0)$ [22, p. 137]. The optimisation problem is then rewritten by Bayers formula

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

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 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 sources 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

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

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 now the posterior mean 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

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

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 support set?/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

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

To minimise the marginal log likelihood $\mathcal{L}(\boldsymbol{\gamma})$ 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:

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

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 $\mathcal{L}(\boldsymbol{\gamma})$ 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 3 M-SBL

1. Given $\mathbf{Y}$ and a dictionary matrix $\mathbf{A}$.
2. Initialise $\boldsymbol{\gamma}$, e.g $\boldsymbol{\gamma} = \mathbf{1}$.
3. Compute the posterior moments $\boldsymbol{\Sigma}$ and $\mathcal{M}$.
4. Update $\boldsymbol{\gamma}$ using EM or fixed-point.
5. Repeat step 3 and 4 until convergence to a fixed point $\boldsymbol{\gamma}^*$.
6. Update the posterior moments $\boldsymbol{\Sigma}^*$ and $\mathcal{M}^*$ with γ^*.
7. Extract support set S
8. Set source matrix estimate $\hat{\mathbf{X}}^* = \mathcal{M}^*_S$

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 4 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 4 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