Bayesian Dictionary Learning for EEG Source Identification

Trine Nyholm Kragh & Laura Nyrup Mogensen
Mathematical Engineering, MATTEK

Master's Thesis

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

AALBORG UNIVERSITY
STUDENT REPORT

Title:
Bayesian Dictionary Learning for EEG Source Identification

Theme:

Project Period:
Fall Semester 2019
Spring Semester 2020

Project Group:
Mattek9b

Participant(s):
Trine Nyholm Kragh
Laura Nyrup Mogensen

Supervisor(s):
Jan Østergaard
Rasmus Waagepetersen

Copies: 1

Page Numbers: 47

Date of Completion:
November 14, 2019

Abstract:

Here is the abstract

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

AALBORG UNIVERSITET
STUDENTERRAPPORT

Titel:
Bayesian Bibliotek Læring for EEG Kilde Identifikation

Tema:

Projektperiode:
Efterårssemestret 2019
Forårssemestret 2020

Projektgruppe:
Mattek9b

Deltager(e):
Trine Nyholm Kragh
Laura Nyrup Mogensen

Vejleder(e):
Jan Østergaard
Rasmus Waagepetersen

Oplagstal: 1

Sidetal: 47

Afleveringsdato:
14. november 2019

Abstract:

Her er resuméet

Preface

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

Aalborg University, November 14, 2019

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

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

Danish Summary

Dansk resume ?

Contents

Introduction

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

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

Chapter 1

Motivation

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

1.1 EEG Measurements

Electroencephalography (EEG) is a technique used within the medical field. It is an imaging technique measuring electric signals on the scalp, caused by brain activity. The human brain consist of an enormous amounts of cells, called neurons. These neurons are mutually connected in neural nets and when a neuron is activated, for instance by a physical stimuli, local current flows are produced [21]. This is what makes a kind of neural interaction across different parts of the brain(?).

EEG measurements are provided by a varies number of metal electrodes, referred to as sensors, carefully placed on a human scalp. Each sensor read the present electrical signals, which are then displayed on a computer, as a sum of sinusoidal waves relative to time.
It takes a large amount of active neurons to generate an electrical signal that is recordable on the scalp as the current have to penetrate the skull, skin and several other thin layers. Hence it is clear that measurements from a single sensor do not correspond to the activity of a single specific neuron in the brain, but rather a collection of many activities within the range of the one sensor. Nor is the range of a single sensor separated from the other sensors thus the same activity can easily be measured by two or more sensors. Furthermore, interfering signals can occur in the measurments resulting from physical movement of e.g. eyes and jawbone [21]. Lastly the transmission of the electric field through the biological tissue to the sensor has an unknown mixing effect on the signal, this process is called volume conduction[18,

p. 68][19].

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

The waves resulting from EEG measurements have been classified into four groups according to the dominant frequency. The delta wave ($0.5 - 4$ Hz) is observed from infants and sleeping adults, the theta wave ($4 - 8$ Hz) is observed from children and sleeping adults, the alpha wave ($8 - 13$ Hz) is the most extensively studied brain rhythm, which is induced by an adult laying down with closed eyes. Lastly the beta wave ($13 - 30$ Hz) is considered the normal brain rhythm for normal adults, associated with active thinking, active attention or solving concrete problems [18, p. 11]. An example of EEG measurements within the four categories is illustrated by figure 1.2.

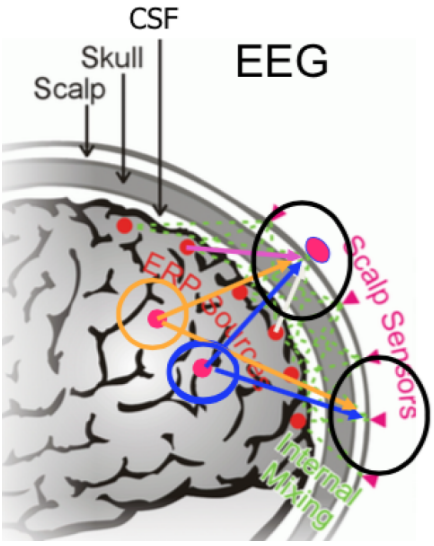

Figure 1.1: Illustration of volume conduction, source [5](we will make our own figure here instead)

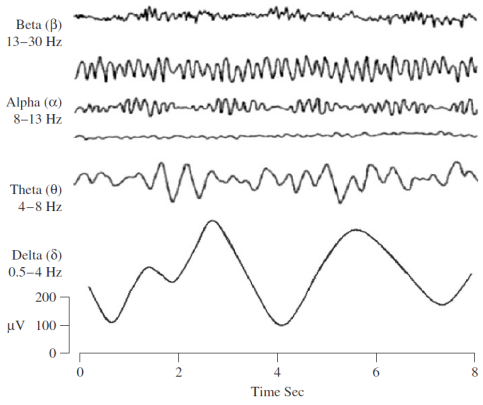

Figure 1.2: Example of time dependent EEG measurements within the four defined categories, source [18]

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

EEG capitalize on the procedure being non-invasive and fast. Neural activity can be measured within fractions of a second after a stimuli has been provided [21, p. 3]. When a person is exposed to a certain stimuli, e.g. visual or audible, the measured activity is said to result from evoked potential.
Over the past two decades, especially functional integration has become an area of interest[11]. Within neurobiology functional integration referrers to the study of the correlation among activities in different regions of the brain. In other words, how do different part of the brain work together to process information and conduct a response[12]. For this purpose separation and localisation of the single sources which contribute to the EEG measurement is of interest. An article from 2016 point out the importance of performing analysis regarding functional integration at source level rather than at EEG level. It is argued through experiments that analysis at EEG level do not allow interpretations about the interaction between sources[19].

The hearing aid industry is one example where this research is highly prioritised. At Eriksholm research center which is a part of the hearing aid manufacture Oticon cognitive hearing science is a research area within fast development[20]. One main purpose of Eriksholm is to make it possible for a hearing aid to identify the attended sound source and hereby exclude noise from elsewhere [2], [6]. This is where EEG and occasionally so called in-ear EEG is interesting, especiallaly in conjunction with the technology of beamforing, which allows for receiving only signals from a specific direction. It is essentially the well known but unsolved cocktail problem which is sought improved by use of EEG. However the focus of this research do consider the correlation between EEG measurements and the sound source rather than localisation of the activated source from the EEG[2]. Hence a source localisation approach could potentially be of interest regarding hearing aids in order to improve the results. (Furthermore, a real-time application to provide feedback from EEG measurements would be essential.)?

1.1.1 Modelling

Considering the issue of localising activated sources from EEG measurements, a known option is to model the observed data by the following linear system

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

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

To solve the EEG inverse problem the concept of compressive sensing makes a solid foundation including sparse signal recovery and dictionary learning. Independent Component Analysis (ICA) is a common applied method to solve the inverse problem [15], [14], here statistical independence between source activity is assumed.
Application of ICA have shown great results regarding source separation of high-density EEG. Furthermore, an enhanced signal-to-noise ratio of the unmixed independent source time series processes allow essential study of the behaviour and relationships between multiple EEG source processes [8].
However a significant flaw to this method is that the EEG measurements are only separated into a number of sources that are equal or less than the number of sensors[3].
This means that the EEG inverse problem can not be over-complete(er det correct i forhold til teorien om ICA?). That is an assumption which undermines the reliability and usability of ICA, as the number of simultaneous active sources easily exceed the number of sensors [5]. This is especially a drawback when low-density EEG are considered, that is EEG equipment with less than 32 sensors. Improved capabilities of low-density EEG devices are desirable due to its relative low cost, mobility and ease to use.

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

1.2 Related Work and Our Contribution

As mentioned above ICA has been a solid method for source localisation in the case where a separation into a number of sources equal to the number of sensors was adequate. To overcome this issue an extension of ICA was suggested, referred to as the ICA mixture model[3]. Instead of identifying one mixing matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$ this approach learns N_{model}(number of sources? or datapoints) different mixing matrices $\mathbf{A}_i \in \mathbb{R}^{M \times M}$. The method was further adapted into the Adaptive Mixture ICA (AMICA) which showed successful results regarding identification of more sources than available sensors [17]. However an assumption of no more than M simultaneously active sources has to be made which is still an essential limitation, especially when considering low-density EEG.
Other types of over-complete ICA algorithms have been proposed to overcome the problem of learning over-complete systems. One is the Restricted ICA (RICA), an efficient method used for unsupervised learning in neural networks [22]. Here the hard orthonormal constraint in ICA is replaced with a soft reconstruction cost.
In 2015 O. Balkan et. al., [3], suggested a new approach also targeting the identification of more sources than sensors regarding EEG. The suggested method, referred

to as Cov-DL, is a covariance based dictionary learning algorithm. The point is to transfer the forward problem(?) into the covariance domain, which has higher dimensionality than the original EEG sensor domain. This can be done when assuming the scalp mixing is linear and using the assumed natural uncorrelation of sources within a certain time-window. The Cov-DL algorithm stands out from the other straight forward dictionary learning methods as it does not relay on the sparsity of active sources, this is an essential advantage when low-density EEG is considered.
Cov-DL was tested on found to outperform both AMICA and RICA[3], thus it is considered the state of the art within the area of source identification.

