Universität Hamburg
Department Mathematik
Bundesstraße 55
20146 Hamburg

Diplomarbeit

Dimensionality Reduction Methods in Independent Subspace Analysis for Signal Detection

Sara Krause-Solberg

August 2011

Betreuer: Prof. Dr. Armin Iske

Contents

Introduction

Signals are present in many different areas of our everyday life. They are used for communication and entertainment, in engineering and medicine, for traffic control, space exploration and data compression. In all these cases signals are used to transmit information. Due to further development during the last decades, in many fields, as for example in multimedia entertainment and information systems, signals have gained even more importance. As a consequence, a wealth of signals is created and thus it might come to a superposition or mixture of signals. In other situations, the information contained in a signal might be encoded such that it is not readily available. Thus, the ability to extract information from a signal has become more and more essential for handling the huge amount of signals.

The extraction of metadata from a signal is used in many applications as for example weather forecasts, where the relevant information needs to be selected from meteorological data and satellite images, or robot control, where a matching of visual, audio and other stimulations is demanded. Most applications, however, refer to audio data, as for example, acoustic echo cancellation and denoising, automatic transcription of music, application of audio effects to single instruments in a mixed recording, speaker separation in video conferences, emotion recognition from speech signals or hearing aids, which are able to accentuate different sources. In all these situations an efficient method to analyze the auditory scene in order to extract the essential information is needed.

In many cases of auditory scene analysis, humans possess the ability of suppressing ambient noises and disturbance sources and to focus on a certain source within a mixture of multiple sound sources. This phenomenon is known as 'cocktail party effect'. Many researchers have focused on the techniques, which humans use to isolate single sources. These techniques are, for example, based on spatial distances between the sources, differences in pitch and quality or visual indicators such as lip reading [50]. Nevertheless, the current state of scientific and technical knowledge is far from attaining similar results to the human auditory system.

In the last decades, some relatively successful separation algorithms appeared, and thus investigation on this topic has been intensified (see [1], [2], [10], [43], [50], [51], [52]). One approach to solve technically the problem of extracting single sources from a mixed signal is Blind Signal Separation. It relies on no assumptions concerning the position of sensors or sources in contrast to geometrical source separation by means of beamforming (e.g. [2]) or similar methods.

Blind Signal Separation (BSS) is a technique which recovers a set of unknown source signals from a set of mixed signals or other observations. The set of observations is usually obtained by a set of sensors recording, each a different combination of the source signals, depending on the position of the sensor. In this context, 'blind' stands for

the fact that the sources themselves are not observed and that there is no information available about the mixing process, i.e. the estimation is performed without almost any knowledge about the sources, as for example location or activity time. This 'blindness' is not a negative property, in contrast, it is precisely the strength of BSS models making them flexible [7].

The core of all BSS methods is the assumption that the observations are a weighted sum of the unknown sources (for non-linear mixing models see for instance [47]). This assumption involves the restriction that there are at least as many observations as sources. But in many applications there is only one sensor recording the mixed signal. This situation is called single-channel problem, and there is a strong demand for methods applicable to single-channel mixtures. To avoid the problem of having less observations as source signals, usually the classical BSS methods are combined with a preprocessing step involving time-frequency analysis in order to construct a set of observations (e.g. a spectrogram).

The different BSS methods can be classified in those operating in the time-amplitude or in the time-frequency domain. But all of them are statistical methods based on the minimization of a certain cost functional. This cost functional might vary from method to method according to the characteristics of the source signals. Since the methods are statistical ones, the cost functionals are based on second or higher order statistics. The second order statistics methods optimize the uncorrelatedness of the sources while the higher order methods involve also moments of higher order and thus optimize the statistical independence of the source signals [16]. This is what Independent Component Analysis (ICA) attempts to do by searching non-Gaussian source signals. A measure of the non-Gaussianity of a signal is for example the negentropy [13].

Recently, Independent Component Analysis (ICA) has become a favourite method in the field of signal separation. Many methods based on ICA (e.g. [1] or [10]) achieve good results for stationary sources and sources with steady-state components. But musical tones and sounds are characterized by their transient effects since musical transients hold much of the perceptual information within a tone [54]. Therefore, in the last years scientists have concentrated on the extraction of transitory acoustic sounds (e.g. [17] or [50]). But since the duration in time of this kind of source signals is very short, it is difficult to separate them from other sources. However, if an algorithm would provide any information about the location in time where such a signal is active, separation would become much easier, since the separation algorithm would have to focus only on those regions. Nevertheless, in many cases already the detection of sources is quite a challenging task, especially if a transitory signal, which has a wide frequency band, is involved.

The objective of this work is to evaluate the usage of dimensionality reduction methods in signal detection and separation algorithms. Recent developments on this subject are presented in [18] and [50], but further investigations on the signal processing and mathematical framework of these algorithms are essentially required [24]. Therefore, we analyze the application of dimensionality reduction methods in the context of Independent Subspace Analysis (ISA). In particular, we focus on the detection of

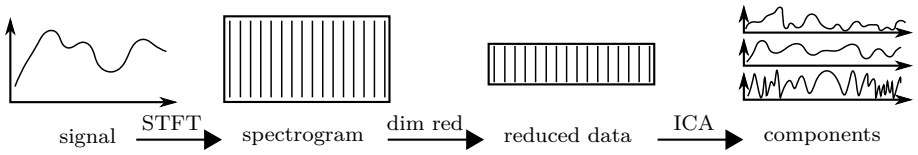


Figure 1.: General proceeding of ISA with dimensionality reduction.

sources in a mixture of transitory signals.

This detection is done by a combination of time-frequency analysis and Independent Component Analysis called ISA. In a preprocessing step, time-frequency analysis is used to obtain a data set from a single-channel signal. This data usually has a high dimensionality which justifies the usage of dimensionality reduction methods in order to process the data adequately and to make the existing algorithms more efficient. This procedure can be improved by analyzing the mathematical background and by running empirical tests. Therefore, the aim of this work is to illustrate how dimensionality reduction can be applied in the field of single-channel problems and especially in combination with ISA. For this purpose we introduce two dimensionality reduction methods, namely Principle Component Analysis (PCA) and Laplacian Eigenmaps (LE) in order to compare how they interact with ISA. The general proceeding is illustrated in Figure 1.

Dimensionality reduction is the transformation of a high-dimensional data set, which lies on a manifold, into a low-dimensional representation of this manifold. In the optimal situation, the dimensionality of the low-dimensional representation corresponds to the intrinsic dimensionality of the data set. In this context, the intrinsic dimensionality is the smallest number of features needed for characterizing the data. Recently, many non-linear dimensionality reduction techniques have been proposed (for an overview see [35] or [39]). In contrast to classical linear techniques such as PCA, the non-linear techniques are able to handle complex non-linear data as for instance the 'Swiss roll', i.e. a set of points that lie on a spiral-like two-dimensional manifold that is embedded in a three-dimensional space.

The purpose of this work is to provide an insight into the underlying concepts and to perform comparative experimental tests of the algorithm. The work is divided into two parts. In Part I the basic methods and concepts are introduced in order to detail the mathematical theory behind the detection algorithm in Figure 1. Chapter 1 deals with time-frequency analysis with emphasis on the short-time Fourier (STFT). Chapter 2 discusses two dimensionality reduction techniques, namely PCA and LE. ICA is concerned in Chapter 3 and ISA, the core concept, in Chapter 4.

Part II is involved with a case study. Chapter 5 overviews the method. In Chapter 6 the different phases of the algorithm are visualized by means of an example and in Chapter 7 the signal separation in the case of PCA is discussed. The work ends with a conclusion in Chapter 8 and a compact disc containing a pdf-version of this work and a MATLAB implementation of the algorithm.

Acknowledgement

This thesis would not have been possible without the support of my advisor, my family and my friends. In particular, I would like to thank Mijail Guillemard for his encouragement and help during the preparation of this work.

Part I.

Theory

1. Time-Frequency Analysis

Time-frequency analysis of signals refers to mathematical transforms of continuous or discrete data and to characterization as well as manipulation of signals whose frequency components might vary in time. This kind of analysis is performed in order to obtain more information about a signal, and with the prospect that also the image of a signal under a certain transform is more easily interpretable and analyzable than the original signal. The benefit of these ideas depends strongly on the choice of the transform and thus on an adequate mathematical model of the signal. Such a method for analysis can be based on a family of functions and performed by using a series expansion or an integral transform. Thus, there are many possible transforms and the challenge is to select an appropriate one. Therefore, we would like the transform to fulfill some elementary conditions such as the continuity of the transform mapping or the conservation of information, i.e. no information should be cut off, lost or hidden as a result of the transform. In [20] these mathematical requirements are formulated as follows:

- The transform should be continuous: Quantitatively small changes in the signal should cause only quantitatively small effects in the transform's image.
- The transform should be continuously invertible.
- There should exist an invertible discrete version of the transform.
- There should exist a stable numerical algorithm.

One of the first analysing systems is the well-known Fourier series, developed in the early 19th century by Jean Baptiste Joseph Fourier [21]. Primarily, Fourier has worked on heat conduction in different solids and proposed an expansion of the initial condition for the temperature in a series of sine terms. Nowadays, what we call the Fourier series of a function $f \in L^p[-\pi, \pi], 1 \leq p < \infty$ is its expansion in series of complex exponentials given by

$$f(x) = \sum_{k \in \mathbb{Z}} \hat{f}(k) e^{ikx},$$

with the Fourier coefficients

$$\hat{f}(k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) e^{-ikt} \mathrm{d}t. \tag{1.1}$$

Since the absolute value of the coefficients $|\hat{f}(k)|$ can be interpreted as the amplitude and the argument $\arg(\hat{f}(k))$ as the phase corresponding to the frequency $k \in \mathbb{Z}$, Fourier

series provide a useful tool for the analysis of periodic functions. But as the coefficients $\hat{f}(k)$ in (1.1) depend strongly on f, each minimal local change of f will cause a global change of the coefficients $\hat{f}(k)$. Such behaviour might cause a dramatically increasing computation time, in particular in the case of integral transforms. For this reason, localized transforms based on families of compactly supported functions are attracting more and more interest. Among these are short-time Fourier analysis and wavelet multiresolution analysis (see [15]). In this work we introduce the short-time Fourier transform on $L^2(\mathbb{R})$ and discuss some of its properties. More information can be found for example in [48], [53] or [55]. In Part II we will use the short-time Fourier transform in the context of signal detection and separation, but any other transform can be used as well (see [10]).

1.1. Fourier Transform

For a better understanding of the short-time Fourier transform we shall start with the Fourier transform. The Fourier transform is an integral transform which is, in some sense, a generalization of the Fourier series as it is defined for all integrable functions $f \in L^1(\mathbb{R})$. In particular, f does not necessarily need to be periodic.

Definition 1.1. For a function $f \in L^1(\mathbb{R})$ its *Fourier transform* $\mathcal{F}f$ is defined by

$$\mathcal{F}f(\omega) = \int_{\mathbb{R}} f(x) e^{-i\omega x} \mathrm{d}x,$$

for all $\omega \in \mathbb{R}$.

Definition 1.2. For a function $g \in L^1(\mathbb{R})$ its *inverse Fourier transform* $\mathcal{F}^{-1}g$ is defined by

$$\mathcal{F}^{-1}g(t) = \frac{1}{2\pi} \int_{\mathbb{R}} g(\omega) e^{i\omega t} \mathrm{d}\omega,$$

for all $t \in \mathbb{R}$.

Theorem 1.1 ([20])**.** *For a function $f \in L^1(\mathbb{R})$ satisfying $\mathcal{F}f \in L^1(\mathbb{R})$ the Fourier inversion formula*

$$f(t) = \frac{1}{2\pi} \int_{\mathbb{R}} \mathcal{F}f(\omega) e^{-i\omega t} \mathrm{d}\omega$$

holds for almost all $t \in \mathbb{R}$ with equality at the points of continuity of f.

Proof. See [20]. □

Remark 1. *The Fourier transform of a signal f is its frequency spectrum. The frequency spectrum provides information about the frequencies which are present in the signal. Usually the Fourier transform is complex-valued and thus the frequency spectrum can be decomposed in amplitude spectrum $|\mathcal{F}f|$ and phase spectrum* $\arg(\mathcal{F}f)$.

