WORKSHEETS

Trine Nyholm Kragh & Laura Nyrup Mogensen
Mathematical Engineering, MATTEK

Master's Thesis

Chapter 1

Sparse Signal Recovery

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

1.1 Linear Algebra

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

$$\mathbf{y} = \mathbf{A}\mathbf{x}, \tag{1.1}$$

where $\mathbf{y} \in \mathbb{R}^M$ is the observed measurement vector consisting of M measurements, $\mathbf{x} \in \mathbb{R}^N$ is an unknown 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 makes a system of linear equations with M equations and N unknowns.

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

By inverting $\mathbf{A}$ from (1.1) the unknown 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.

For an determined system there will exist a unique solution. For an over-determined system there do not exist a solution and for under-determined systems there exist infinitely many solutions [3, p. ix].

As described in chapter **??** the linear model of interest consist of M sensors which makes the observed measurements $\mathbf{y}$ and N sources which makes the unknown vector $\mathbf{x}$. Here it is of interest to find a solution to the case where the system consist of more sources than sensors – hence a solution has to be found within the infinitely solution set.

1.2 Compressive Sensing

Compressive sensing is the theory of efficient recovery or reconstruction 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 described by the linear model (1.1) [5].
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 (1.1). 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 [5].

1.2.1 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 $[N]$ is a set of integers $\{1, 2, \cdots, N\}$ [5, p. 41]. The set of all k-sparse signals is denoted as

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

1.2.2 Optimisation Problem

To find a solution to the linear model (1.1) assuming the solution is k-sparse, it can be view as an optimisation problem. An optimisation problem is defined as

$$\min f_0(\mathbf{x}) \quad \text{subject to} \quad f_i(\mathbf{x}) \leq b_i, \quad i = 1, 2, \cdots, n,$$

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 the optimisation problem can be written as

$$\mathbf{x}^* = \arg\min_{\mathbf{x} \in \mathbf{\Omega}_k} \|\mathbf{x}\|_0 \quad \text{subject to} \quad \mathbf{y} = \mathbf{A}\mathbf{x},$$

The objective function is given by an ℓ_0 norm with the constraint function being the linear model (1.1). 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 be approximated and hence become computational feasible [3, p. 27]

3.2: skal vi indfør z som en approximation til x. og så et nyt omega eller? eller kan vi beholde x

$$\mathbf{x}^* = \arg\min_{\mathbf{x} \in \mathbf{\Omega}_k} \|\mathbf{x}\|_1 \quad \text{subject to} \quad \mathbf{y} = \mathbf{A}\mathbf{x}. \tag{1.2}$$

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 [5, p. 62-63].

Theorem 1.2.1
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 need 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. So when $M \ll N$ a sparse signal is automatically achieved.

Proof
For the case of contradiction, assume that the set $\{\mathbf{a}_j, j \in S\}$ is linearly dependent. Thus there exists a non-zero vector $\mathbf{v} \in \mathbb{R}^N$ supported on S such that $\mathbf{Av} = \mathbf{0}$. Then, for any $t \neq 0$,

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

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

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

It then follows that

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

This is a contradiction, because one can always choose a small $t \neq 0$ such that

$$t \sum_{j \in S} \operatorname{sgn}(x_j^*)v_j \leq 0,$$

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 (1.2) where noise is incorporated. An alternative solution method includes greedy algorithms such as the Orthogonal Matching Pursuit(OMP) [5, P. 65]. At each iteration of the OMP algorithm an index set S is updated by adding the index corresponding to the 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$\{\mathbf{a}_j \mid j \in S\}$. The algorithm for OMP can be seen in 1 where $\mathbf{A}^*$ be the adjoint of a matrix $\mathbf{A}$.

Kan vi benytte $\mathbb{C}$ i algoritmen

Algorithm 1 Orthogonal Matching Pursuit (OMP)

$k = 0$
Initialize $S_{(0)} = \varnothing$
Initialize $x_{(0)} = \mathbf{0}$
procedure OMP($\mathbf{A}, \mathbf{y}$)
    while stopping criteria not meet **do**
        $k = k + 1$
        $j_{(k)} = \arg\max_{j \in [N]} \{|(\mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}_{(k-1)}))_j|\}$
        $S_{(k)} = S_{(k-1)} \cup \{j_{(k)}\}$
        $\mathbf{x}_{(k)} = \arg\min_{\mathbf{z} \in \mathbb{C}^N} \{\|\mathbf{y} - \mathbf{Az}\|_2 \mid \operatorname{supp}(\mathbf{z}) \subset S_{(k)}\}$
    end while
    $\mathbf{x}^* = \mathbf{x}_{(k)}$
end procedure