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

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

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

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

Chapter 2

Problem Statement

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

How can sources of activation within the brain be localised from the EEG inverse problem, in the overcomplete case of less sensors than sources and how can such algorithm be extended to a real-time application providing feedback to improve the intentional listening experience?

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

- How can the over-complete EEG inverse problem be solved by use of compressive sensing included domain transformation?
- How can Cov-DL be used to estimate the mixing matrix $\mathbf{A}$ from the overcomplete EEG inverse problem?
- How can M-SBL be used to estimate the source matrix $\mathbf{X}$ from the overcomplete EEG inverse problem?
- How can an application be formed to constitute this source identification process operating in real-time?
- How can the feedback of the system be used to control the microphone beam of a simulated hearing aid. Especially how to analyse the feedback versus the listening experience in order to improve this.

Chapter 3

Sparse Signal Recovery

This chapter gives an introduction to the concept compressive sensing. Associated theory regarding sparse signal recovery is described along the limitations of the common solution approaches. Finally the state of the art methods regarding non-sparse signal is presented.

3.1 Linear Algebra

Some observed measurement $\mathbf{y}$ can be described as a linear combinations between a coefficient matrix $\mathbf{A}$ and some vector $\mathbf{x}$ such that

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

where $\mathbf{y} \in \mathbb{R}^M$ is the observed measurement vector consisting of M measurements, $\mathbf{x} \in \mathbb{R}^N$ is a source vector of N elements, and $\mathbf{A} \in \mathbb{R}^{M \times N}$ is a coefficient matrix which models the linear measurement process column-wise. The linear model consist of M equations and N unknown.

mention system of linear equations, og evt. vent med model

For the case of $\mathbf{A}$ being a square matrix ($M = N$) a solution can be found to the linear model if the $\mathbf{A}$ has full rank – $\mathbf{A}$ consist of linearly independent columns or rows. For $M > N$ the matrix said to have full rank when the columns are linearly independent. For $M < N$ the matrix has full rank when the rows are linearly independent. For linearly systems/model with $M = N$ is called determined, $M > N$ overdetermined and $M < N$ under-determined.
When full rank do not occur the matrix is then called rank-deficient.

By inverting $\mathbf{A}$ from (3.1) the unknowing vector $\mathbf{x}$ can be achieved. Square matrix is invertible if an only if its has full rank or its determinant $\det(\mathbf{A}) \neq 0$. For rectangular matrices ($M > N$ and $M < N$) left-sided and right-sided inverse exists. With the left-inverse the least norm solution of (3.1) can be found.

For an determined system there will exist a unique solution. For an overdetermined system there do not exist a solution and for under-determined systems there exist infinitely many solutions.
For rank-deficient matrices there do not exist an inverse and therefore (3.1) can not be solve by inverting the model. But rank-deficient matrices meaning have non-empty null space leading to infinitely solutions to (3.1) [7, p. ix].

hvad siger vi her

As described in chapter 1 the linear model of interest consist of M sensors which is known and N sources which is unknown. Furthermore, we also have that $M < N$ – an under-determined system. It is therefore of interest to find a solution in the infinitely solution set.

3.2 Compressive Sensing

Compressive sensing is the theory of efficient recovery or reconstruction of a signal from a minimal number of observed measurements. Assuming linear acquisition of the original information the relation between the measurements and the signal to be recovered is described by the linear model giving in (3.1) [10].

In compressive sensing terminology, $\mathbf{x}$ is the signal of interest which is sought recovered by solving the linear system (3.1). In the typical compressive sensing case where $M < N$ the system becomes under-determined and there are infinitely many solutions, provided that a solution exist. Such system is also referred to as over-complete *(as the number of column basis vectors is greater than the dimension of the input).*
However, by enforcing certain sparsity constraints it is possible to recover the wanted signal [10], hence the term sparse signal recovery.

3.2.1 Sparseness

A signal is said to be k-sparse if the signal has at most k non-zero coefficients. For the purpose of counting the non-zero entries of a vector representing a signal the ℓ_0-norm is defined

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

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

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

where $[N]$ a set of integers $\{1, 2, \cdots, N\}$ [10, p. 41]. The set of all k-sparse signals is denoted as

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

3.2.2 Optimisation Problem

To find a solution to the linear model (3.1) which is k-sparse with $k < M$ an optimisation problem can be used. An optimisation problem is defined as

mangler vi ikke at skrive hvorfor vi skal løse nummerisk og ikke analytisk. er det ikke bare at der i k-sparse tilfælde kun findes en nummerisk løsning?

$$\min f_0(\mathbf{x}) \quad \text{s.t} \quad f_i(\mathbf{x}) \leq b_i, \quad i \in [n],$$

where $f_0 \ : \ \mathbb{R}^N \mapsto \mathbb{R}$ is an objective function and $f_i \ : \ \mathbb{R}^N \mapsto \mathbb{R}$ are the constraint functions.
To find the k-sparse solution our optimisation problem can be written as

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

where $\mathbf{x}$ is all possible candidates to a k-sparse signal $\mathbf{x}^*$. The objective function is given by an ℓ_0 norm with the constraint function been the linear model described in (3.1). Unfortunately, this optimisation problem is non-convex due to the definition of ℓ_0-norm and is therefore difficult to solve – it is a NP-hard problem. Instead by replacing the ℓ_0-norm with the ℓ_1-norm, the optimisation problem can approximated and therefore become computational feasible [7, p. 27]

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

With this optimisation problem we find the best k-term approximation of the signal $\hat{\mathbf{x}}^*$. 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 3.2.1
A measurement matrix $\mathbf{A}$ is defined as $\mathbf{A} \in \mathbb{R}^{M \times N}$ with columns $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_N]$. By assuming uniqueness of a solution $\mathbf{x}^*$ of

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

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 we need to realise that the size of the set $\{\mathbf{a}_j, j \in S\} \leq M$, that we can not have more than M linearly independence columns. So when $M \ll N$ then we automatically achieve a sparse signal.

Proof
By way of contradiction, lets assume that the set $\{\mathbf{a}_j, j \in S\}$ is linearly dependent with $S = \text{supp}(\mathbf{x}^*)$. This means that there exists a non-zero vector $\mathbf{v} \in \mathbb{R}^N$ supported on S such that $\mathbf{Av} = \mathbf{0}$. Then, for any $t \neq 0$,

$$\|\mathbf{x}^*\|_1 < \|\mathbf{x}^* + t\mathbf{v}\|_1 = \sum_{j\in S} |x_j^* + tv_j| = \sum_{j\in S} \text{sgn}(x_j^* + tv_j)(x_j^* + tv_j).$$

If $|t|$ is small enough, namely, $|t| < \min_{j\in S} \frac{|x_j^*|}{\|\mathbf{v}\|_\infty}$, then

$$\text{sgn}(x_j^* + tv_j) = \text{sgn}(x_j^*), \quad \forall j \in S.$$

It the follows that

$$\|\mathbf{x}^*\|_1 < \sum_{j\in S} \text{sgn}(x_j^*)(x_j^* + tv_j) = \sum_{j\in S} \text{sgn}(x_j^*)x_j^* + t\sum_{j\in S} \text{sgn}(x_j^*)v_j = \|\mathbf{x}^*\|_1 + t\sum_{j\in S} \text{sgn}(x_j^*)v_j.$$

This is a contradiction, because we can always choose a small $t \neq 0$ such that $t\sum_{j\in S} \text{sgn}(x_j^*)v_j \leq 0$ and therefore the set $\{\mathbf{a}_j, j \in S\}$ must be linearly independent.■

The Basis Pursuit makes the foundation of several algorithms solving alternative versions of (3.2) where noise is incorporated. A different type of solution method includes a greedy algorithm such as the Orthogonal Matching Pursuit [10, P. 65].

er OMP ikke en basis pursuit metode

Algorithm 1 Orthogonal Matching Pursuit (OMP)

1. Give an measurement matrix $\mathbf{A}$ and a measurement vector $\mathbf{y}$
2. Initial $S^0 = \varnothing$ and $x_0 = \mathbf{0}$
3. Iterate until stopping criterion is met:

$$S^{n+1} = S^n \cup \{j_{n+1}\}, \quad j_{n+1} = \arg\max_{j\in[N]} (|(\mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n))_j|)$$
$$x^{n+1} = \arg\min_{\mathbf{z}\in\mathbb{C}^N} (\|\mathbf{y} - \mathbf{A}\mathbf{z}\|_2, \text{supp}(\mathbf{z}) \subset S^{n+1})$$

4. $\mathbf{x}^* = \mathbf{x}^n$

3.2.3 Conditions on the Dictionary

In section 3.2.2 the measurement matrix $\mathbf{A}$ was known regarding to solve the optimisation problem given in (3.2). But this is not always the case and instead a

dictionary is used to find the measurement matrix $\mathbf{A}$. To ensure an exact or an approximate reconstruction of the sparse signal $\mathbf{x}^*$ from the optimisation problem (3.2) some conditions associated to the matrix $\mathbf{A}$ must be satisfied.