To obtain numerical solutions of complex mathematical problems discretization is necessary. But we have to take into account, that the discrete solution should converge to the solution of the original problem and that its computation can be done efficiently. To begin with, we concentrate on the convergence. Later on, we shall introduce the FFT algorithm which is computationally efficient. The Nyquist-Shannon Sampling Theorem is a fundamental result in signal processing which has been found in the early 20th century. It gives a lower bound on the number of sampling points such that a continuous signal can be reconstructed exactly from a discrete set of samples.

Theorem 1.2 (Nyquist-Shannon Sampling Theorem)**.** *Let $f \in L^1(\mathbb{R}) \cap C(\mathbb{R})$ be band-limited to $[-\pi\delta, \pi\delta]$ for $\delta > 0$, i.e.*

$$f(t) = \frac{1}{2\pi} \int_{-\pi\delta}^{\pi\delta} \mathcal{F}f(\omega) e^{i\omega t} \mathrm{d}\omega,$$

and $\mathcal{F}f \in L^1(\mathbb{R})$. Then, for every $t \in \mathbb{R}$, f can be reconstructed from its sampled values at the points $t_k = \frac{k}{\delta}$, $k \in \mathbb{Z}$, via the formula

$$f(t) = \sum_{k=-\infty}^{\infty} f(t_k) \frac{\sin(\pi\delta(t - t_k))}{\pi\delta(t - t_k)} = \sum_{k=-\infty}^{\infty} f(t_k) \operatorname{sinc}(\delta t - k).$$

Moreover, the series converges uniformly and absolutely on $\mathbb{R}$.

Proof. See [33]. □

Remark 2. *The length of the support of the Fourier transform of f, i.e. $2\pi\delta$ is called bandwidth of f. The sampling frequency δ is known as the Nyquist rate, which is the minimum rate at which the function f needs to be sampled in order to be exactly reconstructible.*

As there is a need of discrete transforms, we now introduce the discrete version of the Fourier transform. In the following, $\mathbb{Z}_n$ denotes the set $\{0, \dots, N-1\} \subset \mathbb{N}$, thus $\ell^\infty(\mathbb{Z}_N) \simeq \mathbb{R}^N$ holds.

Definition 1.3. For a discrete function $f \in \ell^\infty(\mathbb{Z}_N)$ its *discrete Fourier transform* $\mathcal{F}_D f$ is defined by

$$(\mathcal{F}_D f)_j = \sum_{k=0}^{N-1} f_k e^{-\frac{2\pi i j k}{N}}, \tag{1.2}$$

for $j \in \mathbb{Z}_N$.

Definition 1.4. For a discrete function $g \in \ell^\infty(\mathbb{Z}_N)$ its *discrete inverse Fourier transform* $\mathcal{F}_D^{-1} g$ is defined by

$$(\mathcal{F}_D^{-1} g)_k = \frac{1}{N} \sum_{j=0}^{N-1} g_j e^{\frac{2\pi i j k}{N}},$$

for $k \in \mathbb{Z}_N$.

Theorem 1.3. *For a discrete function $f \in \ell^\infty(\mathbb{Z}_N)$ the discrete Fourier inversion formula*

$$f_k = (\mathcal{F}_D^{-1}\mathcal{F}_D f)_k$$

holds for all $k \in \mathbb{Z}_N$.

Proof. See [53]. □

It is easy to see that a straightforward computation of the discrete Fourier transform is of complexity $\mathcal{O}(N^2)$ as the computation for each of the N components is of complexity $\mathcal{O}(N)$. In order to compute the discrete Fourier transform efficiently, we use the so called *Fast Fourier Transform* (FFT). There are different algorithms to perform the FFT, among them the Cooley–Tukey algorithm proposed in 1965 [14]. This algorithm is the most common FFT algorithm and a powerful tool for a fast computation of the discrete Fourier transform because it reduces the complexity to $\mathcal{O}(N \log_2(N))$. It is based on the factorization of the period length N, i.e. we suppose $N = N_1 N_2$ for some $N_1, N_2 \in \mathbb{Z}$. After the choice of N_1 and N_2 we can express the indices j and k from (1.2) as

$$\begin{aligned} j &= j_1 N_1 + j_0, \qquad j_0 \in \mathbb{Z}_{N_1}, \qquad j_1 \in \mathbb{Z}_{N_2} \\ k &= k_1 N_2 + k_0, \qquad k_0 \in \mathbb{Z}_{N_2}, \qquad k_1 \in \mathbb{Z}_{N_1}. \end{aligned}$$

From this decomposition it follows that

$$e^{-\frac{2\pi i j k_1 N_2}{N}} = e^{-\frac{2\pi i (j_1 N_1 + j_0) k_1 N_2}{N}} = e^{-2\pi i j_1 k_1 \frac{N_1 N_2}{N}} e^{-\frac{2\pi i j_0 k_1 N_2}{N}} = e^{-\frac{2\pi i j_0 k_1 N_2}{N}}.$$

Hence, we can decompose the sum in (1.2) and thus we get

$$\begin{aligned} (\mathcal{F}_D f)_j &= (\mathcal{F}_D f)_{j_1 N_1 + j_0} \\ &= \sum_{k_0=0}^{N_2-1} \sum_{k_1=0}^{N_1-1} f_{k_1 N_2 + k_0} e^{-\frac{2\pi i (j_1 N_1 + j_0)(k_1 N_2 + k_0)}{N}} \\ &= \sum_{k_0=0}^{N_2-1} \sum_{k_1=0}^{N_1-1} f_{k_1 N_2 + k_0} e^{-\frac{2\pi i j_0 k_1 N_2}{N}} e^{-\frac{2\pi i (j_1 N_1 + j_0) k_0}{N}} \\ &= \sum_{k_0=0}^{N_2-1} e^{-\frac{2\pi i (j_1 N_1 + j_0) k_0}{N}} \sum_{k_1=0}^{N_1-1} f_{k_1 N_2 + k_0} e^{-\frac{2\pi i j_0 k_1 N_2}{N}} \\ &= \sum_{k_0=0}^{N_2-1} \tilde{f}_{j_0, k_0} e^{-\frac{2\pi i (j_1 N_1 + j_0) k_0}{N}}, \end{aligned}$$

where $\tilde{f}_{j_0,k_0} = \sum_{k_1=0}^{N_1-1} f_{k_1 N_2 + k_0} e^{-\frac{2\pi i j_0 k_1 N_2}{N}}$. Since $\tilde{f}$ has $N_1 N_2 = N$ elements, we need $\mathcal{O}(N N_1)$ operations to compute $\tilde{f}$. For similar reasons, it takes $\mathcal{O}(N N_2)$ operations to obtain $\mathcal{F}_D f$ from $\tilde{f}$. Thus, the composed algorithm requires $\mathcal{O}(N(N_1 + N_2))$ operations

to calculate $\mathcal{F}_D f$ from f. A factorization of N into more than two integers allows a successive application of this procedure resulting in $\mathcal{O}(N(n_1+n_2+\cdots+n_m))$ operations for $N = n_1 n_2 \dots n_m$. As a consequence, for $N = 2^m$ we obtain $m = \log_2(N)$ which leads to a complexity of $\mathcal{O}(N \log_2(N))$. Therefore, it is reasonable to take 2^m, $m \in \mathbb{Z}$, samples of a given continuous signal.

In a more global context, we can consider the Fourier transform as an operator $\mathcal{F} : L^2(\mathbb{R}^n) \to L^2(\mathbb{R}^n)$. The operator $\mathcal{F}$ is linear and continuous as we can see from the following theorem.

Theorem 1.4 (The Rayleigh-Plancherel theorem)**.** *Let $f \in L^1(\mathbb{R})$. If either f or its Fourier transform is square integrable over the real line, i.e. $f \in L^2(\mathbb{R})$ or $\mathcal{F}f \in L^2(\mathbb{R})$, then we have*

$$\|f\|^2_{L^2(\mathbb{R})} = \frac{1}{2\pi}\|\mathcal{F}f\|^2_{L^2(\mathbb{R})}.$$

Proof. See [45]. □

Remark 3. *In the discrete case*

$$\sum_{k=0}^{N-1} |f_k|^2 = \frac{1}{N}\sum_{k=0}^{N-1} |(\mathcal{F}_D f)_k|^2$$

holds for any discrete function $f \in \ell^\infty(\mathbb{Z}_N)$ (see [53]).

1.2. Short-Time Fourier Transform

The Fourier transform uses non-compactly supported functions for the analysis of signals. As we have already stated before, this might cause instability with respect to local manipulation in the time or frequency domain. In order to avoid this phenomenon, we multiply the function f by a *window function* φ and apply the Fourier transform to their product.

Definition 1.5 ([20])**.** Assume that $\varphi \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ and $f \in L^2(\mathbb{R})$. For $\tau \in \mathbb{R}$ and $\omega \in \mathbb{R}$ we define

$$\mathcal{F}_\varphi f(\omega, \tau) = \int_{\mathbb{R}} f(t)\varphi(t-\tau)e^{-i\omega t}\mathrm{d}t.$$

Then $\mathcal{F}_\varphi f$ is called the *short-time Fourier transform* (STFT) of f.

This localization gives us the frequency content of the signal in a concrete window φ with center τ such that the short-time Fourier transform depends on two variables, the frequency ω and the center of localization τ. It can be shown (see [20]) that $f\varphi(\cdot - \tau) \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ and thus the STFT has properties analogue to the properties of the Fourier transform.

In the previous definition we mentioned a window function φ. How does such a function look like? Usually, a window function is a continuous, compactly supported, non-negative and symmetric function. In fact, this definition can be generalized claiming

that the function decreases sufficiently fast to zero away from the origin. The STFT was first used by Gábor in 1946. In [22], Gábor considers the Gaussian window

$$\phi_\sigma(t) = \frac{1}{2\sqrt{\pi\sigma}} e^{-\frac{t^2}{4\sigma}}$$

with $\sigma > 0$. Due to the importance of the STFT in many applications, the STFT using this special window is called *Gabor transform*.

From the huge class of window functions we like to introduce the *Hann window*

$$h(t) = \frac{1}{2}\left(1 + \cos\left(\frac{2\pi t}{T}\right)\right),$$

where T is the window size, i.e. $\operatorname{supp} h \subset [-\frac{T}{2}, \frac{T}{2}]$ (see [5]).

Definition 1.6. For a function $g \in L^1(\mathbb{R}^2)$ the *inverse short-time Fourier transform* $\mathcal{F}_\varphi^{-1} g$ is defined by

$$\mathcal{F}_\varphi^{-1} g(t) = \frac{1}{2\pi c} \int_{\mathbb{R}} \int_{\mathbb{R}} g(\omega, \tau) e^{i\omega t} \mathrm{d}\omega \mathrm{d}\tau,$$

where $c = \int_{\mathbb{R}} \varphi(t) \mathrm{d}t$.

Theorem 1.5. *For a function $f \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ satisfying $\mathcal{F}f \in L^1(\mathbb{R})$, the inversion formula*

$$f(t) = \mathcal{F}_\varphi^{-1} \mathcal{F}_\varphi f(t)$$

holds for all t at which f is continuous.