1.2.3 Conditions on the Mixing Matrix

In section 1.2.2 the mixing matrix $\mathbf{A}$ was assumed known, in order to solve the optimisation problem (1.2). However, in practise it is only the measurement vector $\mathbf{y}$ which is known. In this case the mixing matrix $\mathbf{A}$ is considered a estimate of the true mixing matrix. In general compressive sensing terminology $\mathbf{A}$ is also referred to as a dictionary matrix.
To ensure exact or approximately reconstruction of the sparse signal $\mathbf{x}$, the mixing matrix must be constructed with certain conditions in mind.

Null Space Condition

The the null space property is a necessary and sufficient condition to $\mathbf{A}$ for exact reconstruction of every sparse signal $\mathbf{x}$ that solves the optimisation problem (1.2)[5, p. 77]. The null space of the matrix $\mathbf{A}$ is defined as

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

The null space property is defined as

Definition 1.1 (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

$$\|\mathbf{v}_S\|_1 < \|\mathbf{v}_{\overline{S}}\|_1 \quad \text{for all} \quad \mathbf{v} \in \text{null}(\mathbf{A} \setminus \{\mathbf{0}\}), \tag{1.3}$$

where the vector $\mathbf{v}_S$ is the restriction of $\mathbf{v}$ to the indices in S, and $\bar{S}$ is the set $[N] \setminus S$.

Theorem 1.2.2
Given a matrix $\mathbf{A} \in \mathbb{R}^{M\times N}$, every vector $\mathbf{x} \in \mathbb{R}^N$ with $\text{supp}(\mathbf{x}) \subset S$ is the unique solution of (1.2) with $\mathbf{y} = \mathbf{A}\mathbf{x}$ if and only if $\mathbf{A}$ satisfies the null space property relative to S.

Proof
$\Rightarrow$:
Let $S \subseteq [N]$ be a fixed index set. Assume that any vector $\mathbf{x} \in \mathbb{R}^N$ with support $\text{supp}(\mathbf{x}) \subset S$ is the unique minimizer of $\|\mathbf{z}\|_1$ with respect to $\mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x}$. Thus, for any vector $\mathbf{v} \in \text{null}(\mathbf{A}) \setminus \{\mathbf{0}\}$, the vector $\mathbf{v}_S$ is the unique minimizer of $\|\mathbf{z}\|_1$ with respect to $\mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{v}_S$. But

$$\mathbf{0} = \mathbf{A}(\mathbf{v}_S + \mathbf{v}_{\overline{S}}) \implies \mathbf{A}\mathbf{v}_S = \mathbf{A}(-\mathbf{v}_{\overline{S}}), \quad \text{with} \ -\mathbf{v}_{\overline{S}} \neq \mathbf{v}_S,$$

or else $\mathbf{v} = \mathbf{0}$. It is concluded that $\|\mathbf{v}_S\|_1 < \|\mathbf{v}_{\overline{S}}\|_1$. This establishes the null space property relative to S.

$\Leftarrow$:
Conversely, assume that the null space property relative to S holds. Given an index set $S \subseteq [N]$ with null space property and a vector $\mathbf{x} \in \mathbb{R}^N$ with $\text{supp}(\mathbf{x}) \subset S$. Furthermore, given a vector $\mathbf{z} \in \mathbb{R}^N$ where $\mathbf{z} \neq \mathbf{x}$, such that $\mathbf{Az} = \mathbf{Ax}$. Consider then a vector $\mathbf{v}$ given by $\mathbf{v} := \mathbf{x} - \mathbf{z} \in \text{null}(\mathbf{A}) \setminus \{\mathbf{0}\}$. From the null space property, the following is obtained:

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

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

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

Coherence

The null space property provide a unique solution to the optimisation problem (1.2), but it is unfortunately complicated to investigate. Instead an alternative measure is presented.
Coherence is a measure of quality, it determines whether a matrix $\mathbf{A}$ is a good choice for the optimisation problem (1.2). A small coherence describes the performance of a recovery algorithm as good with that choice of $\mathbf{A}$.