Null Space Condition

One condition of $\mathbf{A}$ is the null space property. The null space of the matrix A is defined as

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

The null space property is defined as

Definition 3.1 (Null Space Property)
A matrix $\mathbf{A} \in \mathbb{R}^{M\times N}$ is said to satisfy the null space property relative to a set $S \subset [N]$ if

$$\|\mathbf{v}_S\|_1 < \|\mathbf{v}_{\overline{S}}\|_1 \quad \text{for all} \quad \mathbf{v} \in \ker(\mathbf{A} \setminus \{\mathbf{0}\}), \tag{3.3}$$

where the vector $\mathbf{v}_S$ is the restriction of $\mathbf{v}$ to the indices in S.

Theorem 3.2.2
Given a matrix $\mathbf{A} \in \mathbb{R}^{M\times N}$, every vector $\mathbf{x} \in \mathbb{R}^N$ supported on a set S is the unique solution of (3.2)) with $\mathbf{y} = \mathbf{A}\mathbf{x}$ if and only if $\mathbf{A}$ satisfies the null space property relative to S.

Proof
$\Rightarrow$: Given a fixed index set $S \subseteq [N]$, lets first assume that every vector $\mathbf{x} \in \mathbb{R}^N$ supported on S is the unique minimiser of $\|\mathbf{z}\|_1$ subject to $\mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x}$. Thus, for any $\mathbf{v} \in \ker(\mathbf{A}) \setminus \{\mathbf{0}\}$, the vector $\mathbf{v}_S$ is the unique minimizer $\|\mathbf{z}\|_1$ subject to $\mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{v}_S$. But we have that $\mathbf{0} = \mathbf{A}(\mathbf{v}_S + \mathbf{v}_{\overline{S}}) \implies \mathbf{A}\mathbf{v}_S = \mathbf{A}(-\mathbf{v}_{\overline{S}}$ and $-\mathbf{v}_{\overline{S}} \neq \mathbf{v}_S$ or else $\mathbf{v} = \mathbf{0}$. We conclude that $\|\mathbf{v}_S\|_1 < \|\mathbf{v}_{\overline{S}}\|_1$. This establishes the null space property relative to S.

$\Leftarrow$: Conversely, lets assume that the null space property relative to S holds. Then, given $S \subseteq [N]$ with null space property and a vector $\mathbf{x} \in \mathbb{R}^N$ supported on S and a vector $\mathbf{z} \in \mathbb{R}^N$, $\mathbf{z} \neq \mathbf{x}$, satisfying $\mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x}$, we consider the vector $\mathbf{v} := \mathbf{x} - \mathbf{z} \in \ker(\mathbf{A}) \setminus \{\mathbf{0}\}$. In view of the null space property, we obtain

$$\begin{aligned} \|\mathbf{x}\|_1 \leq \|\mathbf{x} - \mathbf{z}_S\|_1 &= \|\mathbf{v}_S\|_1 + \|\mathbf{z}_S\|_1 \\ &< \|\mathbf{v}_{\overline{S}}\|_1 + \|\mathbf{z}_S\|_1 \\ &= \|-\mathbf{z}_{\overline{S}}\|_1 + \|\mathbf{z}_S\|_1 = \|\mathbf{z}\|_1 \end{aligned}$$

This establishes the required minimality of $\|\mathbf{x}\|_1$. ■

Unfortunately, this is a property/condition which is hard to check in practice.

Coherence

The null space property provide a unique solution to the optimisation problem, (3.2), but is unfortunately complicated to investigate. Instead an alternative measure used for sparsity is presented.
Coherence is a measure of quality and determine if the matrix $\mathbf{A}$ is a good choice for the optimisation problem (3.2). A small coherence describe the performance of a recovery algorithm as good with that choice of $\mathbf{A}$.

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

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

Restricted Isometry Condition

Restricted isometry condition is a stronger condition concerning the orthogonality of the matrix $\mathbf{A}$.

Definition 3.3 (Restricted Isometry Property)
A matrix A satisfies the RIP of order k if there exists a $\delta_k \in (0,1)$ such that

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

Theorem 3.2.3
Suppose that the $2s$-th restricted isometry constant of the matrix $\mathbf{A} \in \mathbb{R}^{M\times N}$ satisfies

$$\delta_{2s} < \frac{1}{3}.$$

Then every s-sparse vector $\mathbf{x}^* \in \mathbb{R}^N$ is the unique solution of

$$\min_{\mathbf{z}\in\mathbb{R}^N} \|\mathbf{z}\|_1 \quad \text{subject to} \quad \mathbf{Az} = \mathbf{Ax}.$$

Proof
To proof the theorem we only need to show the null space condition:

$$\|\mathbf{v}\|_1 < \frac{1}{2}\|\mathbf{v}\|_1, \quad \forall\ \mathbf{v} \in \ker(\mathbf{A}) \setminus \{\mathbf{0}\},\ S \subseteq [N],\ \mathrm{card}(S) \leq s.$$

Cf. Cauchy-Schwarz or $\|\mathbf{v}_S\|_1 \leq \|\mathbf{v}_S\|_2\sqrt{s}$, we only need to show

$$\|\mathbf{v}_S\|_2 \leq \frac{\rho}{2\sqrt{s}}\|\mathbf{v}\|_1$$
$$\rho = \frac{2\delta_{2s}}{1-\delta_{2s}} < 1,$$

whenever $\delta_{2s} < 1/3$. Given $\mathbf{v} \in \ker(\mathbf{A}) \setminus \{\mathbf{0}\}$, it is enough to consider an index set $S = S_0$ of s largest absolute entries of the vector $\mathbf{v}$. The complement $\overline{S_0}$ of S_0 in $[N]$ is partition as $S_0 = S_1 \cup S_2 \cup \cdots$, where

$$\begin{aligned} S_1 &: \text{ index set of } s \text{ largest absolute entries of } \mathbf{v} \text{ in } \overline{S_0}, \\ S_2 &: \text{ index set of } s \text{ largest absolute entries of } \mathbf{v} \text{ in } \overline{S_0 \cup S_1}. \end{aligned}$$

With $\mathbf{v} \in \ker(\mathbf{A})$:

$$\mathbf{A}(\mathbf{v}_{S_0}) = \mathbf{A}(-\mathbf{v}_{S_1} - \mathbf{v}_{S_2} - \cdots),$$

so that

$$\begin{aligned} \|\mathbf{v}_{S_0}\|_2^2 &\leq \frac{1}{1-\delta_{2s}}\|\mathbf{A}(\mathbf{v}_{S_0})\|_2^2 = \frac{1}{1-\delta_{2s}}\langle \mathbf{A}(\mathbf{v}_{S_0}), \mathbf{A}(-\mathbf{v}_{S_1}) + \mathbf{A}(-\mathbf{v}_{S_2}) + \cdots \rangle \\ &= \frac{1}{1-\delta_{2s}} \sum_{k\geq 1} \langle \mathbf{A}(\mathbf{v}_{S_0}), \mathbf{A}(-\mathbf{v}_{S_k}) \rangle. \end{aligned} \tag{3.4}$$

According to Proposition **??**, we also have

$$\langle \mathbf{A}(\mathbf{v}_{S_0}), \mathbf{A}(-\mathbf{v}_{S_k}) \rangle. \leq \delta_{2s}\|\mathbf{v}_{S_0}\|_2\|\mathbf{v}_{S_k}\|_2. \tag{3.5}$$

Substituting (3.5) into (3.4) and dividing by $\|\mathbf{v}_{S_0}\|_2 > 0$ ■

3.2.4 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 consist 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 $\mathbf{E} \in \mathbb{R}^{M\times L}$:

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{E}. \tag{3.6}$$

L denote the number of observed measurement vectors each consisting of M measurements. For $L = 1$ the linear model will just be the SMV model (3.1).
The matrix $\mathbf{X}$ consist of $\{\mathbf{x}_i\}_{i=1}^{L}$ sparse vectors which has been stack column-wise leading to that $\mathbf{X}$ consist of at most k non-zero rows. As for the SMV model (3.1) the MMV model (3.6) is under-determined with $M \ll N$ and the number of samples is less than M, $L < M$ [7, p. 42].

vi skal lige have s på om vi kun anta identisk sparsness søjler, og i så fald data vi arbejde m den form

hvorfor L<M?