Proof. A simple computation gives the result:

$$\begin{aligned}
\mathcal{F}_\varphi^{-1} \mathcal{F}_\varphi f(t) &= \frac{1}{2\pi c} \int_{\mathbb{R}} \int_{\mathbb{R}} \mathcal{F}_\varphi f(\omega, \tau) e^{i\omega t} \mathrm{d}\omega \mathrm{d}\tau \\
&= \frac{1}{2\pi c} \int_{\mathbb{R}} \int_{\mathbb{R}} \int_{\mathbb{R}} f(s) \varphi(s - \tau) e^{-i\omega s} \mathrm{d}s e^{i\omega t} \mathrm{d}\omega \mathrm{d}\tau \\
&= \frac{1}{2\pi c} \int_{\mathbb{R}} \varphi(\tau) \mathrm{d}\tau \int_{\mathbb{R}} \int_{\mathbb{R}} f(s) e^{-i\omega s} \mathrm{d}s e^{i\omega t} \mathrm{d}\omega \\
&= \frac{1}{2\pi} \int_{\mathbb{R}} \mathcal{F}f(\omega) e^{i\omega t} \mathrm{d}\omega \\
&= \mathcal{F}^{-1} \mathcal{F} f(t) \\
&= f(t).
\end{aligned}$$

□

Like in the case of the continuous Fourier transform, we introduce a discrete version of the STFT (see [44]). As shown in Figure 1.1, we consider a segmentation of the

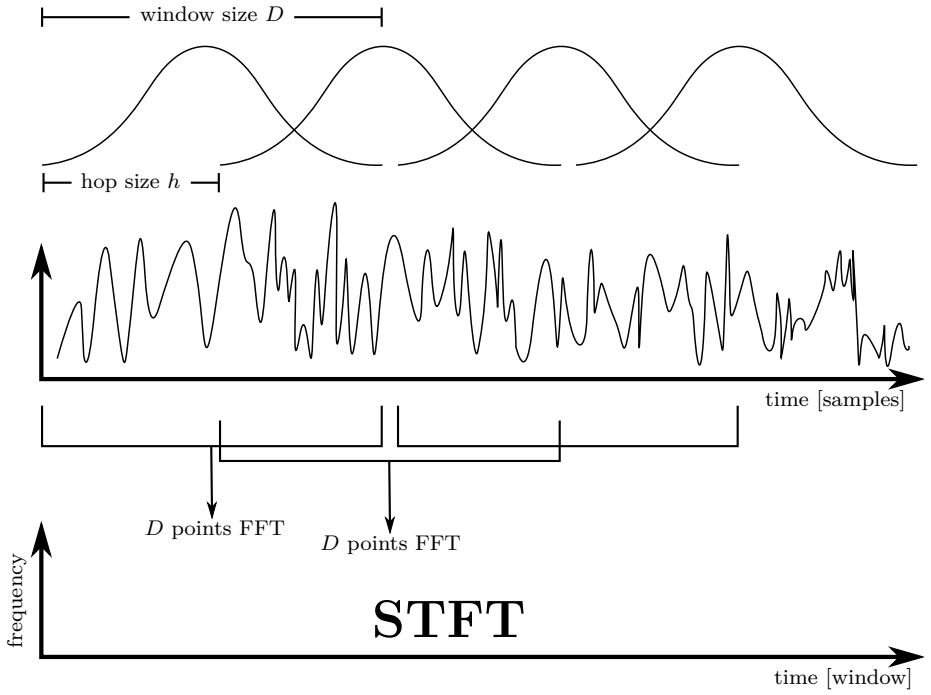


Figure 1.1.: Short-time Fourier transform and construction of spectrogram.

signal into small segments of length D at distance h. This segmentation is obtained by multiplication of the signal by a discrete, compactly supported window of length D with center lh. Subsequently, the FFT algorithm is applied to the segments in order to compute a discrete spectrogram.

Definition 1.7. Assume that $\varphi \in \ell^\infty(\mathbb{Z}_D)$ is a discrete window with $\varphi_k \neq 0$ and $f \in \ell^\infty(\mathbb{Z}_N)$. For n and $h \in \mathbb{N}$ with $(n-1)h = N - 1 - D$, we define the *discrete short-time Fourier transform* $\mathcal{F}_{\varphi,D} f$ of f by

$$\left(\mathcal{F}_{\varphi,D} f\right)_{j,l} = \sum_{k=0}^{D-1} f_{k+lh} \varphi_k e^{-\frac{2\pi i j k}{D}} = \left(\mathcal{F}_D \left(f_{k+lh}\varphi_k\right)_{k=0}^{D-1}\right)_j,$$

for $j \in \mathbb{Z}_D$ and $l \in \mathbb{Z}_n$. The parameter h is called *hop size* and D is the *window length*.

Definition 1.8. For φ, h and n as in Definition 1.7 with $h \leq D$ and $g \in \ell^\infty(\mathbb{Z}_D \times \mathbb{Z}_n)$ the *discrete inverse short-time Fourier transform* is defined by

$$\left(\mathcal{F}_{\varphi,D}^{-1} g\right)_k = \frac{1}{c_k} \sum_{(j,l)\in\mathbb{Z}_D\times\mathbb{Z}_n : j+lh=k} \left(\mathcal{F}_D^{-1}\left(g_{i,l}\right)_{i=0}^{D-1}\right)_j,$$

for $k \in \mathbb{Z}_N$, where

$$c_k = \sum_{(j,l)\in\mathbb{Z}_D\times\mathbb{Z}_n : j+lh=k} \varphi_j.$$

Remark 4. *The sum in Definition 1.8 is not empty if $h \leq D$. This follows from the decomposition of k by Euclidean division by h. This seems reasonable since otherwise the hop size would be larger than the window size and application of the discrete STFT would cause the loss of parts of the function f.*

Theorem 1.6. *For a function $f \in \ell^\infty(\mathbb{Z}_N)$ and φ, h and n as in Definition 1.7 with $h \leq D$ the inversion formula*

$$f_k = \left(\mathcal{F}_{\varphi,D}^{-1}\mathcal{F}_{\varphi,D}f\right)_k$$

holds for all $k \in \mathbb{Z}_N$.

Proof. Computation leads to

$$\begin{aligned}
\left(\mathcal{F}_{\varphi,D}^{-1}\mathcal{F}_{\varphi,D}f\right)_k &= \frac{1}{c_k} \sum_{(j,l)\in\mathbb{Z}_D\times\mathbb{Z}_n : j+lh=k} \left(\mathcal{F}_D^{-1}\left(\left(\mathcal{F}_{\varphi,D}f\right)_{i,l}\right)_{i=0}^{D-1}\right)_j \\
&= \frac{1}{c_k} \sum_{(j,l)\in\mathbb{Z}_D\times\mathbb{Z}_n : j+lh=k} \left(\mathcal{F}_D^{-1}\left(\left(\mathcal{F}_D\left(f_{m+lh}\varphi_m\right)_{m=0}^{D-1}\right)\right)\right)_j \\
&= \frac{1}{c_k} \sum_{(j,l)\in\mathbb{Z}_D\times\mathbb{Z}_n : j+lh=k} f_{j+lh}\varphi_j \\
&= \frac{1}{c_k} \sum_{(j,l)\in\mathbb{Z}_D\times\mathbb{Z}_n : j+lh=k} f_k\varphi_j \\
&= f_k.
\end{aligned}$$

□

Remark 5. *By means of the short-time Fourier transform we get the frequency range of a signal f as a function of time: the spectrogram of f. The spectrogram displays the values $|\mathcal{F}_\varphi f(\omega,\tau)|$ in a time-frequency diagram. The values $|\mathcal{F}_\varphi f(\omega,\tau)|$ can be interpreted as the frequency range of f at time τ. Compared to the common frequency spectrum (Remark 1) the spectrogram makes more information that is contained in f accessible. In order to reconstruct the signal from the spectrogram by the inversion formula, the phase spectrogram $\arg(\mathcal{F}_\varphi f(\omega,\tau))$ is needed as well.*

2. Dimensionality Reduction

Since real life data is diverse and complex, a given data set remains very high-dimensional even after discretization. Analysis and interpretation of this kind of data sets pose some mathematical and computational challenges where traditional statistical methods might fail. In recent years, controlling and processing of such data has taken on a greater significance as it has become possible to easily transfer, store and record a huge amount of data due to extremely powerful and efficient computers and the expansion of storage capacity. Therefore, in many scientific disciplines such as physics, geography, medicine, musicology, biology and social sciences, to mention just a few, large quantities of raw data have to be handled and hence high-dimensional data sets occur frequently in the fields of data analysis and machine learning.

To understand, visualize and process the structure of this data many new methods known as dimensionality reduction methods have been developed during the last decades. These innovative analytical and numerical tools for data analysis are mainly based on geometrical concepts.

In our case, a basic characteristic of time-frequency data obtained from a signal transform, such as short-time Fourier transform (STFT) or a similar transform, is the high dimensionality of the Euclidean space in which the data is embedded. In this context, a logical consequence is that for many applications a reduction of the data's dimensionality might improve the quality and speed up the computation of the data analysis. We observe that in many cases less than all information contained in the data points is relevant for understanding the underlying characteristics or properties of the data. Also low-dimensional data sets are much easier to operate with in case of classification, visualization or compression.

As a consequence, we would like to reduce the dimensionality of the given data. At this point dimensionality reduction comes in. Dimensionality reduction means to embed the data into a significant manifold of lower dimension within the higher dimensional space in order to encode important information of the data set. This lower dimension should ideally correspond to the intrinsic dimensionality of the data and different strategies are available for estimating this dimensionality (see [36] or [39])).

There are two major types of dimensionality reduction methods: linear and non-linear ones. In this context, linearity refers to the idea that each data point on the manifold is a linear combination of the original data points, i.e. we assume the manifold $\mathcal{M}$ to be a linear subspace (see [19]). Non-linear techniques are mainly based on at least one of the following qualities (see [39]):

1. Preservation of global properties or structures of the data set in the low-dimensional data set,

2. Preservation of local properties or structures,

3. Composition of linear techniques.

In the following sections we discuss two classical dimensionality reduction techniques. We first present the well-known and frequently used Principal Component Analysis (PCA) method. Later, we introduce Laplacian Eigenmaps as a generalization of this concept.

2.1. Basic Notations

Mathematically, the above problem can be formulated as in [25]: Let $X = \{x_k\}_{k=1}^n \subset \mathbb{R}^D$ be a data set of dimensionality D, also called a point cloud data. If much of the information described by X is redundant and can be neglected we try to find a low-dimensional data set $Y \subset \mathbb{R}^d$ which best represents X conserving the characteristics of the data. The dimensionality d of Y is called intrinsic dimensionality of the data and assumed to satisfy $d \ll D$. This process is called *dimensionality reduction*.

An additional concept is the idea of *manifold learning*. In this context, the data is assumed to lie on (or nearby) a (smooth) manifold $\mathcal{M}$ embedded in a D-dimensional space. More precisely, we assume X to be sampled from $\mathcal{M}$, a p-dimensional smooth compact manifold of $\mathbb{R}^D$. In mathematical terms, we search a homeomorphism $\mathcal{B} : \mathbb{R}^D \supset \mathcal{M} \to \Omega \subset \mathbb{R}^d$, where Ω is a p-dimensional submanifold of $\mathbb{R}^d$. Recall that due to the Whitney Embedding Theorem any smooth p-dimensional connected manifold can be embedded in $\mathbb{R}^d$, for all d with $d \geq 2p + 1$ (see [34]).

The objective is to construct a low-dimensional data set Y representing X and its structure using the geometrical informations given by $\mathcal{M}$ (see Figure 2.1). The homeomorphism $\mathcal{B}$ maps the data set X with dimensionality D onto a new data set Y with dimensionality d preserving the main structure of the data.

$$\begin{array}{ccccc} X & \subset & \mathcal{M} & \subset & \mathbb{R}^D \\ \downarrow{\scriptstyle P} & & \downarrow{\scriptstyle \mathcal{B}} & & \\ Y & \subset & \Omega & \subset & \mathbb{R}^d \end{array}$$

In practice, neither the manifold $\mathcal{M}$ nor its low-dimensional representation Ω is known. Therefore, we can only approximate the homeomorphism $\mathcal{B}$ by a dimensionality reduction mapping P as shown in the diagram above.

Combining the key concepts of dimensionality reduction and manifold learning allows the development of more sophisticated dimensionality reduction algorithms. The PCA method and Laplacian Eigenmaps are two examples for such methods. In this setting usually neither the parameter d nor the manifold $\mathcal{M}$ is known.

Remark 6. *In our situation the existence of such a manifold is a reasonable assumption since each source signal has a characteristic frequency range, which does not include all frequencies, i.e. the considered signals are band-limited.*

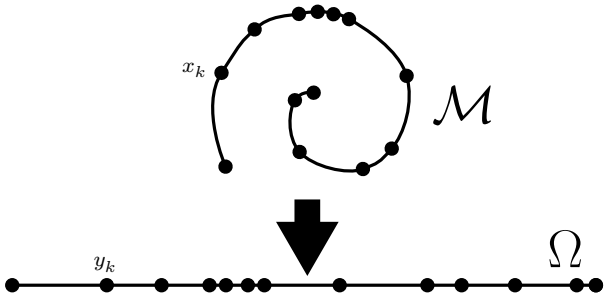


Figure 2.1.: A manifold $\mathcal{M} \subset \mathbb{R}^D$ is embedded in a low-dimensional space.