Definition 1.2 (Coherence)
Coherence of the matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$, denoted as $\mu(\mathbf{A})$, with columns $\mathbf{a}_1, \ldots, \mathbf{a}_N$ for all $i \in [N]$ is given as

$$\mu(\mathbf{A}) = \max_{1 \leq i < j \leq n} \frac{|\langle \mathbf{a}_i, \mathbf{a}_j \rangle|}{\|\mathbf{a}_i\|_2 \|\mathbf{a}_j\|_2}.$$

Restricted Isometry Condition

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

Definition 1.3 (Restricted Isometry Property (RIP))
A matrix $\mathbf{A}$ satisfies the RIP of order k if there exists a $\delta_k \in (0,1)$ such that

$$(1-\delta_k)\|\mathbf{x}\|_2^2 \le \|\mathbf{A}\mathbf{x}\|_2^2 \le (1+\delta_k)\|\mathbf{x}\|_2^2,$$

Theorem 1.2.3
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

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

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

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

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

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

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

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

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

$$\begin{aligned}\|\mathbf{v}_{S_0}\|_2^2 &\le \frac{1}{1-\delta_{2s}}\|\mathbf{A}(\mathbf{v}_{S_0})\|_2^2 = \frac{1}{1-\delta_{2s}}\langle \mathbf{A}(\mathbf{v}_{S_0}), \mathbf{A}(-\mathbf{v}_{S_1}) + \mathbf{A}(-\mathbf{v}_{S_2}) + \cdots\rangle \\ &= \frac{1}{1-\delta_{2s}}\sum_{k\ge 1}\langle \mathbf{A}(\mathbf{v}_{S_0}), \mathbf{A}(-\mathbf{v}_{S_k})\rangle. \end{aligned} \tag{1.4}$$

According to Proposition 6.3 [5, p. 135], one also have

$$\langle \mathbf{A}(\mathbf{v}_{S_0}), \mathbf{A}(-\mathbf{v}_{S_k}) \rangle . \leq \delta_{2s} \|\mathbf{v}_{S_0}\|_2 \|\mathbf{v}_{S_k}\|_2. \tag{1.5}$$

Substituting (1.5) into (1.4) and dividing by $\|\mathbf{v}_{S_0}\|_2 > 0$ ■

1.2.4 Multiple Measurement Vector Model

The linear model (1.1) is also referred to as a single measurement vector (SMV) model. In order to adapt the model (1.1) 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 $\mathbf{E} \in \mathbb{R}^{M \times L}$:

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{E}. \tag{1.6}$$

L denote the number of observed measurement vectors each consisting of M measurements, that is L samples is given. For $L = 1$ the linear model will just be the SMV model (1.1).
The matrix $\mathbf{X}$ consist of $\{\mathbf{x}_i\}_{i=1}^{L}$ k-sparse vectors which has been stacked column-wise such that $\mathbf{X}$ consist of at most k non-zero rows. As for the SMV model (1.1) the MMV model (1.6) is under-determined with $M \ll N$ and $k < M$[3, p. 42].

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 for all columns. 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 $\mathbf{A}$ which is the minimum set of linearly dependent columns, it is possible to set some conditions on the system to ensure recovery.
When $|\text{supp}(\mathbf{X})| = k$ then $\text{rank}(\mathbf{X}) \leq k$. If $\text{rank}(\mathbf{X}) = 1$ then are the k-sparse vectors $\{\mathbf{x}_i\}_{i=1}^{L}$ multiples of each other and the joint information can not be taken advantage of. But for large rank it is possible to exploit the diversity of the columns in $\mathbf{X}$. This can be defined as a sufficient and necessary condition of the MMV model (1.6). MMV system $\mathbf{Y} = \mathbf{A}\mathbf{X}$ must have

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

such that $\mathbf{X}$ can uniquely be determined.
This result says that a row-sparse matrix $\mathbf{X}$ with large rank can be recovered from fewer measurement [3, p. 43].

Skal "fewer" udspecificeres yderligere?

1.3 Dictionary learning

As clarified in section 1.2.3 the estimation of 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 $\mathbf{x}$[4]. 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[4]. 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 $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_L]$, $\mathbf{y}_i \in \mathbb{R}^M$ 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 $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_L]$, $\mathbf{x}_i \in \mathbb{R}^N$. For a known sparsity constraint k this can be defined by an optimisation problem similar to the general compressive sensing problem of multiple measurements [4]

$$\min_{\mathbf{A},\mathbf{X}} \sum_{i=1}^{L} \|\mathbf{y}_i - \mathbf{A}\mathbf{x}_i\|_2^2 \quad st.\ \|\mathbf{x}_i\|_0 \le k,\ 1 \le i \le L. \tag{1.7}$$

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 $k < M$[2]. 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[4]. Again the ℓ_0-norm lead to an NP-hard problem an heuristic methods are need.

fungerer disse to uniqueness parameter sammen?