The support of $\mathbf{X}$ denote 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 mention before $\mathbf{X}$ has at most k non-zero rows, the non-zero values occur in common location. By using this joint information it is possible to recover $\mathbf{X}$ from fewer measurements.
By using the rank of $\mathbf{X}$ which give us information of the amount of linearly independent rows or columns and the spark of $\mathbf{A}$ the minimum set of linearly dependent columns it is possible to set some condition on the measurement to ensure recovery. When the support of $\mathbf{X}$ is equal to the amount of non-zero rows in $\mathbf{X}$, $|\text{supp}(\mathbf{X})| = k$ then the rank of $\mathbf{X}$ would be $\text{rank}(\mathbf{X}) \leq k$. If $\text{rank}(\mathbf{X}) = 1$ then are the k-sparse vectors $\{\mathbf{x}_i\}_{i=1}^{L}$ multiple of each other the joint information cannot be taken advantage of. But for large rank we can exploit the diversity of the columns in $\mathbf{X}$. This can be defined in a sufficient and necessary condition of the MMV model (3.6) which must have

antal, og er det ikke altid det?

$$|\text{supp}(\mathbf{X})| < \frac{\text{Spark}(\mathbf{A}) - 1 + \text{rank}(\mathbf{X})}{2}$$

such that $\mathbf{X}$ can uniquely be determined.
This result lead to that row-sparse matrix $\mathbf{X}$ with large rank can be recovered from fewer measurement [7, p. 43].

3.3 Dictionary learning

As clarified in section 3.2.3 the choice of dictionary matrix $\mathbf{A}$ is essential to achieve the best recovery of a 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 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].

Consider now $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_L]$, $\mathbf{y}_i \in \mathbb{R}^M$ as a training database, created by $\mathbf{y}_i = \mathbf{A}\mathbf{x}_i$ for which we want to learn the best suitable dictionary $\mathbf{A}$ and sparse representation $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_L]$, $\mathbf{x}_i \in \mathbb{R}^N$. 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]

$$\min_{\mathbf{A},\mathbf{X}} \sum_{i=1}^{L} \|\mathbf{y}_i - \mathbf{A}\mathbf{x}_i\|_2^2 \quad st.\ \|\mathbf{x}_i\|_0,\ 1 \le i \le L. \tag{3.7}$$

The learning consist 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 $k < M$[5]. Furthermore, consider $\mathbf{A}_0$ such that every training signal can be represented by $k_0 < \text{spark}(\mathbf{A}_0)/2$ columns of $\mathbf{A}_0$, then $\mathbf{A}_0$ is a unique dictionary, up to scaling and permutation of columns[9]. Again the ℓ_0-norm lead to an NP-hard problem an heuristic methods are need.

fungerer disse to uniqueness parameter sammen?

3.3.1 K-SVD

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 constrict $k = 1$ and the representation reduced to a binary scalar $x = 1, 0$. Further instead of computing the mean of K sub-sets the K-SVD algorithm computes the SVD factorisation of the K different sub-matrices that correspond to the K columns of $\mathbf{A}$.

maybe this should come in the end

The dictionary learning algorithm K-SVD provide an update rule which is applied to each column of $\mathbf{A}_0 = [\mathbf{a}_0, \cdots, \mathbf{a}_N]$. Updating first $\mathbf{a}_i$ and then the corresponding coefficients in $\mathbf{X}$ which it is multiplied with, that is the i^{th} row in $\mathbf{X}$ denoted by $\mathbf{x}_i^T$. Let $\mathbf{a}_{i_0}$ be the column to be updated and let the remaining columns be fixed. By rewriting the objective function in (3.7) using matrix notation it is possible to isolate the contribution from $\mathbf{a}_{i_0}$.

$$\begin{aligned} \|\mathbf{Y} - \mathbf{A}\mathbf{X}\|_F^2 &= \left\| \mathbf{Y} - \sum_{i=1}^{M} \mathbf{a}_i \mathbf{x}_i^T \right\|_F^2 \\ &= \left\| \left(\mathbf{Y} - \sum_{i \neq i_i}^{M} \mathbf{a}_i \mathbf{x}_i^T \right) - \mathbf{a}_{i_0} \mathbf{x}_{i_0}^T \right\|_F^2, \end{aligned} \tag{3.8}$$

where F is the Frobenius norm that works on matrices

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

In (3.8) the term in the parenthesis makes the an error matrix $\mathbf{E}_{i_0}$ without the contribution from i_0, hence minimising (3.8) with respect to $\mathbf{a}_{i_0}$ and $\mathbf{x}_{i_0}^T$ leads to the optimal contribution from i_0 (can I say it this way..?).

$$\min_{\mathbf{a}_{i_0},\mathbf{x}_{i_0}^T} \left\| \mathbf{E}_{i_0} - \mathbf{a}_{i_0},\mathbf{x}_{i_0}^T \right\|_F^2 \tag{3.9}$$

The optimal solution to (3.9) is known to be the rank-1 approximation of $\mathbf{E}_{i_0}$. 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 $\mathbf{E}_{i_0}$.
That is specifically that for $\mathbf{E}_{i_0} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \in \mathbb{R}^{M\times N},\ M \leq N$ with

$$\mathbf{U} = [\mathbf{u}_1, \cdots, \mathbf{u}_M] \in \mathbb{R}^{M\times M}, \quad \mathbf{\Sigma} = \operatorname{diag}[\sigma_1, \cdots, \sigma_m] \in \mathbb{R}^{M\times N}, \quad \mathbf{V} = [\mathbf{v}_1, \cdots, \mathbf{v}_N] \in \mathbb{R}^{N\times N}$$

where $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices, i.e. $\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{I}$, and σ_j is the non-negative singular values of $\mathbf{E}_{i_0}$ such that $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$. The best k-rank approximation to $\mathbf{E}_{i_0}$, with $k < rank(\mathbf{E}_{i_0})$ is then given by[**Wiki..**]

$$\mathbf{E}_{i_0}^{(k)} = \sum_{j=1}^{k} \sigma_j \mathbf{u}_j \mathbf{v}_j^T.$$

Since the outer product always have rank-1 letting $\mathbf{a}_{i_0} = \mathbf{u}_1$ and $\mathbf{x}_{i_0}^T = \sigma_i \mathbf{v}_1^T$ solves the optimisation problem (3.9). However in order to preserve the sparsity in $\mathbf{X}$ while optimising, only the non-zero entries in $\mathbf{x}_{i_0}^T$ are allowed to vary. For this purpose only a subset of columns in $\mathbf{E}_{i_0}$ is considered, those which correspond to the non-zero entries of $\mathbf{x}_{i_0}^T$. A matrix $\mathbf{P}_{i_0}$ is defined such that $\mathbf{x}_{i_0}^{T(R)} = \mathbf{x}_{i_0}^T\mathbf{P}_{i_0}$ is restricted to contain only the M_{j_0} non-zero entries of $\mathbf{x}_{i_0}^T$. By applying SVD to the sub-matrix $\mathbf{E}_{i_0}\mathbf{P}_{i_0}$ and updating $\mathbf{a}_{i_0}$ and $\mathbf{x}_{i_0}^{T(R)}$ the rank-1 approximation is found and the original representation vector is updated as $\mathbf{x}_{i_0}^T = \mathbf{x}_{i_0}^{T(R)}\mathbf{P}_{i_0}^T$.

The main steps of K-SVD is described in algorithm 2.

Algorithm 2 K-SVD

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

procedure K-SVD($\mathbf{A}_{(0)}$)
  normilize columns of $\mathbf{A}_{(0)}$
  while $error \geq limit$ **do**
    $j = j + 1$
    for $j \leftarrow 1, 2, \cdots, L$ **do** ▷ updating each col. in $\mathbf{X}_{(k)}$
      $\hat{\mathbf{x}}_j = \min_{\mathbf{x}} \|\mathbf{y}_j - \mathbf{A}_{(k-1)}\mathbf{x}_j\| \quad s.t. \quad \|\mathbf{x}_j\| \leq k_0$
    end for
    $\mathbf{X}_{(k)} = \{\hat{\mathbf{x}}_j\}_{j=1}^{L}$
    for $i_0 \leftarrow 1, 2, \cdots, N$ **do**
      $\Omega_{i_0} = \{j | 1 \leq j \leq L, \mathbf{X}_{(k)}[i_0, j] \neq 0\}$
      From Ω_{i_0} define $\mathbf{P}_{i_0}$
      $\mathbf{E}_{i_0} = \mathbf{Y} - \sum_{i \neq i_0}^{M} \mathbf{a}_i \mathbf{x}_i^T$
      $\mathbf{E}_{i_0}^R = \mathbf{E}_{i_0} \mathbf{P}_{i_0}$
      $\mathbf{E}_{i_0}^R = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ ▷ perform SVD
      $\mathbf{a}_{i_0} \leftarrow \mathbf{u}_1$ ▷ update the i_0 col. in $\mathbf{A}_{(k)}$
      $\left(\mathbf{x}_{i_0}^T\right)^R \leftarrow \sigma_1 \mathbf{v}_1$
      $\mathbf{x}_{i_0}^T \leftarrow \left(\mathbf{x}_{i_0}^T\right)^R \mathbf{P}_{i_0}^T$ ▷ 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

3.4 Independent Component Analysis

Independent component analysis (ICA) is a method that applies to the general problem of decomposition of a measurement vector into a source vector and a mixing matrix. The intention of ICA is to separate a multivariate signal into statistical independent and non-Gaussian signals and furthermore identify the mixing matrix $\mathbf{A}$, given only the observed measurements $\mathbf{Y}$. A well known application example of source separation is the cocktail party problem, where it is sought to listen to one specific person speaking in a room full of people having interfering conversations. Let $\mathbf{y} \in \mathbb{R}^M$ be a single measurement from M microphones containing a linear mixture of all the speak signal that are present in the room. When additional noise is not

considered the problem can be described as the familiar linear model,

$$\mathbf{y} = \mathbf{A}\mathbf{x} \tag{3.10}$$

where $\mathbf{x} \in \mathbb{R}^N$ contain the N underlying speak signals and $\mathbf{A}$ is a mixing matrix where the coefficients depends (more or less?) on the distance from the source to the microphone. As such each y_i is a weighted sum of all the present sources of speak.