2.2. PCA - Principal Component Analysis

Principal component analysis (PCA) is probably one of the most frequently used methods in multivariate data analysis. As PCA has many applications, it was discovered independently in different scientific fields and improved by many scientists. It was first introduced by Pearson [42] in 1901 in a biological framework. In the field of stochastic processes PCA is also known as the Karhunen-Loève transform.

As stated before, we consider a data set $X = \{x_k\}_{k=1}^n \subset \mathbb{R}^D$. In the concept of PCA the data points are assumed to lie on or nearby a linear subspace of $\mathbb{R}^D$. The set Y is obtained by projecting the set X onto this subspace. The aim is to find a projection which preserves as much information as possible and discards the redundancy in terms of correlation. This is done by using a principal axis transformation. The redundancy can be measured by the covariance matrix of the data. In this section we proceed as Lee and Verleysen in [35].

2.2.1. Preprocessing

Let us suppose the data set X to be n realizations of a random vector $\mathcal{X} = (\mathcal{X}_1, \ldots, \mathcal{X}_D)^T$. It might be more convenient to use a matrix notation $X = (x_1, \ldots, x_n) \in \mathbb{R}^{D \times n}$. In the following we switch between the data set and the matrix notation depending on context and situation. The idea of PCA is to assume the variables of $\mathcal{X}$ to result from a linear transform $W \in \mathbb{R}^{D \times d}$ of d latent variables $\mathcal{Y} = (\mathcal{Y}_1, \ldots, \mathcal{Y}_d)^T$:

$$\mathcal{X} = W\mathcal{Y}. \tag{2.1}$$

The variables in $\mathcal{Y}$ are assumed to have a Gaussian distribution. The matrix W is supposed to represent an axis transformation, i.e. its columns are normalized and orthogonal to each other. Therefore, we have

$$W^T W = I_d,$$

where I_d is the unit matrix of dimension d (note that WW^T is in general not identical to I_D). We call a matrix *orthonormal* even though only its columns *or* rows are orthonormal, i.e. the matrix W has not to be quadratic.

For the further considerations we need the random variables $\mathcal{X}_i$ to be centered, i.e. of zero-mean. Since in real situations this is not very likely, we center $\mathcal{X}$ in a preprocessing step by subtracting the expectation $E(\mathcal{X})$. This expectation depends on the distribution of $\mathcal{X}$, which is unknown. Hence, the expectation must be approximated by the sample mean:

$$E(\mathcal{X}) \approx \frac{1}{n}\sum_{k=1}^{n} x_k = \frac{1}{n} X \mathbf{1}_n,$$

where $\mathbf{1}_n$ is the column vector of length n containing ones. The centered data set is obtained by

$$X - \frac{1}{n} X \mathbf{1}_n \mathbf{1}_n^T.$$

The task is now to identify W and d.

2.2.2. Criteria leading to PCA

The orthonormal transform W^T converts the given set of possibly correlated variables $\mathcal{X}_i$ into a set of uncorrelated variables called principal components. Uncorrelated means that there is no linear dependency between them, i.e. their covariance is zero. The first principal component is the axis in whose direction the widest variability or scattering of the data occurs. As the variance of a variable is a measure for its range of spread, the first component is the one with the largest variance. The following components each have the largest variance under the constraint of being orthogonal to the previous ones. This leads to an uncorrelated set of variables $\mathcal{Y}$. In other words, the axes of the coordinate system are rotated in such a way that the covariance matrix $C_{\mathcal{X}} = E(\mathcal{X}\mathcal{X}^T)$ is diagonalized.

The (i, j)th element of the covariance matrix $C_{\mathcal{X}}$ is the covariance between $\mathcal{X}_i$ and $\mathcal{X}_j$. Therefore, the covariance matrix is symmetric and the diagonal elements represent the variances of the random variables $\mathcal{X}_i$. Since the covariance matrix is also positive semidefinite there exists a decomposition

$$C_{\mathcal{X}} = V \Lambda V^T,$$

where $V \in \mathbb{R}^{D \times D}$ is an orthonormal matrix whose ith column is an eigenvector v_i of $C_{\mathcal{X}}$ and Λ is the diagonal matrix whose diagonal elements are the corresponding real eigenvalues $\lambda_i \geq 0$ in descending order.

A key requirement for successful dimensionality reduction is to loose as little information of the data as possible by applying the transform W^T. That means to preserve as much of the global variance $\mathrm{var}(\mathcal{X}) = \mathrm{tr}(C_{\mathcal{X}})$ of the data as possible, i.e.

$$\mathrm{var}(\mathcal{X}) = \mathrm{tr}(C_{\mathcal{X}}) \approx \mathrm{tr}(C_{\mathcal{Y}}) = \mathrm{var}(\mathcal{Y}).$$

Lemma 2.1. *In the above setting*

$$\mathrm{var}(\mathcal{X}) \geq \mathrm{var}(\mathcal{Y})$$

holds.

Proof. From (2.1) and the linearity of the expectation E we deduce

$$\begin{aligned} C_{\mathcal{X}} &= E(\mathcal{X}\mathcal{X}^T) \\ &= E(W\mathcal{Y}\mathcal{Y}^T W^T) \\ &= WE(\mathcal{Y}\mathcal{Y}^T)W^T \\ &= WC_{\mathcal{Y}}W^T. \end{aligned} \tag{2.2}$$

Let u_j be an eigenvector of $C_{\mathcal{Y}}$ and let μ_j be the corresponding eigenvalue. With (2.2) and

$$C_{\mathcal{X}}Wu_j = WC_{\mathcal{Y}}W^T Wu_j = W\mu_j u_j = \mu_j Wu_j$$

we observe that the eigenvalues of $C_{\mathcal{Y}}$ are eigenvalues of $C_{\mathcal{X}}$ as well and hence

$$\operatorname{var}(\mathcal{Y}) = \operatorname{tr}(C_{\mathcal{Y}}) = \sum_{j=1}^{d} \mu_j \leq \operatorname{tr}(C_{\mathcal{X}}) = \operatorname{var}(\mathcal{X}).$$

□

We can conclude that the unknown variables in $\mathcal{Y}$ can be assumed to be uncorrelated. This leads to a diagonal covariance matrix $C_{\mathcal{Y}}$ of the centered $\mathcal{Y}$. Our aim is to identify the d unknown uncorrelated variables in $\mathcal{Y}$ from the given covariance matrix of $\mathcal{X}$, i.e. we have to find W, such that $\operatorname{var}(\mathcal{Y})$ is maximal for a given d (see Lemma 2.1). From the next lemma it follows that this can be achieved for $W = VI_{D\times d}$.

Lemma 2.2. *Let $F \in \mathbb{R}^{D\times D}$ be symmetric and positive semidefinite. Among all orthonormal matrices $W \in \mathbb{R}^{D\times d}$, the trace of $D = W^T FW \in \mathbb{R}^{d\times d}$ is maximal if*

$$W = VI_{D\times d}$$

holds, where V contains the eigenvectors of F ordered by the size of the corresponding eigenvalues.

Proof. Since $F = VLV^T$, where L is diagonal containing the eigenvalues l_i of F ordered by size and $V^T V = VV^T = I_D$ we get for $C = V^T W$

$$\begin{aligned} \operatorname{tr}(D) &= \operatorname{tr}(W^T FW) \\ &= \operatorname{tr}(W^T VV^T FVV^T W) \\ &= \operatorname{tr}(W^T VLV^T W) \\ &= \operatorname{tr}(C^T LC) \\ &= \sum_{j=1}^{d}\sum_{i=1}^{D} l_i c_{ij}^2. \end{aligned} \tag{2.3}$$

The matrix C is orthonormal because of the orthonormality of V and W:

$$C^T C = W^T VV^T W = I_d.$$

Hence, we can expand C to an orthonormal square matrix $\tilde{C}$ and thus for the rows of $\tilde{C}$ it follows that $\tilde{c}_i \tilde{c}_i^T = 1$, for $i = 1, \dots D$, which leads to

$$\sum_{j=1}^{d} c_{ij}^2 \leq 1. \tag{2.4}$$

Using (2.3), (2.4) and $l_i \geq l_{i+1}$ yields

$$\begin{aligned}
\mathrm{tr}(D) &= \sum_{i=1}^{D} l_i \sum_{j=1}^{d} c_{ij}^2 \\
&= \sum_{i=1}^{d} l_i \sum_{j=1}^{d} c_{ij}^2 + \sum_{i=d+1}^{D} l_i \sum_{j=1}^{d} c_{ij}^2 \\
&\leq \sum_{i=1}^{d} l_i \sum_{j=1}^{d} c_{ij}^2 + \sum_{i=d+1}^{D} l_d \sum_{j=1}^{d} c_{ij}^2 \\
&= \sum_{i=1}^{d} l_i \sum_{j=1}^{d} c_{ij}^2 + l_d \sum_{j=1}^{d} \sum_{i=d+1}^{D} c_{ij}^2 \\
&= \sum_{i=1}^{d} l_i \sum_{j=1}^{d} c_{ij}^2 + l_d \sum_{j=1}^{d} \left(1 - \sum_{i=1}^{d} c_{ij}^2\right) \\
&= \sum_{i=1}^{d} \left(l_i \sum_{j=1}^{d} c_{ij}^2 + l_d \left(1 - \sum_{j=1}^{d} c_{ij}^2\right)\right) \\
&\leq \sum_{i=1}^{d} \left(l_i \left(\sum_{j=1}^{d} c_{ij}^2 + 1 - \sum_{j=1}^{d} c_{ij}^2\right)\right) \\
&= \sum_{i=1}^{d} l_i.
\end{aligned}$$

For $W = V I_{D \times d}$ we get $C = I_{D \times d}$ and thus with (2.3)

$$\mathrm{tr}(D) = \sum_{i=1}^{D} l_i \sum_{j=1}^{d} c_{ij}^2 = \sum_{i=1}^{d} l_i$$

holds, which finishes the proof. □

Remark 7. *If the PCA model in* (2.1) *is fully respected, the smallest* d *is given by* $D - \dim(\ker(C_{\mathcal{X}}))$. *The trace* $\mathrm{tr}(C_{\mathcal{Y}})$ *is maximal if we keep all eigenvalues of* $C_{\mathcal{X}}$ *apart from the zero eigenvalues. The number of zero eigenvalues of* $C_{\mathcal{X}}$ *is given by* $\dim(\ker(C_{\mathcal{X}}))$

In real situations we often observe some noise and thus the PCA model (2.1) *might be not fully respected. This can result in a situation where all eigenvalues of* $C_{\mathcal{X}}$ *are larger than zero. In this case,* d *cannot be estimated without loss of information. But*

assuming that the variances of the unknown variables $\mathcal{Y}$ are larger than the variance of the noise, it is a natural procedure to choose the eigenvectors associated to the largest eigenvalues. Thus, we have almost the same situation as before.

Remark 8. *In the above setting the random variables $\mathcal{X}_i$ and their probability densities are unknown. Therefore, we need to estimate the covariance matrix $C_{\mathcal{X}}$ using the given data X. As known from empirical statistics, for a random vector $\mathcal{X}$ this can be done by $C_{\mathcal{X}} \approx \frac{1}{n} XX^T$. The factor $\frac{1}{n}$ is neither changing the algebraic multiplicity of zero eigenvalues nor the eigenspaces, thus it is sufficient to consider XX^T.*

According to the above explained background, PCA can be performed by a singular value decomposition of X, i.e. by an eigenvalue decomposition of the data's covariance matrix XX^T. Therefore, the subspace on which we project the data set X is given by the linear span of eigenvectors of the covariance matrix.

As the shortest distance from a point to a subspace is the distance from this point to its orthogonal projection into the subspace, minimizing the sum of these distances is an alternative formulation of the above explained problem.

The following lemma summarizes these observations.