1.3.1 K-SVD

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 (1.7) using matrix notation it is possible to isolate

the contribution from $\mathbf{a}_{i_0}$.

$$\begin{aligned}\|\mathbf{Y}-\mathbf{A}\mathbf{X}\|_F^2 &= \left\|\mathbf{Y}-\sum_{i=1}^{M}\mathbf{a}_i\mathbf{x}_i^T\right\|_F^2 \\ &= \left\|\left(\mathbf{Y}-\sum_{i\neq i_i}^{M}\mathbf{a}_i\mathbf{x}_i^T\right)-\mathbf{a}_{i_0}\mathbf{x}_{i_0}^T\right\|_F^2,\end{aligned} \tag{1.8}$$

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 (1.8) the term in the parenthesis makes the an error matrix $\mathbf{E}_{i_0}$ without the contribution from i_0, hence minimising (1.8) 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..?).

$$\min_{\mathbf{a}_{i_0},\mathbf{x}_{i_0}^T}\left\|\ \mathbf{E}_{i_0}-\mathbf{a}_{i_0},\mathbf{x}_{i_0}^T\right\|_F^2 \tag{1.9}$$

The optimal solution to (1.9) 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..**]

find kilde

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

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 (1.9). 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

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 $\mathbf{A}$.

1.4 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,

$$\mathbf{y} = \mathbf{A}\mathbf{x} \tag{1.10}$$

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[6, 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.

1.4.1 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 $\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}$[6, 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 scala[6, p. 154].
For further simplification it is assumed without loss of generality that $\mathbb{E}[\mathbf{y}] = 0$ and $\mathbb{E}[\mathbf{x}] = 0$[6, p. 154]. In case this assumption is not true, the measurements can be centred by subtracting the mean as preprocessing before doing ICA.
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 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 $\mathbf{V}$,

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

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

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 [6, 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 [6, p. 159] .

se udkommentering herunder?

1.4.2 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 $\mathbf{y} = \mathbf{Ax}$ 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

$$x_j = \mathbf{b}_j^T \mathbf{y} = \sum_i b_i y_i \tag{1.11}$$

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 (1.11)

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

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[6, 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[6, p. 166].

1.4.3 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 [6, 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 [6, 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.

hvordan kommer dette frem?

Due to the described preprocessing $\mathbf{b}$ is assumed to be white and it can be shown that $\|\mathbf{q}\| = \|\mathbf{b}_j\|$[6, p. 174]. This show that constraining $\|\mathbf{q}\|$ to one is similar to constraining $\|\mathbf{b}_j\|$ to one.

1.4.4 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

$$\frac{\partial|\text{kurt}(\mathbf{b}_j^T\mathbf{y})|}{\partial\mathbf{b}_j} = 4\text{sign}(\text{kurt}(\mathbf{b}_j^T\mathbf{y}))(\mathbb{E}[\mathbf{y}(\mathbf{b}_j^T\mathbf{y})^3] - 3\mathbf{y}\mathbb{E}[(\mathbf{b}_j^T\mathbf{y})^2]) \tag{1.12}$$

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 and serving as the learning rate of the gradient method.

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

1.4.5 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

1.4.6 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. Extension to the basic ICA algorithm can be found in appendix 2.

1.5 Limitations of compressive sensing

Through this chapter the concept of sparse signal recovery have been explained. It is seen that to recover the a sparse signal from the
The essential limitation of signal recovery from an under-determined system is that $k \leq M$ is necessary in order to uniquely recover the k-sparse signal $\mathbf{X} \in \mathbb{R}^N$ from the measurements $\mathbf{Y} \in \mathbb{R}^M$. That is the number of measurements must be greater than the number of active sources within the signal to be recovered. Similarly it is not possible to recover the true dictionary $\mathbf{A}$ by dictionary learning methods if $k > M$. Because in that case any random dictionary of full rank can be used to create $\mathbf{y}$ from $\geq M$ basis vectors[2, p. 30].
When considering source recovery from EEG measurements, described in section **??**, 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 2

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.

2.1 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 1.4.4. 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 point of 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 (1.12) be equal to $\mathbf{w}$ and considering the same simplifications again suggests the new update step as [6, 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 [6, p. 179].

2.1.1 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 [6, 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 $\mathbf{y}$ [6, 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 [6, p. 182].

2.1.2 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 [6, 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 [6, 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

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

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

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

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

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