By ICA both the mixing matrix $\mathbf{A}$ and the sources signals $\mathbf{x}$ are sought estimated from the observed measurements $\mathbf{y}$. The main attribute of ICA is the assumption that the sources in $\mathbf{x}$ are statistical independent and non-Gaussian distributed, hence the name independent components.

when we assume independence it is enough to solve system, why?

By independence, one means that changes in one source signal do not affect the other source signals. Theoretically that is the joint probability density function (pdf) of $\mathbf{x}$ can be factorised into the product of the marginal pdfs of the components x_i

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

herover er ukommenteret et afsnit jeg ikke forstår

The possibility of separating a signal into independent and non-Gaussian components originates from the central limit theorem[13, p. 34]. The theorem state that the distribution any linear mixture of two or more independent random variables tents toward a Gaussian distribution, under certain conditions. Thus, when a non-Gaussian distribution of the independent components is achieved through optimization it must be the original sources.

3.4.1 Assumptions and Preprocessing

For simplicity assume $\mathbf{A}$ is square i.e. $M = N$ and invertible. As such when $\mathbf{A}$ has been estimated the inverse is computed the components can simply be estimated as $\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}$[13, p. 152-153].
As both $\mathbf{A}$ and $\mathbf{x}$ are unknown the variances of the independent components can not be determined. However it is reasonable to assume that $\mathbf{x}$ has unit variance, as $\mathbf{A}$ will adapt to this restriction. Any scalar multiplier within a source can be cancelled out by dividing the corresponding column in $\mathbf{A}$ with the same scalar.
This of course must also apply to the mixing matrix $\mathbf{A}$ which will be restricted in the recovery method, described in section 3.4.2. For further simplification it is assumed without loss of generality that $\mathbb{E}[\mathbf{y}] = 0$ and $\mathbb{E}[\mathbf{x}] = 0$[13, p. 154]. In case this assumption is not true, the measurements can be centred by subtracting the mean as preprocessing before doing ICA.

check up on A restrictions

A preprocessing step central to ICA is to whiten the measurements $\mathbf{y}$. By the whitening process any correlation in the measurements are removed and unit variance is ensured. This ensures that the independent components $\mathbf{x}$ are uncorrelated and have unit variance(true?). Furthermore, this reduces the complexity of ICA and therefore

whiteining is a linear change of coordinates of the mixed datahttp://arnauddelorme.com/ica_for_dummies/ "By rotating the axis and minimizing Gaussianity of the projection in the first scatter plot, ICA is able to recover the original sources which are statistically independent

simplifies the recovering process.
Whitening is a linear transformation of the observed data. That is multiplying the measurement vector $\mathbf{y}$ with a whitening matrix $\mathbf{V}$,

$$\mathbf{y}_{white} = \mathbf{V}\mathbf{y}$$

to obtain a new measurement vector $\mathbf{y}_{white}$ that is white. To obtain a whitening matrix the eigenvalue decomposition (EVD) of the covariance matrix can be used,

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

here $\mathbf{D}$ is a diagonal matrix of eigenvalues and $\mathbf{E}$ is the associated eigenvectors. From $\mathbf{E}$ and $\mathbf{D}$ a whitening matrix is constructed [13, p.159].

$$\mathbf{V} = \mathbf{E}\mathbf{D}^{-1/2}\mathbf{E}^T.$$

Where $\mathbf{D}^{-1/2} = diagd_1^{-1/2}, \cdots, d_n^{-1/2}$ is a componentwise operation.
By multiplying the measurement vector $\mathbf{y}$ with a whitening matrix $\mathbf{V}$ the data becomes white

$$\mathbf{y}_{\text{white}} = \mathbf{V}\mathbf{y} = \mathbf{V}\mathbf{A}\mathbf{x} = \mathbf{A}_{\text{white}}\mathbf{x}$$

Furthermore the mixing matrix $\mathbf{A}_{\text{white}}$ becomes orthogonal

$$\mathbb{E}[\mathbf{y}_{\text{white}}\mathbf{y}_{\text{white}}^T] = \mathbf{A}_{\text{white}}\mathbb{E}[\mathbf{x}\mathbf{x}^T]\mathbf{A}_{\text{white}}^T = \mathbf{A}_{\text{white}}\mathbf{A}_{\text{white}}^T = \mathbf{I}.$$

Consequently ICA can restrict its search for the mixing matrix to the orthogonal matrix space – That is instead of estimating n^2 parameters ICA now only has to estimate an orthogonal matrix which has $n(n-1)/2$ parameters/degrees of freedom [13, p. 159] .

se udkommentering herunder?

3.4.2 Recovery of the Independent Components

Now the ICA model is established, the next step is the estimation of the mixing coefficients a_{ij} and independent components x_i. The simple and intuitive method is to take advantage of the assumption of non-Gaussian independent components. Consider again the ICA model of a single measurement vector $\mathbf{y} = \mathbf{A}\mathbf{x}$ where the independent components can be estimated by the inverted model $\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}$. Let $\mathbf{A}^{-1} = \mathbf{B}$, now a single independent component can be seen as the linear combination

$$x_j = \mathbf{b}_j^T\mathbf{y} = \sum_i b_i y_i \tag{3.11}$$

where $\mathbf{b}_j^T$ is the j^{th} row of $\mathbf{B}$. The issue is now to determined $\mathbf{b}_j$ such that it equals the j^{th} row from the inverse $\mathbf{A}$. As $\mathbf{A}$ is unknown it is not possible to determine

$\mathbf{b}_j$ exactly, but an estimate can be found to make a good approximation. Rewriting (3.11)

$$x_j = \mathbf{b}_j^T \mathbf{y} = \mathbf{b}_j^T \mathbf{A}\mathbf{x} = \mathbf{q}^T \mathbf{x} = \sum_{i=1} q_i x_i$$

it is seen how x_j is a linear combination of all $\mathbf{x}_i$, thus the equality only holds true when $\mathbf{q}$ consist of only one non-zero element that equals 1. Due to the central limit theorem the distribution of $\mathbf{q}^T\mathbf{x}$ most non-Gaussian when it equals one of the independent components which was assumed non-Guassian. Then, since $\mathbf{q}^T\mathbf{x} = \mathbf{b}_j^T\mathbf{y}$, it is possible to vary the coefficients in $\mathbf{b}$ and look at the distribution of $\mathbf{b}_j^T\mathbf{y}$. Finding the vector $\mathbf{b}$ that maximize the non-Gaussianity would then corresponds to $\mathbf{q} = \mathbf{A}^T\mathbf{b}$ having only a single non-zero element. Thus maximizing the non-Gaussianity of $\mathbf{b}_j^T\mathbf{y}$ results in one of the independent components[13, p. 166].
Considering the n-dimensional space of vectors $\mathbf{b}$ there exist $2n$ local maxima, corresponding to x_i and $-x_i$ for all n independent components[13, p. 166].

3.4.3 Kurtosis

To maximize the non-gaussianity a measure for gaussianity is needed. Kurtosis is a quantitative measure used for nongaussianity of random variables. Kurtosis of a random variable y is the fourth-order cumulant denotet by $\text{kurt}(y)$. For y with zero mean and unit variance kurtosis reduces to

uddyb? og tilføj kilde til kurtosis

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

It is seen that the kurtosis is a normalized version of the fourth-order moment defined as $\mathbb{E}[y^4]$. For a Gaussian random variable the fourth-order moment equals $3(\mathbb{E}[y^2])^2$ hence the corresponding kurtosis will be zero [13, p. 171]. Consequently the kurtosis of non-Gaussian random variables will almost always be different from zero.
The kurtosis is a common measure for non-Gaussianity due to its simplicity both theoretical and computational. The kurtosis can be estimated computationally by the forth-order moment of sample data when the variance is constant. Furthermore, for two independent random variables x_1, x_2 the following linear properties applies to the kurtosis of the sum

$$\text{kurt}(x_1 + x_2) = \text{kurt}(x_1) + \text{kurt}(x_2) \quad \text{and} \quad \text{kurt}(\alpha x_1) = \alpha^4 \text{kurt}(x_1)$$

However, one complication concerning kurtosis as a measure is that kurtosis is sensitive to outliers [13, p. 182].