Lemma 2.3. *Let $X = (x_1, \dots, x_n) \in \mathbb{R}^{D\times n}$ be a matrix whose columns represent a centered data set and let V be a matrix containing the eigenvectors of XX^T ordered by the size of the corresponding eigenvalues.*

i) The global variance of the reduced data set $Y = W^T X$ is maximized among all orthogonal matrices $W \in \mathbb{R}^{D\times d}$ if $W = VI_{D\times d}$, i.e.

$$\operatorname{var}(Y) = \operatorname{tr}(YY^T) = \max_{\tilde{W}} \operatorname{tr}(\tilde{W}^T XX^T \tilde{W}) = \max_{\tilde{W}} \sum_{k=1}^{n} \|\tilde{W}^T x_k\|^2.$$

ii) The sum of distances from the data points x_k to their images $W^T x_k$ is minimized among all orthogonal matrices $W \in \mathbb{R}^{D\times d}$ if $W = VI_{D\times d}$, i.e.

$$\operatorname{err}_{PCA}(W, X) = \min_{\tilde{W}} \sum_{k=1}^{n} \|x_k - \tilde{W}^T x_k\|^2.$$

Proof. The statement *i)* is a direct consequence of Lemma 2.2.

ii). Bearing in mind the orthogonality property of W the well-known Pythagoras' Theorem gives us

$$\|x_k\|^2 = \|x_k - W^T x_k\|^2 + \|W^T x_k\|^2, \quad \text{for } k = 1, \dots, n.$$

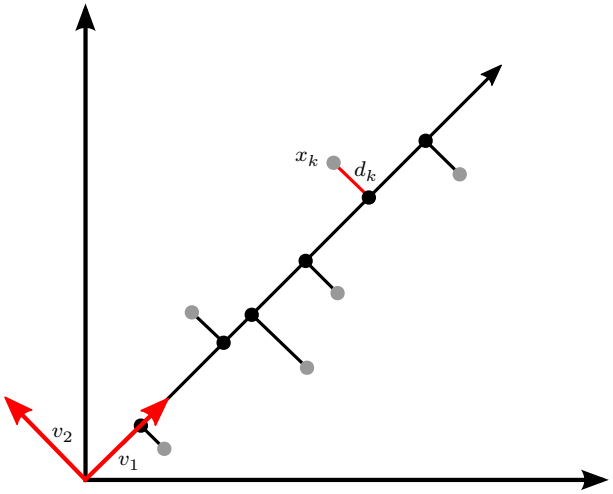


Figure 2.2.: Identification of the principal components as the orthogonal directions in which the data is scattering the most.

Hence, we deduce the relation between the sum of distances and the global variance as

$$\begin{aligned}\operatorname{var}(W^T X) &= \sum_{k=1}^{n} \|W^T x_k\|^2 \\ &= \sum_{k=1}^{n} \|x_k\|^2 - \|x_k - W^T x_k\|^2 \\ &= \sum_{k=1}^{n} \|x_k\|^2 - \operatorname{err}_{PCA}(W, X).\end{aligned}$$

From this equation it follows directly that a minimization of $\operatorname{err}_{PCA}(W, X)$ is equivalent to a maximization of $\operatorname{var}(W^T X)$. □

Figure 2.2 serves to illustrate the general idea of PCA for the case of $D = 2$ and $d = 1$. It depicts the distances d_k whose sum has to be minimized and also the directions v_i in which the data is distributed with maximum variance. We have learned from Lemma 2.3 that this problem can be solved by considering the eigenvalue decomposition of the covariance matrix.

In the case where Equation (2.1) is fully respected, it is obvious that there exists a back projection, projecting the data from the subspace back to the original high dimensional space using W. But if (2.1) is not fully respected, i.e. the data is not lying in the subspace (just near by), it might be difficult to find an exact back projection and sometimes it does not even exist. To deal with this problem, we assume the data to lie in the subspace, and if it does not, we neglect the error.

In this section we have introduced a dimensionality reduction method $P = W^T$ (compare Section 2.1). To conclude this section we define PCA as follows.

Definition 2.1 ([13])**.** The PCA of a random vector $\mathcal{X}$ of size D with finite covariance $C_{\mathcal{X}}$ is a pair $\{W, C_{\mathcal{Y}}\}$ of matrices such that

i) the covariance factorizes into

$$C_{\mathcal{X}} = W C_{\mathcal{Y}} W^T,$$

where $C_{\mathcal{Y}}$ is diagonal with positive entries and W has full column rank d.

ii) W is a $D \times d$ matrix whose columns are orthogonal to each other, i.e. $W^T W$ is diagonal.

2.3. LE - Laplacian Eigenmaps

The second dimensionality reduction method used in this work is called *Laplacian Eigenmaps* (LE). This method belongs to the family of non-linear dimensionality reduction techniques and is, as PCA, based on spectral decomposition. LE can be seen as a generalization of PCA from linear subspaces to arbitrary smooth p-manifolds. In this context, a p-manifold is a topological space, which is locally homeomorphic to the Euclidean space $\mathbb{R}^p$. LE is a topology preserving method, i.e. it reduces the dimension of a given data set by preserving rather its topology than its pairwise distances (as for example Isomap [49]). In contrast to other methods, LE is a local technique whose operating principle is the use of '*information contained in the data in order to establish the topology of the data set and compute the shape and topology of the embedding accordingly*' [35]. LE was first used in 2002 by Belkin and Niyogi [3].

In order to preserve the topological structure of the data set (compare Figure 2.3), we try to minimize the distances between neighboring data points. This is done using graph theoretical tools like the Laplacian operator on a graph which is closely related to the graph's adjacency matrix.

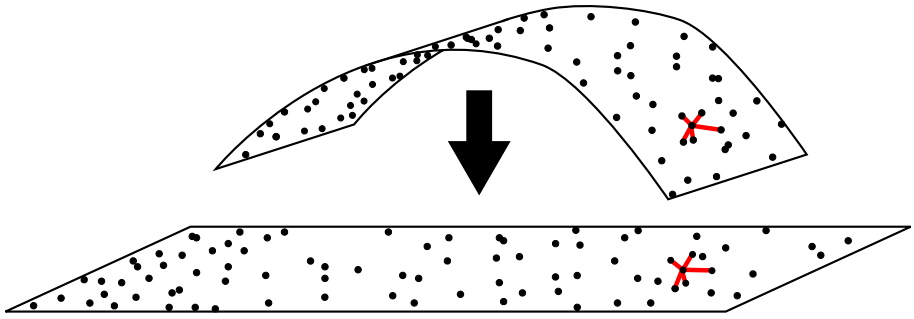


Figure 2.3.: Laplacian Eigenmaps preserves the neighborhood structure.

2.3.1. Neigborhood graph

As for the PCA method, we consider a data set $X = \{x_k\}_{k=1}^n \subset \mathbb{R}^D$ and we search a low dimensional representation $Y = \{y_k\}_{k=1}^n \subset \mathbb{R}^d$ of X. Further, we assume this data

to lie on or near by a (unknown) smooth p-manifold.

Let us consider the data points as vertices $v_k \in V_n$ of an undirected graph $\mathcal{G}$. Two vertices are connected if the corresponding data points are adjacent. If the set of edges is denoted by E, the graph $\mathcal{G}$ is given by

$$\mathcal{G} = (V_n, E)\,.$$

Let us denote by $e(k,i)$ the edge connecting the vertices v_k and v_i. Although the manifold is unknown, it can be represented with good accuracy by the graph $\mathcal{G}$ if n is large enough. In this context, adjacent refers either to ϵ-ball neighborhoods or to r-ary neighborhoods. The parameter ϵ, or r respectively, have to be chosen in such a way that the constructed graph $\mathcal{G}$ has no isolated vertices. The neighborhood relationship is usually described by a (sparse) adjacency matrix $A \in \mathbb{R}^{n\times n}$ with

$$a_{ki} = \begin{cases} 1 & \text{if } e(k,i) \in E \\ 0 & \text{otherwise.} \end{cases}$$

It is obvious that A is symmetric and has no zero column. As before, the purpose is now to map the given data set X into a data set Y of lower dimensionality with the same adjacency relationships. Therefore, we have to define a criterion to measure the accuracy of the mapping (see [3]):

$$\mathrm{err}_{LE}(W,Y) = \frac{1}{2}\sum_{k,i=1}^{n} \|y_k - y_i\|^2 w_{ki} \tag{2.5}$$

with $W = (w_{ki})_{k,i=1,\dots,n}$ being a symmetric weight matrix. The matrix W is defined by the adjacency matrix A as $w_{ki} = 0$, if $a_{ki} = 0$ and $w_{ki} \geq 0$ otherwise. Thus $\mathcal{G}$ becomes a weighted graph. The weights w_{ki} to be determined shall ensure the preservation of the topology, i.e. they can be interpreted as penalties. These penalties should be heavier for edges $e(k,i)$ between vertices associated to close data points x_k and x_i, and small for edges between vertices associated to far away data points. To minimize the error term we consider the following alternative characterization.

Lemma 2.4. *In the above setting let D be the diagonal matrix defined by $D_{kk} = \sum_{i=1}^{n} w_{ki}$. Then it holds*

$$\mathrm{err}_{LE}(W,Y) = \mathrm{tr}(YLY^T),$$

where $L = D - W$ is the Laplacian matrix of the graph $\mathcal{G}$. By Y we denote the data matrix containing column by column the low-dimensional representation of X.

Proof. Let us denote by $\bar{y}_j$ the transposed of the jth row of the data matrix Y. Using the definition of the norm, the Binomial Theorem and the fact that W is symmetric,

we get

$$\begin{aligned}
\mathrm{err}_{LE}(W,Y) &= \frac{1}{2}\sum_{k,i=1}^{n} \|y_k - y_i\|^2 w_{ki} \\
&= \frac{1}{2}\sum_{k,i=1}^{n}\sum_{j=1}^{d} \left(y_{jk} - y_{ji}\right)^2 w_{ki} \\
&= \frac{1}{2}\sum_{j=1}^{d}\sum_{k,i=1}^{n} \left(y_{jk}^2 + y_{ji}^2 - 2y_{jk}y_{ji}\right) w_{ki} \\
&= \frac{1}{2}\sum_{j=1}^{d}\left(\sum_{k=1}^{n} y_{jk}^2 \sum_{i=1}^{n} w_{ki} + \sum_{i=1}^{n} y_{ji}^2 \sum_{k=1}^{n} w_{ki} - 2\sum_{k,i=1}^{n} y_{jk}y_{ji}w_{ki}\right) \\
&= \frac{1}{2}\sum_{j=1}^{d}\left(\sum_{k=1}^{n} y_{jk}^2 D_{kk} + \sum_{i=1}^{n} y_{ji}^2 D_{ii} - 2\bar{y}_j^T W \bar{y}_j\right) \\
&= \frac{1}{2}\sum_{j=1}^{d}\left(2\sum_{k=1}^{n} y_{jk}^2 D_{kk} - 2\bar{y}_j^T W \bar{y}_j\right) \\
&= \sum_{j=1}^{d}\left(\bar{y}_j^T D \bar{y}_j - \bar{y}_j^T W \bar{y}_j\right) \\
&= \sum_{j=1}^{d} \bar{y}_j^T L \bar{y}_j \\
&= \mathrm{tr}(YLY^T).
\end{aligned}$$

□

Remark 9. *We observe that the Laplacian matrix is symmetric and positive semidefinite. This follows from the symmetry of W and D and the proof of Lemma 2.4 since*

$$v^T L v = \frac{1}{2}\sum_{k,i=1}^{n} \left(v_k - v_i\right)^2 w_{ki} \geq 0,$$

for all $v \in \mathbb{R}^n$.

Therefore, minimizing the error err_{LE} is equivalent to finding an Y which minimizes $\mathrm{tr}(YLY^T)$. Since D is positive definite (because the graph $\mathcal{G}$ is connected) it induces an inner product

$$\langle x_k, x_i \rangle = x_k^T D x_i$$

on $\mathbb{R}^n$. In the following orthogonality refers to this inner product.