Consider again the vector $\mathbf{q} = \mathbf{A}^T\mathbf{b}$ such that $\mathbf{b}_j^T\mathbf{y} = \sum_{i=1} q_i x_i$. By the additive property of kurtosis

$$\text{kurt}\left(\mathbf{b}_j^T\mathbf{y}\right) = \sum_{i=1} q_i^4 \text{kurt}(x_i).$$

Then the assumption of the independent components having unit variance results in $\mathbb{E}[x_j] = \sum_{i=1} q_i^2 = 1$. That is geometrically that $\mathbf{q}$ is constraint to the unit sphere, $\|\mathbf{q}\|^2 = 1$. By this the optimisation problem of maximising the kurtosis of $\mathbf{b}_j^T\mathbf{y}$ is similar to maximizing $|\text{kurt}(x_j)| = |\sum_{i=1} q_i^4 \text{kurt}(x_i)|$ on the unit sphere.
Due to the described preprocessing $\mathbf{b}$ is assumed to be white and it can be shown that $\|\mathbf{q}\| = \|\mathbf{b}_j\|$[13, p. 174]. This show that constraining $\|\mathbf{q}\|$ to one is similar to constraining $\|\mathbf{b}_j\|$ to one.

hvordan kommer dette frem?

3.4.4 The Gradient Algorithm with Kurtosis

In practise, to recover the mixing matrix $\mathbf{A}$ by maximizing the kurtosis of $\mathbf{b}_j^T\mathbf{y}$, gradient optimisation methods are used.
The general idea behind a gradient algorithm is to determine the direction for which $\text{kurt}(\mathbf{b}_j^T\mathbf{y})$ is growing the most, based on the gradient.

The gradient of $|\text{kurt}(\mathbf{b}_j^T\mathbf{y})|$ is computed as

$$\frac{\partial|\text{kurt}(\mathbf{b}_j^T\mathbf{y})|}{\partial\mathbf{b}_j} = 4\text{sign}(\text{kurt}(\mathbf{b}_j^T\mathbf{y}))(\mathbb{E}[\mathbf{y}(\mathbf{b}_j^T\mathbf{y})^3] - 3\mathbf{y}\mathbb{E}[(\mathbf{b}_j^T\mathbf{y})^2]) \tag{3.12}$$

As $\mathbb{E}[(\mathbf{b}_j^T\mathbf{y})^2] = \|\mathbf{y}\|^2$) for whitened data the corresponding term does only affect the norm of $\mathbf{b}_j$ within the gradient algorithm. Thus, as it is only the direction that is of interest, this term can be omitted. Because the optimisation is restricted to the unit sphere a projection of $\mathbf{b}_j$ onto the unit sphere must be performed in every step of the gradient method. This is done by dividing $\mathbf{b}_j$ by its norm. This gives update step

$$\begin{aligned}\Delta\mathbf{b}_j &\propto \text{sign}\left(\text{kurt}(\mathbf{b}_j^T\mathbf{y})\right)\mathbb{E}[\mathbf{y}(\mathbf{b}_j^T\mathbf{y})^3] \\ \mathbf{b}_j &\leftarrow \mathbf{b}_j/\|\mathbf{b}_j\|\end{aligned}$$

The expectation operator can be omitted in order to achieve an adaptive version of the algorithm, now using every measurement $\mathbf{y}$. However, the expectation operator from the definition of kurtosis can not be omitted and must therefore be estimated. This can be done by a time-average estimate, denoted as γ:

$$\Delta\gamma \propto ((\mathbf{b}_j^T\mathbf{y})^4 - 3) - \gamma$$

3.4.5 Basic ICA algorithm

Algorithm 3 combined the above theory, to give an overview of the ICA procedure. Estimating the mixing matrix and the corresponding independent components, from the given measurements.

Algorithm 3 Basis ICA

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 ICA($\mathbf{y}$)
    $k = 0$
    Initialise random vector $\mathbf{b}_{j(k)}$ ▷ unit norm
    Initialise random value $\gamma_{(k)}$
    for $j \leftarrow 1, 2, \cdots, N$ **do**
        while convergance critia not meet **do**
            $k = k + 1$
            $\mathbf{b}_{j(k)} \leftarrow \text{sign}\gamma_{(k-1)}\mathbf{y}(\mathbf{b}_j^T\mathbf{y})^3$
            $\mathbf{b}_{j(k)} \leftarrow \mathbf{b}_j / \|\mathbf{b}_j\|$
            $\gamma_{(k)} \leftarrow ((\mathbf{b}_j^T\mathbf{y})^4 - 3) - \gamma_{(k-1)}$
        end while
        $x_j = \mathbf{b}_j^T\mathbf{y}$
    end for
end procedure

3.4.6 ICA for sparse signal recovery

ICA is widely used within sparse signal recovery. When ICA is applied to a measurement vector $\mathbf{y} \in \mathbb{R}^M$ it is possible to separate the mixed signal into M or less independent components. However, by assuming that the independent components makes a k-sparse signal it is possible to apply ICA within sparse signal recovery of cases where $M < N$ and $k \leq M$.
To apply ICA to such case the independent components are obtained by the pseudo-inverse solution

$$\hat{\mathbf{x}} = \mathbf{A}_S^{\dagger}\mathbf{y}$$

where $\mathbf{A}_S$ is derived from the dictionary matrix $\mathbf{A}$ by containing only the columns associated with the non-zero entries of $\mathbf{x}$, specified by the support set S.

3.5 Limitations of compressive sensing

Through this chapter the concept of sparse signal recovery have been explained. The essential limitation of source recovery is that for $k > M$ it is not possible to recover $\mathbf{x}$ as the system becomes underdetermined. Similarly we can not find the

true dictionary $\mathbf{A}$ by dictionary learning methods because when $k > M$, any random dictionary can be used to create $\mathbf{y}$ from $\geq M$ basis vectors.
When considering source recovery from EEG measurements, described in section 1.1, it is not reasonable to assume that $k < M$ especially not in the case of low density EEG measurements. This motives the next two chapters where the possibility of sources recovery for $k > M$ is explored. The methods, proposed recently by O. Balkan, are taking advantage of the covariance domain and..

skal dette argumenteres yderligere, som værende uafhængig af motivations kapitlet?

Chapter 4

Cov-DL

4.1 Covariance-Domain Dictionary Learning

Covariance-domain dictionary learning (Cov-DL) is an algorithm proposed in [3] which claims to be able to identify more sources N than available observations M for the linear model described in (3.6).
The section is inspired by chapter 3 in [5] and the article [3].

Consider the multiple measurement vector model as above

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

which consist of multiple snapshots (L) in time such that the model consist of several measurements. The observed data becomes a matrix of size $\mathbf{Y} \in \mathbb{R}^{M \times L}$, the sources a matrix of size $\mathbf{X} \in \mathbb{R}^{N \times L}$ and the additional noise is of size $\mathbf{E} \in \mathbb{R}^{M \times L}$.
Let s be the index of time segments that the observed data $\mathbf{Y}$, is divided into and let f_s be the sample frequency the observed data has been sampled in. As such the observed data is divided into segments $\mathbf{Y}_s \in \mathbb{R}^{M \times t_s f_S}$, 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

snakker vi ikke lidt dobbelt her? L er længden af vores snapshot men længere nede dele vi hvor data ind i samples rate gange tid og dropper L notationen

Her blev jeg i tvivl :)

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

Cov-DL takes advantage of the dictionary framework and the transformation into another domain – covariance domain – to recover the mixing matrix $\mathbf{A}$ from the observed data $\mathbf{Y}_s$. An important aspect of this method is the prior assumption that the sources are statistical independent within their columns within the defined time segments. This implies the entries in $\mathbf{X}_s$ to be uncorrelated because of the independent columns .

Lige lidt hjælp her. Kan ikke huske snakken om og det kun skulle være independnet

As the Cov-DL algorithm only focus on finding the mixing matrix $\mathbf{A}$ the algorithm

must collaborate with another method/algorithm to determine the source matrix $\mathbf{X}_s$ such that a approximative representation of $\mathbf{Y}_s$ is possible. In this thesis the Multiple Sparse Bayesian Learning (M-SBL) algorithm is used for the source recovery and is described in section **??** .

en ny section. Tjek lige om det er bedre formuleret

In this section we assume that $\mathbf{X}$ is known and a random sparse matrix is used to represent the sources.

4.1.1 Covariances domain representation

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

Tjek lige om det er rigtig. Vi nævner jo M som værende observationer og L som snapshot (før stod der at L var observationer).

$$\sigma_{jk} = \frac{1}{L}\sum_{i=1}^{L}(y_{ji} - \bar{y}_j)(y_{ki} - \bar{y}_k).$$

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

Hvad menes der her?

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

$$\begin{aligned}\hat{\mathbf{\Sigma}}_{\mathbf{X}_s} &= \frac{1}{L}\mathbf{X}_s\mathbf{X}_s^T \\ &= \mathbf{\Lambda}_s\end{aligned}$$

From the assumption of uncorrelated sources in the entries of $\mathbf{X}_s$ the sample covariance matrix is expected to be nearly diagonal, thus it can be expressed as $\mathbf{\Lambda}_s$ which is a diagonal matrix consisting of the diagonal entries of $\hat{\mathbf{\Sigma}}_{\mathbf{X}_s}$ [3].

Each segment of observations can be modelled as

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

$$\begin{aligned}\hat{\mathbf{\Sigma}}_{\mathbf{Y}_s} = \frac{1}{L}\mathbf{Y}_s\mathbf{Y}_s^T &= \frac{1}{L}\left(\mathbf{A}\mathbf{X}_s\mathbf{X}_s^T\mathbf{A} + \mathbf{E}_s\mathbf{E}_s^T + 2\mathbf{A}\mathbf{X}_s\mathbf{E}_s^T\right) \\ &= \mathbf{A}\mathbf{\Lambda}_s\mathbf{A}^T + \mathbf{\Sigma}_{\mathbf{E}_s} \\ &= \mathbf{A}\mathbf{\Sigma}_{\mathbf{X}_s}\mathbf{A}^T + \mathbf{A}\mathbf{\Phi}_s\mathbf{A}^T \\ &= \mathbf{A}\mathbf{\Lambda}_s\mathbf{A}^T + \mathbf{A}\mathbf{\Phi}_s\mathbf{A}^T \\ &= \sum_{i=1}^{N}\mathbf{\Lambda}_{s_{ii}}\mathbf{a}_i\mathbf{a}_i^T + \mathbf{A}\mathbf{\Phi}_s\mathbf{A}^T \end{aligned} \tag{4.1}$$

Remember that N is the dimension of $\mathbf{X}_s$ hence the number of possible sources and k is the number of active sources (non-zero entries of $\mathbf{X}_s$).

It has been shown that by this model it is possible to identify $\mathcal{O}(M^2)$ sources given the true dictionary[16](dog er vektoriseringen også includeret i resultatet - så måske flyttes udtalelsen?). The purpose of the Cov-DL algorithm is instead to find the dictionary $\mathbf{A}$ from this expression and then still allow for $\mathcal{O}(M^2)$ sources to be identified.

Mangler dette stykke

4.1.2 Determination of the Dictionary

In order to enable the possibility of identifying $\mathcal{O}(M^2)$ sources and learning the corresponding dictionary $\mathbf{A}$ the model in (4.1) is rewritten.
At first both sides of (4.1) is vectorized. 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 $\text{vec}(\cdot)$ is defined to map a symmetric $M \times M$ matrix into a vector of size $\frac{M(M+1)}{2}$ making a vectorization of its lower triangular part. Furthermore, let $\text{vec}^{-1}(\cdot)$ be the inverse function.

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

$$\begin{aligned}\text{vec}(\mathbf{\Sigma}_{\mathbf{Y}_s}) &= \sum_{i=1}^{N} \mathbf{\Lambda}_{s_{ii}} \text{vec}(\mathbf{a}_i \mathbf{a}_i^T) + \text{vec}(\mathbf{E}_s) \\ &= \sum_{i=1}^{N} \mathbf{d}_i \mathbf{\Lambda}_{s_{ii}} + \text{vec}(\mathbf{E}_s) \\ &= \mathbf{D}\boldsymbol{\delta}_s + \text{vec}(\mathbf{E}_s), \quad \forall s. \end{aligned} \tag{4.2}$$

Note that $\mathbf{\Lambda}_{s_{ii}}$ is a scalar hence not vectorized. The vector $\boldsymbol{\delta}_s \in \mathbb{R}^N$ 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 $\mathbf{d}_i = \text{vec}(\mathbf{a}_i \mathbf{a}_i^T)$. Note that $\mathbf{D}$ and $\boldsymbol{\delta}_s$ are unknown while $\text{vec}(\mathbf{\Sigma}_{\mathbf{Y}_s})$ is known from the observed data.

The unknown variables $\mathbf{D}$ and $\boldsymbol{\delta}_s$ need to be found in order to make a full recovery. This can be done with dictionary learning methods such as K-SVD as mentioned in 3.3. With the learned variables (4.2) can be used to find the mixing matrix $\mathbf{A}$. Comparing this expression to the original compressive sensing problem (3.6) it is clear that higher dimensionality is achieved by representing the problem in the covariance domain.
From (4.2) we now have a problem with N unknowns in the sample covariance matrix $\mathbf{\Lambda}$ instead of (3.6) which have $N \cdot L$ unknowns in $\mathbf{X}$.

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

Cov-DL1 – Over-complete D

The case where $N > \frac{M(M+1)}{2}$ results in an under-determined system similar to the original system being under-determined when $N > M$. Though, 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. Note that this sparsity constraint is weaker than the original constraint $k < M$, as it allows for identification of remarkably more sources than within the original domain as it is not necessarily violated when $k > M$. Assuming a sufficient sparsity on $\boldsymbol{\delta}_s$ it is possible to learn the dictionary matrix of the covariance domain $\mathbf{D}$ by traditional dictionary learning methods, as introduced in section 3.3, applied to the observations represented in the covariance domain $\text{vec}(\boldsymbol{\Sigma}_{\mathbf{Y}_s})$ for all s. When $\mathbf{D}$ is known it is possible to find the original mixing matrix $\mathbf{A}$ through the relation $\mathbf{d}_i = \text{vec}(\mathbf{a}_i\mathbf{a}_i^T)$.
To do so each column is found by the optimisation problem

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

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

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

redegørelse for resultatet her skal laves

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

Algorithm 4 Cov-DL1 – over-complete **D**

1. Segmentation of measurements
2. Transformation to covariance domain
3. Vectorising
4. Learn D, by ?
5. Find A

Cov-DL2 – undercomplete D

kan ve udelade denne med det argument at vi altid vil finde så mange sources som muligt. fordi vi ikke ved hvor mange der der

Chapter 5

Multiple Sparse Bayesian Learning

As described in earlier chapters, the support set of sources is wish recovered. A way to recover the set is to use sparse bayesian learning (M-SBL) on the MMV model. To insure full recovery some sufficient condition must be applied on the dictionary matrix $\mathbf{A}$ and the sources.
The chapter is inspired by [**??**] .

indsæt ref for bayesian phd

- *$M \leq N$ (more sources than sensors)*
- *$k \leq N$ activations within the sources*
- *Instantaneous mixing – no time delay between samples*
- *The conditions on the support set are:*
  - *Orthogonality/uncorrelated of the active sources k.*
  - *Constraint on the dictionary matrix $\mathbf{A}$*

5.1 Maximum a Posterior Estimation

NOTES:

- *Finding sparse solutions to our optimisation problem can be view in Bayesian framework. We can find a sparse X by finding its estimates with maximum a posterior (MAP) estimation with use of a fixed and sparsity induced prior. This is also called empirical Bayesian with used a parameterized prior to encourage sparsity.*
- *With a global SBL minimum we always achieve a maximal sparse solution.*

- *By assuming some prior belif that Y has been generated by sparse coefficeinet expansion such that most ellemens in X are zero – called sparsity-induciong prior*
- *ALternative we can use empirical priors*
- *One problem – they all ensuing the inverse problem from Y to X becoming non-linear.*

As mention in section 3.2.4 a multiple measurement model (MMV) $\mathbf{Y} \in \mathbb{R}^{M \times L}$

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

models the mixing of a unknown source matrix $\mathbf{X} \in \mathbb{R}^{N \times L}$ with a mixing matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$ and some additional noise $\mathbf{E} \in \mathbb{R}^{M \times L}$ for L snapshots in time. The mixing matrix $\mathbf{A}$ is in this section known and is found with the Cov-DL as described in section XX . The non-zeros rows of $\mathbf{X}$ is referred to as the active sources and there are at most $k \leq M$ activations.

Sæt reference ind til Cov-DL

The goal is to find a source matrix $\mathbf{X}$ which is as sparse as possible to ensure a full recovery of our MMV (3.6).