2.3.2. Criteria leading to LE

We would like the solution Y to be unique in the sense of translation invariance. This requirement results from the fact that a translation of the whole data set Y does not change the distances between the data points, i.e. if Y is a minimizer, then $Y + C\mathbf{1}_{d\times n}$, where $C \in \mathbb{R}^{d\times d}$ is a diagonal matrix, is also a minimizer. Therefore, we additionally demand that

$$YD\mathbf{1}_{n\times 1} = 0. \tag{2.6}$$

This constraint is reasonable since for two solutions Y_1 and $Y_2 = Y_1 + C\mathbf{1}_{d\times n}$ which fulfill the constraint (2.6),

$$Y_1 D\mathbf{1}_{n\times 1} = 0 = (Y_1 + C\mathbf{1}_{d\times n})D\mathbf{1}_{n\times 1} = C\mathbf{1}_{d\times n}D\mathbf{1}_{n\times 1} \tag{2.7}$$

holds. Since in (2.7) we have $D \neq 0$, it follows $C = 0$ and hence $Y_1 = Y_2$. The constraint $YD\mathbf{1}_{n\times 1} = 0$ implies that each row of Y is orthogonal to the constant vector $(1, \ldots, 1)$.

Furthermore, from (2.5) it follows that the minimization problem has a trivial solution where all data points are mapped on a single point. Such a solution has the form $Y = C\mathbf{1}_{d\times n}$, where $C \in \mathbb{R}^{d\times d}$ is a diagonal matrix. This solution should be excluded. Since the problem is already unique in the sense of translation invariance, we only need to exclude the solution $Y = 0$. This can be done together with a normalization (to remove an arbitrary scaling factor in the embedding) by $YDY^T = I_{d\times d}$ (see [4]).

Theorem 2.1. *The solution of*

$$\operatorname{argmin}_{\substack{YDY^T = I_{d\times d} \\ YD\mathbf{1}_{n\times 1} = 0}} \operatorname{tr}(YLY^T) \tag{2.8}$$

is provided by a matrix of eigenvectors corresponding to the d smallest non-zero eigenvalues of the generalized eigenvalue problem

$$Lv = \lambda Dv,$$

i.e. $Y = (f_1, \ldots f_d)^T$ with $Lf_j = \lambda_j Df_j$ and $\lambda_j \neq 0$.

For the proof of this theorem we need two standard statements concerning necessary and sufficient conditions of optimization problems, which we present without their proofs.

Theorem 2.2 (Karush-Kuhn-Tucker)**.** *Let $x^\star$ be a local solution of*

$$\textit{Minimize } f(x) \textit{ on } M = \{x \in \mathbb{R}^n : (g, h)(x) = 0\},$$

let the objective function $f : \mathbb{R}^n \to \mathbb{R}$ and the constraint functions $(g, h) : \mathbb{R}^n \to \mathbb{R}^l \times \mathbb{R}^m$ be continuously differentiable on a neighborhood of $x^\star$. If $\operatorname{rg}\left(J(g, h)(x^\star)\right) = \operatorname{rg}\left(J(g), J(h)\right)(x^\star)) = m+l$, there exists a pair of Lagrange multipliers $(\Lambda, \mu) \in \mathbb{R}^l \times \mathbb{R}^m$ with

$$\nabla f(x^\star) + J(g)(x^\star)^T \Lambda + J(h)(x^\star)^T \mu = 0.$$

In this context, J denotes the Jacobian matrix.

Proof. See [23]. □

Theorem 2.3. *Given the optimization problem*

$$\text{Minimize } f(x) \text{ on } M = \{x \in \mathbb{R}^n : (g,h)(x) = 0\}. \tag{$*$}$$

Let the functions f, h and g defined as before be twice continuously differentiable in $x^\star \in M$. If there exists a pair (Λ, μ) as in Theorem 2.2 with

$$v^T \left[\operatorname{Hess}(f(x^\star)) + \sum_{j=1}^{l} \Lambda_j \operatorname{Hess}(g_j) + \sum_{j=1}^{m} \mu_j \operatorname{Hess}(h_j) \right] v > 0$$

for all $v \in \ker\left(J(g,h)\right)(x^\star) \setminus \{0\}$, then $x^\star$ is an isolated local solution of $(*)$.

Proof. See [23]. □

Proof of Theorem 2.1. In order to determine Y such that $\operatorname{tr}(YLY^T)$ under the given constraints is minimal, we use the well-known method of Lagrange multipliers. Let us consider the Lagrange function

$$\mathcal{L}(Y, \Lambda, \mu) = \operatorname{tr}(YLY^T) + \sum_{i,j=1}^{d} \lambda_{ij}(\bar{y}_i^T D \bar{y}_j - \delta_{ij}) + \sum_{j=1}^{d} \mu_j \bar{y}_j^T D \mathbf{1}_{n\times 1},$$

with $\mathcal{L} : \mathbb{R}^{d\times n} \times \mathbb{R}^{d\times d} \times \mathbb{R}^d \to \mathbb{R}$, $\Lambda = (\lambda_{ij})_{i,j=1,\dots d}$, $\mu \in \mathbb{R}^d$ and $\bar{y}_j$, for $j = 1, \dots d$, the transposed of the jth row of the data matrix Y. Since the condition $\bar{y}_i^T D \bar{y}_j = \delta_{ij}$ is symmetric in i and j, the matrix Λ is symmetric.

For the moment we consider the auxiliary problem, where only the diagonal entries of the constraint $YDY^T = I_{d\times d}$ are kept, i.e. Λ is diagonal.

We can write a given vector $\eta \in \mathbb{R}^{dn}$ as a $d \times n$ matrix as follows:

$$Y = (y_{pq})_{\substack{p=1,\dots,d \\ q=1,\dots,n}} = \begin{pmatrix} \eta_1 & \eta_{d+1} & \cdots & \eta_{d(n-1)+1} \\ \vdots & \vdots & \ddots & \vdots \\ \eta_d & \eta_{d2} & \cdots & \eta_{dn} \end{pmatrix}.$$

Using the Euclidean division we can find a unique decomposition $m = d(q-1)+p$, with $1 \le p \le d$ and $1 \le q \le n$. This provides a natural correspondence between η and Y: $\eta_m = y_{pq}$. For simplicity's sake, we use the following vector notation for the involved functions:

$$\begin{aligned} f : \mathbb{R}^{dn} &\to \mathbb{R} \\ \eta &\mapsto \operatorname{tr}\left(YLY^T\right) \\ g_j : \mathbb{R}^{dn} &\to \mathbb{R} \\ \eta &\mapsto \left(YDY^T\right)_{jj} - 1 \\ h_j : \mathbb{R}^{dn} &\to \mathbb{R} \\ \eta &\mapsto \left(YD\mathbf{1}_{n\times 1}\right)_j, \end{aligned}$$

for $j = 1, \ldots, d$.

To apply the necessary condition of Theorem 2.2 we have to show that the rank of the Jacobian matrix $J(g,h)(\eta^\star)$ of the constraints is $2d$. With the notation introduced above, the Jacobian matrix $J(g,h)(\eta)$ is given by

$$J(g,h)(\eta) = \left(\begin{array}{cccc|cccc|c|cccc} 2\eta_1 D_{11} & & 0 & & 2\eta_{d+1} D_{22} & & 0 & & & 2\eta_{d(n-1)+1} D_{nn} & & 0 & \\ & \ddots & & & & \ddots & & & \cdots & & \ddots & & \\ 0 & & 2\eta_d D_{11} & & 0 & & 2\eta_{2d} D_{22} & & & 0 & & 2\eta_{dn} D_{nn} & \\ \hline D_{11} & & 0 & & D_{22} & & 0 & & & D_{nn} & & 0 & \\ & \ddots & & & & \ddots & & & \cdots & & \ddots & & \\ 0 & & D_{11} & & 0 & & D_{22} & & & 0 & & D_{nn} & \end{array}\right).$$

From the constraint

$$g_j(\eta^\star) = \sum_{k=1}^{n} (\eta^\star_{d(k-1)+j})^2 D_{kk} - 1 = 0$$

for a solution $\eta^\star$ like in Theorem 2.2 it follows that there exists a k such that $\eta^\star_{d(k-1)+j} \neq 0$, for $j = 1, \ldots, d$. These non-zero entries of $J(g,h)(\eta^\star)$ are in the first d rows on different positions and therefore the first d rows are linearly independent. From the structure of $J(g,h)(\eta^\star)$ and the fact that $\mathcal{G}$ is connected, i.e. D is positive definite, it follows that for $1 \leq j \leq d$ only the jth and the $(j+d)$th row of $J(g,h)(\eta^\star)$ can be linearly dependent. Linear dependence would imply that for all j the entry $\eta_{(k-1)d+j}$ has the same value for all $k = 1, \ldots, n$, i.e. all data points are mapped on the same point. This was excluded, and thus the row rank of $J(g,h)(\eta^\star)$ is $2d$. Hence, for each minimum $\eta^\star$ of f we get Lagrange multipliers (Λ, μ) such that

$$\nabla \left(f(\eta^\star) + \sum_{j=1}^{d} \lambda_{jj} g_j(\eta^\star) + \sum_{j=1}^{d} \mu_j h_j(\eta^\star) \right) = \nabla \mathcal{L}(\eta^\star, \Lambda, \mu) = 0. \tag{2.9}$$

In order to find the solutions $\eta^\star$, we differentiate the function $\mathcal{L}$ in direction of $\eta_m = y_{pq}$, with $m = d(q-1) + p$ using the fact that L is symmetric:

$$\begin{aligned} \frac{\partial}{\partial y_{pq}} \mathcal{L}(Y, \Lambda, \mu) &= \frac{\partial}{\partial y_{pq}} \operatorname{tr}(YLY^T) + \frac{\partial}{\partial y_{pq}} \sum_{i,j=1}^{d} \lambda_{ij} (\bar{y}_i^T D \bar{y}_j - \delta_{ij}) + \frac{\partial}{\partial y_{pq}} \sum_{j=1}^{d} \mu_j \bar{y}_j^T D \mathbf{1}_{n\times 1} \\ &= \frac{\partial}{\partial y_{pq}} \sum_{j=1}^{d} \sum_{k,l=1}^{n} y_{jk} L_{kl} y_{jl} + \frac{\partial}{\partial y_{pq}} \sum_{i,j=1}^{d} \lambda_{ij} \left(\sum_{k=1}^{n} y_{ik} y_{jk} D_{kk} - \delta_{ij} \right) \\ &\qquad + \frac{\partial}{\partial y_{pq}} \sum_{j=1}^{d} \mu_j \sum_{k=1}^{n} y_{jk} D_{kk} \\ &= \sum_{k=1}^{n} L_{kq} y_{pk} + \sum_{l=1}^{n} L_{ql} y_{pl} + \sum_{i=1}^{d} \lambda_{ip} y_{iq} D_{qq} + \sum_{j=1}^{d} \lambda_{pj} y_{jq} D_{qq} + \mu_p D_{qq} \\ &= (YL)_{pq} + (YL^T)_{pq} + \sum_{j=1}^{d} (\lambda_{jp} + \lambda_{pj}) y_{jq} D_{qq} + (\mu \mathbf{1}_{1\times n} D)_{pq} \end{aligned}$$

$$= \left(Y(L+L^T)\right)_{pq} + \left((\Lambda+\Lambda^T)YD\right)_{pq} + (\mu \mathbf{1}_{1\times n} D)_{pq}$$
$$= (2YL)_{pq} + (2\Lambda YD)_{pq} + (\mu \mathbf{1}_{1\times n} D)_{pq}\,.$$

Hence, we get for the gradient of the Lagrange function $\mathcal{L}$ in the direction of Y:

$$\nabla_Y \mathcal{L}(Y,\Lambda,\mu) = 2YL + 2\Lambda YD + \mu \mathbf{1}_{1\times n} D.$$

Thus, solving (2.9) leads to solving

$$2YL + 2\Lambda YD + \mu \mathbf{1}_{1\times n} D = 0, \tag{2.10}$$

or column by column

$$2L\bar{y}_j + 2\lambda_{jj} D\bar{y}_j + \mu_j D\mathbf{1}_{n\times 1} = 0.$$

Multiplying by $\mathbf{1}_{1\times n}$ gives

$$2\mathbf{1}_{1\times n} L\bar{y}_j + 2\lambda_{jj}\mathbf{1}_{1\times n} D\bar{y}_j + \mu_j \mathbf{1}_{1\times n} D\mathbf{1}_{n\times 1} = 0. \tag{2.11}$$

Due to the special form of L the column sums $\mathbf{1}_{1\times n} L$ of L (and also the row sums) are zero. Thus, the first term vanishes for all j. The same holds for the middle term since the constraint $YD\mathbf{1}_{n\times 1} = 0$ implies the orthogonality of the transposed rows $\bar{y}_j$ of Y to the vector $\mathbf{1}_{n\times 1}$. Therefore, Equation (2.11) turns into

$$\mu_j \mathbf{1}_{1\times n} D\mathbf{1}_{n\times 1} = \mu_j \operatorname{tr}(D) = 0,$$

such that we can deduce $\mu_j = 0$, for all j, because $\mathcal{G}$ is connected. Thus, solving Equation (2.10) reduces to solving the generalized eigenvalue problem

$$YL = \Lambda YD,$$

i.e. Y contains row by row generalized eigenvectors of L. Additionally, we observe that the diagonal entries of Λ are d generalized non-zero eigenvalues of L. To conclude that the zero eigenvalue is excluded, we use a graph theoretical statement, which says that the dimension of the eigenspace Eig(0) to the generalized zero eigenvalue of L equals the number of components of $\mathcal{G}$. Therefore, the multiplicity of the generalized zero eigenvalue is one and $\mathrm{Eig}(0) = \mathrm{span}(\mathbf{1}_{n\times 1})$. Hence, the generalized zero eigenvalue is not contained in Λ, since the corresponding eigenvector is excluded by the constraints of the minimization problem.

The eigenvectors $\bar{y}_j$ fulfill in a natural way the orthogonality condition $\bar{y}_j D\bar{y}_j^T = 0$. Furthermore, we show that the solutions $(\eta^\star, \Lambda, \mu)$ are isolated local minima. Therefore, we consider the Hessian matrices of f, $\sum_{j=1}^d g_j$ and $\sum_{j=1}^d h_j$:

$$\mathrm{Hess}(f) = \nabla^2 f = 2\left(\begin{array}{c|c|c|c} L_{11}I_{d\times d} & L_{21}I_{d\times d} & \cdots & L_{n1}I_{d\times d} \\ \hline L_{12}I_{d\times d} & L_{22}I_{d\times d} & & L_{n2}I_{d\times d} \\ \hline \vdots & & \ddots & \vdots \\ \hline L_{1n}I_{d\times d} & L_{2n}I_{d\times d} & \cdots & L_{nn}I_{d\times d} \end{array}\right)$$

$$\mathrm{Hess}(\sum_{j=1}^{d} g_j) = \nabla^2 \sum_{j=1}^{d} g_j = 2 \left(\begin{array}{c|c|c|c} \lambda_1 D_{11} I_{d\times d} & 0 & \cdots & 0 \\ \hline 0 & \lambda_2 D_{22} I_{d\times d} & & \vdots \\ \hline \vdots & & \ddots & 0 \\ \hline 0 & \cdots & 0 & \lambda_n D_{nn} I_{d\times d} \end{array}\right)$$

$$\mathrm{Hess}(\sum_{j=1}^{d} h_j) = 0.$$

The matrix $\mathrm{Hess}(f)$ is positive semidefinite. This follows from Remark 9 because it has the same eigenvalues as L just with d-fold multiplicity. The matrix $\mathrm{Hess}(\sum_{j=1}^{d} g_j)$ is positive definite because it is diagonal and the diagonal entries do not vanish. Thus, we can deduce the positive definiteness of

$$\mathrm{Hess}(\mathcal{L}) = \mathrm{Hess}(f) + \mathrm{Hess}(\sum_{j=1}^{d} g_j) + \mathrm{Hess}(\sum_{j=1}^{d} h_j).$$

Hence, we can apply the sufficient condition of Theorem 2.3 and deduce that $\eta^\star$ are isolated local minima if the corresponding Y contains row by row d generalized eigenvectors of L.

The remaining task is now to find a global minimum among these. Using the constraint $YDY^T = I_{d\times d}$ for a local minimum

$$\begin{aligned} \mathrm{tr}(YLY^T) &= \sum_{j=1}^{d} \bar{y}_j^T L \bar{y}_j \\ &= \sum_{j=1}^{d} \lambda_j \bar{y}_j^T D \bar{y}_j \\ &= \sum_{j=1}^{d} \lambda_j \end{aligned}$$

holds. And thus, the trace is minimized for Y containing the eigenvectors corresponding to the d smallest non-zero eigenvalues.

Until here we have found a global solution for the auxiliary problem. Since Y is composed of generalized eigenvectors, the additional constraints $\bar{y}_i^T D \bar{y}_j = 0$, for $i \neq j$ are also fulfilled. Therefore, this solution is also a candidate for the initial problem (2.8). The global minimum of the initial problem is taken on, because the set $\{Y \in \mathbb{R}^{d\times n} | YDY^T = I_{d\times d} \text{ and } YD\mathbf{1}_{n\times 1} = 0\}$ is compact. Thus, each global minimum of the auxiliary problem is also a global minimum of the initial problem. □

Remark 10 ([35])**.** *An alternative way to find a low dimensional embedding consists of normalizing the Laplacian matrix by*

$$L' = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = \left[\frac{L_{kl}}{\sqrt{D_{kk} D_{ll}}}\right]_{1\leq k,l\leq n}$$

and calculating its eigenvectors by

$$L' = U\Gamma U^T.$$

The eigenvectors corresponding to the d smallest non-zero eigenvalues will give us a d-dimensional embedding of the data set.

Theorem 2.4. *The eigenvalues Γ are the same as the eigenvalues of the generalized eigenvalue problem of Theorem 2.1, i.e. the d-dimensional embedding is (up to a scaling) the same as the solution obtained by Theorem 2.1.*

Proof. Let f_k, for $k = 1, \ldots, n$, be the n generalized eigenvectors of L given by $Lf_k = \lambda_k D f_k$. Then, for $v_k = D^{\frac{1}{2}} f_k$,

$$L'v_k = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} D^{\frac{1}{2}} f_k = \lambda_k v_k$$

holds. Since the diagonal entries of D are non-zero, the multiplication with $D^{\frac{1}{2}}$ causes only a component-wise scaling of the eigenvectors. □

Remark 11. *There are different possibilities for the choice of the weights w_{ij}. These weights will ensure that close points are mapped onto close points. Here, we mention the two possibilities proposed in [3].*

- *Heat kernel. For $t \in \mathbb{R}$*

$$w_{ij} = \begin{cases} e^{-\frac{\|x_i - x_j\|^2}{t}} & \text{if } e(i,j) \in E \\ 0 & \text{otherwise} \end{cases}.$$

  This choice is motivated by the analogy of the Laplacian matrix of a graph to the Laplace-Beltrami operator on manifolds. More details can be found in [4].

- *Simple weights.*

$$w_{ij} = \begin{cases} 1 & \text{if } e(i,j) \in E \\ 0 & \text{otherwise} \end{cases}.$$

  This corresponds to the first option for $t = \infty$.

Remark 12. *The reconstruction mapping (or back transform) to the high-dimensional space, is not as easy as it is in the case of PCA even in the case where the LE model is fully respected.*

3. ICA - Independent Component Analysis

Independent component analysis (ICA) is a stochastical method for decomposing a given data set into a set of statistically (i.e. mutually) independent components. This statistical independence can be achieved by maximization of the non-Gaussianity or by minimization of the mutual information. This idea is based on the Central Limit Theorem which says that the distribution of a sum of independent random variables tends to a Gaussian distribution.

ICA is frequently used for blind source separation (BSS). In this context we assume a signal to be a linear mixture of different unknown source signals. The aim is to retrieve these source signals without knowing anything about the mixing process.

The problem of ICA was first proposed and so named by Herault and Jutten in [26] around 1986 because of its similarities to PCA. ICA is closely related to the so called *cocktail party effect*. This effect describes the phenomenon of selective listening. Suppose the conversation of two people being recorded with two different microphones then, depending on its position, each microphone registers a signal. The weighted sums of these two signals correspond to the two source signals. The problem of determining these source signals and the weights leads to a set of linear equations with more unknown variables than equations. In the above example we obtain two equations with six unknown variables.

The core idea of ICA is to make some statistical assumptions on the source signals in order to balance the disproportion of equations and unknowns. In concrete terms we assume the signals to be statistically independent. This does not need to be completly true in practice [30].

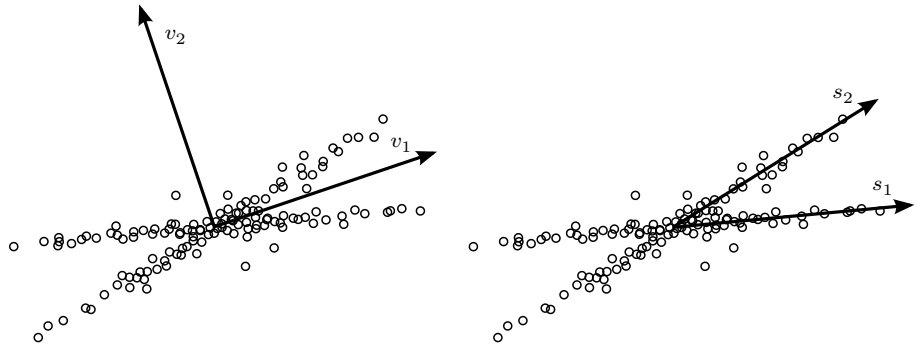
To give a mathematical formulation of the just explained situation, we follow [13] and consider d weighted sums $y_1, \dots, y_d$ of ρ source signals $s_1, \dots, s_\rho$ called independent components

$$y_i = \sum_{j=1}^{\rho} a_{ij} s_j, \qquad d \geq \rho.$$

The functions of time s_j and y_i can be interpreted as the realization of random variables as we did in Section 2.2. This leads to the following linear statistical model:

$$\mathcal{Y} = A\mathcal{S},$$

where $\mathcal{Y}$ and $\mathcal{S}$ are random vectors with values in $\mathbb{R}^d$ and $\mathbb{R}^\rho$ respectively and $A \in \mathbb{R}^{d \times \rho}$. The components of the vector $\mathcal{S}$ are maximizing a 'contrast' function. The contrast of a vector is maximal if its components are statistically independent. Both $\mathcal{Y}$ and $\mathcal{S}$ are assumed to have zero mean and a finite covariance. Thus, the ICA of a random



(a) Components computed by PCA **(b)** Components computed by ICA

Figure 3.1.: ICA recovers the structure of the data better than PCA, because the independent components are not required to be orthogonal.

vector consists of searching a linear transformation such that the statistical dependence between its components is minimized.

Given n realizations of the random vector $\mathcal{Y}$ we aim to estimate both, A and the corresponding realizations of $\mathcal{S}$. In the trivial case where A is known we simply compute its pseudoinverse $G = A^{-1}$ and thus

$$\mathcal{S} = G\mathcal{Y}. \tag{3.1}$$

But since A is unknown we aim to find another way to estimate G.

One method we already know is PCA. While PCA uses only statistics of second order, i.e. the covariance matrix, ICA however uses statistics of all orders. As a consequence PCA can only impose independence up to the second order and hence it defines orthogonal directions. But in practice there are situations, where this is not sufficient as can be seen in Figure 3.1. With ICA the data is not only uncorrelated but also as statistically independent as possible. ICA can thus be seen as an extension of PCA.

Definition 3.1 ([13])**.** The *ICA* of a random vector $\mathcal{Y}$ of length d with finite covariance $C_{\mathcal{Y}}$ is a pair $\{A, \Lambda\}$ of matrices such that

i) the covariance factorizes into

$$C_{\mathcal{Y}} = A\Lambda^2 A^T,$$

where Λ is diagonal with real positive entries and A has full column rank ρ.

ii) the vector $\mathcal{Y}$ can be written as $\mathcal{Y} = A\mathcal{S}$, where $\mathcal{S}$ is a $\rho \times 1$ random vector with covariance Λ^2 and whose components are 'the most independent possible', in the sense of maximization of a given 'contrast function' that will be defined later on.

Remark 13. *If we formulate the problem of PCA like in Definition 2.1, it becomes obvious that ICA is an extension of PCA.*

Remark 14. *The decomposition of the covariance $C_{\mathcal{Y}}$ defined above is not unique. Multiplying the components of the random vector with non-zero scalar factors or changing the order of the components is not affecting the statistical independence of the components. The Definition 3.1 characterizes in effect an equivalence class of decompositions rather than a single one.*

Against the background of computation equivalence classes are not easy to handle. For this reason we intend to define a unique representative of each of these equivalence classes. In order to do so we have to assume some additional constraints. These constraints are arbitrary. We use the three constraints proposed in [13].