One way to find the source matrix $\mathbf{X}$ is by finding its estimate, e.g. from the maximum a posterior (MAP) which have priors which induce sparsity of $\mathbf{X}$ given the measurement matrix $\mathbf{Y}$ and mixing matrix $\mathbf{A}$. By maximising the likelihood $p(\mathbf{Y}|\mathbf{X})$ and effective solution for an estimate of $\mathbf{X}$ is achieved when $M \ll N$. It is not always the case where $\mathbf{X}$ has a smaller dimensionality than $\mathbf{Y}$ because of the wanted sparsity of $\mathbf{X}$ leading to a matrix of greater dimensionality because of the added zeros. Hence the estimation becomes complicate as the MMV model becomes under-determined – an infinitely number of solutions with equal likelihoods.
As the optimisation problem of the MMV model is NP-hard another estimation method must be used.

indsæt her hvad tanken bag Bayesian egenligt er

Bayesian probability ... With this Bayesian framework the source matrix $\mathbf{X}$ seen as a variable can be drawn from some distribution $p(\mathbf{X})$ such that the infinitely solution space is narrowed.
E.g. $\mathbf{X}$ could be drawn from a Gaussian prior with zero-mean and covariance $\sigma_X^2\mathbf{I}$ where the additional noise $\mathbf{E}$ is independently Gaussian with covariance $\sigma_E^2\mathbf{I}$. The MAP estimator could then be rewritten to

$$\hat{\mathbf{X}} = \arg\max_{\mathbf{X}} p(\mathbf{Y}|\mathbf{X})p(\mathbf{X}) = \mathbf{A}^T(\lambda\mathbf{I} + \mathbf{A}\mathbf{A}^T)^{-1}\mathbf{Y},$$

with $\lambda = \sigma_E^2/\sigma_X^2$. With this distribution the estimate of the source matrix $\hat{\mathbf{X}}$ would have a large number of small non-zero coefficients.

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 $p(\mathbf{X}) \propto \exp(-\|\mathbf{X}\|_0)$ [**??**]. The MAP estimation problem is then rewritten to

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

5.2 Empirical Bayesian Estimation

From earlier MAP estimation approaches some problems occurs 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 . In other cases the prior can become to sparse and then lead to a combinatorial problem when looking for the global optima.
By using automatic relevance determination (ARD) the problem of using sparse prior can be overcome .

Skriv kort her hvad ARD er

With ARD we start looking at using a empirical prior for the MAP estimation as this prior is flexible and dependent on the unknown hyperparameter γ.

Let $p(\mathbf{Y}|\mathbf{X})$ be a Gaussian prior with a known noise variance λ. Then for each columns in $\mathbf{Y}$ and $\mathbf{X}$ the likelihood is written as

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

With 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

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

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

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

with the mean and covariance given as

$$\boldsymbol{\Sigma} = \mathrm{Cov}(\mathbf{x}_{\cdot j}|\mathbf{y}_{\cdot j};\boldsymbol{\gamma}) = \boldsymbol{\Gamma} - \boldsymbol{\Gamma}\mathbf{A}^T\left(\lambda\mathbf{I} + \mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^T\right)^{-1}\mathbf{A}\boldsymbol{\Gamma}, \quad \forall j = 1, \ldots, L \tag{5.1}$$

$$\mathcal{M} = [\boldsymbol{\mu}_{\cdot 1}, \ldots, \boldsymbol{\mu}_{\cdot L}] = \mathbb{E}[\mathbf{X}|\mathbf{Y};\boldsymbol{\gamma}] = \boldsymbol{\Gamma}\mathbf{A}^T\left(\lambda\mathbf{I} + \mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^T\right)^{-1}\mathbf{Y}, \tag{5.2}$$

where $\boldsymbol{\Gamma} = \mathrm{diag}(\boldsymbol{\gamma})$.

The row sparsity is achieved whenever $\gamma_i = 0$ leading to that the posterior must have the following probability

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

which ensure that the posterior mean $\mathcal{M}$ of the i-th row, $\boldsymbol{\mu}_{i\cdot}$, will be zero. Instead of estimating our source matrix $\mathbf{X}$ we instead estimate the hyperparameter γ_i.

Each of the hyperparameters γ_i correspond to different hypothesis for the prior distribution of the underlying generation of $\mathbf{Y}$. Therefore the determining of γ_i must be seen as a model selection in with we can use the empirical Bayesian stragetry. This evolve the task of treating the unknown source matrix $\mathbf{X}$ as nuisance parameters and integrating them out.

pestilens, plage, gene – de brugte dette ord i phden

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

$$\begin{aligned}\mathcal{L}(\gamma) &= -2\log(p(\mathbf{Y};\boldsymbol{\gamma})) = -2\log\left(\int p(\mathbf{Y}|\mathbf{X})p(\mathbf{X};\boldsymbol{\gamma})\ d\mathbf{X}\right) \\ &= L\log(|\lambda\mathbf{I} + \mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^T|) + \sum_{j=1}^{L}\mathbf{y}_{\cdot j}^T\left(\lambda\mathbf{I} + \mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^T\right)^{-1}\mathbf{y}_{\cdot j}\end{aligned}$$

To minimise the marginal 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 as mention in (5.1) while the M-step is a 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. The fixed point updating step is achieved by

taking the derivative of the marginal likelihood $\mathcal{L}(\boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$ and equating it with zero. This lead to the updating equation which can replace the one from M-step in the EM-algorithm:

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

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

Algorithm 5 M-SBL – See page 148 in PHD

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. ...

Notes:

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

Bibliography

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

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

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

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

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

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

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

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

[9] 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". In: *Measurement science review* 2 (2002).

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

Appendix A

Extended ICA Algorithms

Intro

A.1 Fixed-Point Algorithm - FastICA

A fixed-point algorithm to maximise the nongaussianity is more efficient than the gradient algorithm as the gradient algorithm converge slow depending on the choice of γ. The fixed-point algorithm is an alternative that could be used. By using the gradient given in (3.12) and then set the equation equal with $\mathbf{w}$:

$$\mathbf{w} \propto (\mathbb{E}[\mathbf{y}_{\text{white}}(\mathbf{w}^T\mathbf{y}_{\text{white}})^3] - 3\|\mathbf{w}\|^2\mathbf{w})$$

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

$$\mathbf{w} = \mathbb{E}[\mathbf{y}_{\text{white}}(\mathbf{w}^T\mathbf{y}_{\text{white}})^3] - 3\mathbf{w}$$

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

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

Algorithm 6 Fixed-Point Algorithm with Kurtosis

1. Center the observed data to achieve zero mean
2. Whiten the centered data
3. Create the initial random vector $\mathbf{w}$
4. Compute $\mathbf{w} = \mathbb{E}[\mathbf{y}_{\text{white}}(\mathbf{w}^T\mathbf{y}_{\text{white}})^3] - 3\mathbf{w}$
5. Normalise $\mathbf{w} = \frac{\mathbf{w}}{\|\mathbf{w}\|}$
6. Update $\mathbf{w}$
7. Repeat until convergence
8. Independent components are found as $\mathbf{x} = \mathbf{w}\mathbf{y}_{\text{white}}$

The fixed-point algorithm is also called for FastICA as the algorithm has shown to converge fast and reliably, then the current and previous $\mathbf{w}$ laid in the same direction [13, p. 179].

A.1.1 Negentropy

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

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

The entropy describes the information of a random variable. For variables becoming more random the entropy becomes larger, e.g. gaussian random variables have a high entropy, in fact gaussian random variables have the highest entropy among the random variables of the same variance [13, p. 182].

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

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

with $\mathbf{y}_{\text{gaus}}$ 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. Instead it could be an idea to use an approximation of the negentropy.

A.1.2 Approximation of Negentropy

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

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

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

For the generalised approximation the use of expectations of nonquadratic functions is introduced. The polynomial functions y^3 and y^4 from (A.1) are replaced by G^i with i being an index and G being some function. The approximation in (A.1) then becomes

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

The choice of G can lead to a better approximation than (A.1) and by choosing one which do not grow to fast more robust estimators can be obtained. The choice of G could be the two following functions

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

A.1.3 Algorithm with Negentropy

As described in section 3.4.2 the gradient algorithm is used to maximising negentropy. The gradient of the approximated negentropy is given as

$$\mathbf{w} = \gamma \mathbf{y}_{\text{white}} g(\mathbf{w}^T \mathbf{y}_{\text{white}})$$

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

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

For the choice of g the derivative of the functions presented in (A.2) could be use to achieve a robust result. Alternative a derivative which correspond to the fourth-power as seen in the kurtosis could be used. The functions g could be

$$\begin{aligned} g_1(y) &= \tanh(a_1 y), \quad 1 \leq a_1 \leq 2 \\ g_2(y) &= y \exp\left(\frac{-y^2}{2}\right) \\ g_3(y) &= y^3 \end{aligned} \tag{A.2}$$

Algorithm 7 Gradient Algorithm

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

A.1.4 Fixed-Point Algorithm with Negentropy

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

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

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

Algorithm 8 Fixed-Point Algorithm with Negentropy (FastICA)

1. Center the observed data to achieve zero mean
2. Whiten the centered data
3. Create the initial random vector $\mathbf{w}$
4. Update
$$\mathbf{w} = \mathbb{E}[\mathbf{y}_{\text{white}} g(\mathbf{w}^T \mathbf{y}_{\text{white}})] - \mathbb{E}[g'(\mathbf{w}^T \mathbf{y}_{\text{white}})]\mathbf{w}$$
5. Normalise $\mathbf{w}$
$$\mathbf{w} = \frac{\mathbf{w}}{\|\mathbf{w}\|}$$
6. Repeat from 4. until convergence
7. Independent components are found as $\mathbf{x} = \mathbf{w}\mathbf{y}_{\text{white}}$