Definition 3.2. The constraints we use to guarantee the uniqueness of ICA are

i) the columns of A have unit norm,

ii) the entries of Λ are sorted in decreasing order,

iii) the entry of largest modulus in each column of A is positive.

Remark 15. *The constraints in Definition 3.2 ensure also the uniqueness of PCA.*

Now that we have formulated the problem of ICA, we will derive a way to solve it in the next section. This solution is found by minimizing the mutual information, which turns out to be a numerically efficient and accurate way of doing ICA.

3.1. On statistics and contrast functions

In Definition 3.1 the objective vector $\mathcal{S}$ is stated to be 'the most independent possible'. Since the contrast of a vector is a measure for its statistically independence this is achieved by maximization of an appropriate contrast criterion. Before proposing such a criterion (see [13]), we will give a heuristic motivation for a certain contrast function. Thereafter, at the end of this section, we shall show that this function is indeed a contrast.

But first we start with some notes on the standardization of random vectors and a formal definition of *contrast function*.

Definition 3.3 ([13])**.** We denote by

i) $\mathbb{E}^d$ the space of random vectors or multivariate random variables with values in $\mathbb{R}^d$.

ii) $\mathbb{E}_r^d$ the Euclidean subspace of $\mathbb{E}^d$ spanned by variables with finite moments up to order r, for any $r \geq 2$, provided with the inner product $\langle \mathcal{X}, \mathcal{Y} \rangle = E(\mathcal{X}^T \mathcal{Y})$.

iii) $\tilde{\mathbb{E}}_2^d$ the subset of $\mathbb{E}_2^d$ of variables having an invertible covariance matrix.

In the above definition the d^r moments of order r of $\mathcal{X}$ are defined by

$$\mu_{\mathcal{X}}(r_1, \ldots, r_d) = E\left(\prod_{i=1}^{d} \mathcal{X}_i^{r_i}\right),$$

where $r_1 + r_2 + \cdots + r_d = r$. In general the observed data corresponds to a random vector which lies in $\mathbb{E}_2^d$ (note that $\mathbb{E}_r^d \subseteq \mathbb{E}_2^d$).

If we talk about standardization, we refer to transforming a random vector $\mathcal{X} \in \mathbb{E}_r^d$ into another, denoted by $\tilde{\mathcal{X}}$, that has a unit covariance. In particular if the covariance $C_{\mathcal{X}}$ of $\mathcal{X}$ is not invertible, the length of $\mathcal{X}$ and $\tilde{\mathcal{X}}$ cannot be the same because $C_{\mathcal{X}}$ has at least one zero eigenvalue. In this case we have to project $\mathcal{X}$ on the range space of $C_{\mathcal{X}}$. The projection and standardization can be done by PCA.

Without loss of generality we may assume in the following that the observed variable belongs to $\tilde{\mathbb{E}}_2^d$. For a complete discussion, see [13].

Definition 3.4 ([13])**.** A *contrast function* is a mapping Ψ from the set of probability densities $\{p_{\mathcal{X}}, \mathcal{X} \in \mathbb{E}^d\}$ to $\mathbb{R}$

$$\Psi : \{p_{\mathcal{X}}, \mathcal{X} \in \mathbb{E}^d\} \longrightarrow \mathbb{R}$$

satisfying the following three requirements

i) $\Psi(p_{P\mathcal{X}}) = \Psi(p_{\mathcal{X}})$ for all permutations P, i.e. $\Psi(p_{\mathcal{X}})$ does not change if the components $\mathcal{X}_i$ are permuted.

ii) $\Psi(p_{\Delta\mathcal{X}}) = \Psi(p_{\mathcal{X}})$ for all invertible diagonal matrices Δ, i.e. Ψ is invariant by 'scale' change.

iii) If $\mathcal{X}$ has independent components, then $\Psi(p_{A\mathcal{X}}) \leq \Psi(p_{\mathcal{X}})$ for all invertible matrices A.

We see that the contrast of a random vector $\mathcal{X}$ is maximal if its components are statistically independent. Therefore, we look for an appropriate contrast function. But first we will briefly recall the definition of statistical independence.

Definition 3.5. The components of a random vector $\mathcal{X} = (\mathcal{X}_1, \ldots, \mathcal{X}_d)^T$ with probability density function $p_{\mathcal{X}}(x)$ are *mutually* or *statistically independent* if and only if for the joint density function holds

$$p_{\mathcal{X}}(x) = \prod_{i=1}^{d} p_{\mathcal{X}_i}(x_i).$$

This definition provides a natural way of measuring the degree of statistical independence of the components of a random vector by comparing the joint density $p_{\mathcal{X}}$ and the marginal densities $p_{\mathcal{X}_i}$. In other words we search a distance measure δ for density functions:

$$\delta\left(p_{\mathcal{X}}, \prod_{i=1}^{d} p_{\mathcal{X}_i}\right).$$

In this context, we do not always talk about proper distances since some are not symmetric as for example the *Kullback-Leibler divergence*. This divergence was introduced by Kullback and Leibler [32] in 1951.

Definition 3.6. The *Kullback-Leibler divergence* is defined as

$$\delta(p_{\mathcal{X}}, p_{\mathcal{Z}}) = \int p_{\mathcal{X}}(x) \ln\left(\frac{p_{\mathcal{X}}(x)}{p_{\mathcal{Z}}(x)}\right) \mathrm{d}x.$$

Lemma 3.1. *The Kullback-Leibler divergence satisfies*

$$\delta(p_{\mathcal{X}}, p_{\mathcal{Z}}) \geq 0,$$

where equality is obtained if and only if $p_{\mathcal{X}} = p_{\mathcal{Z}}$ almost everywhere.

Proof. For the proof we use $\ln(x) \leq x - 1$, which follows from the concavity of the logarithm, and the property of density functions $\int p(x)\mathrm{d}x = 1$:

$$\begin{aligned}
\int p_{\mathcal{X}}(x) \ln\left(\frac{p_{\mathcal{X}}(x)}{p_{\mathcal{Z}}(x)}\right) \mathrm{d}x &= -\int p_{\mathcal{X}}(x) \ln\left(\frac{p_{\mathcal{Z}}(x)}{p_{\mathcal{X}}(x)}\right) \mathrm{d}x \\
&\geq \int p_{\mathcal{X}}(x) \left(1 - \frac{p_{\mathcal{Z}}(x)}{p_{\mathcal{X}}(x)}\right) \mathrm{d}x \\
&= \left(-\int p_{\mathcal{Z}}(x)\mathrm{d}x + \int p_{\mathcal{X}}(x)\mathrm{d}x\right) \\
&= -1 + 1 = 0.
\end{aligned}$$

For the equality to hold we require

$$\frac{p_{\mathcal{Z}}}{p_{\mathcal{X}}} = 1 \qquad \text{almost everywhere,}$$

such that $\ln(x) = x - 1$, i.e. $p_{\mathcal{Z}}(x) = p_{\mathcal{X}}(x)$. □

As we have stated before we want $\mathcal{X}$ to be as statistically independent as possible. From Definition 3.5 it is clear that choosing $p_{\mathcal{X}}$ such that the distance between $p_{\mathcal{X}}$ and $\prod_{i=1}^{d} p_{\mathcal{X}_i}$ is minimized gives the requested result. This awareness leads us to take a closer look at the specific Kullback-Leibler distance

$$I(p_{\mathcal{X}}) = \delta(p_{\mathcal{X}}, \prod_{i=1}^{d} p_{\mathcal{X}_i}) = \int p_{\mathcal{X}}(x) \ln\left(\frac{p_{\mathcal{X}}(x)}{\prod_{i=1}^{d} p_{\mathcal{X}_i}(x_i)}\right) \mathrm{d}x, \tag{3.2}$$

which is called *mutual information*. The key problem in the analysis of the mutual information of a random variable is the computation of the density $p_{\mathcal{X}}$ itself. Since it is usually unknown further consideration is needed. In this context a useful definition is the one of *differential entropy*.

Definition 3.7. The *differential entropy* of a random variable $\mathcal{X}$ is defined as

$$S(p_{\mathcal{X}}) = -\int p_{\mathcal{X}}(x) \ln p_{\mathcal{X}}(x)\mathrm{d}x.$$

The differential entropy, also known as continuous entropy, is a measure of the average information content of a random variable. While the well-known Shannon entropy is only defined for discrete random variables, the differential entropy is a concept for continuous ones. Hence it can be interpreted as an extension of the Shannon entropy even though it has not all the nice properties, as for example, invariance under linear invertible changes of coordinates. Nevertheless we would like to express (3.2) using entropy terms. But first we recall that among all densities of $\tilde{\mathbb{E}}_2^d$ with given covariance $C_\mathcal{X}$ the multivariate Gaussian density given by

$$\phi_\mathcal{X}(x) = \frac{1}{(2\pi)^{d/2}\det(C_\mathcal{X})^{1/2}} \exp\left(-\frac{1}{2}x^T C_\mathcal{X}^{-1} x\right),$$

has the largest entropy.

Lemma 3.2. *For all $\mathcal{X}, \mathcal{Y} \in \tilde{\mathbb{E}}_2^d$ with given covariance $C_\mathcal{X}$ we have*

$$S(\phi_\mathcal{X}) \geq S(p_\mathcal{Y}),$$

with equality if and only if $\phi_\mathcal{X} = p_\mathcal{Y}$ almost everywhere. Furthermore it holds

$$S(\phi_\mathcal{X}) = \frac{1}{2}\left(d + d\ln(2\pi) + \ln(\det(C_\mathcal{X}))\right).$$

Proof. From Lemma 3.1 we derive for $p_\mathcal{Y}$ and $\phi_\mathcal{X}$

$$0 \leq \int p_\mathcal{Y}(x)\ln\left(\frac{p_\mathcal{Y}(x)}{\phi_\mathcal{X}(x)}\right)\mathrm{d}x = \int p_\mathcal{Y}(x)\ln\left(p_\mathcal{Y}(x)\right)\mathrm{d}x - \int p_\mathcal{Y}(x)\ln(\phi_\mathcal{X}(x))\mathrm{d}x,$$

which leads to

$$-\int p_\mathcal{Y}(x)\ln(p_\mathcal{Y}(x))\mathrm{d}x \leq -\int p_\mathcal{Y}(x)\ln(\phi_\mathcal{X}(x))\mathrm{d}x \tag{3.3}$$

with equality if and only if $p_\mathcal{X} = \phi_\mathcal{X}$ almost everywhere. Using the definition of the density function of the multivariate Gaussian distribution the right-hand side turns into

$$\begin{aligned}
-\int p_\mathcal{Y}(x)\ln(\phi_\mathcal{X}(x))\mathrm{d}x &= -\int p_\mathcal{Y}(x)\left(\ln\left(\frac{1}{(2\pi)^{d/2}\det(C_\mathcal{X})^{1/2}}\right) - \frac{x^T C_\mathcal{X}^{-1}x}{2}\right) \\
&= \ln\left((2\pi)^{d/2}\det(C_\mathcal{X})^{1/2}\right)\int p_\mathcal{Y}(x)\mathrm{d}x + \frac{1}{2}\int p_\mathcal{Y}(x)x^T C_\mathcal{X}^{-1}x\mathrm{d}x \\
&= \frac{1}{2}\left(d\ln(2\pi) + \ln(\det(C_\mathcal{X}))\right) + \frac{1}{2}\int p_\mathcal{Y}(x)\sum_{i,j=1}^{d} x_i (C_\mathcal{X}^{-1})_{ij} x_j \mathrm{d}x \\
&= \frac{1}{2}\left(d\ln(2\pi) + \ln(\det(C_\mathcal{X}))\right) + \frac{1}{2}\sum_{i,j=1}^{d}(C_\mathcal{X}^{-1})_{ij}\int p_\mathcal{Y}(x)x_i x_j \mathrm{d}x \\
&= \frac{1}{2}\left(d\ln(2\pi) + \ln(\det(C_\mathcal{X}))\right) + \frac{1}{2}\sum_{i,j=1}^{d}(C_\mathcal{X}^{-1})_{ij}(C_\mathcal{X})_{ji}
\end{aligned}$$

$$= \frac{1}{2}\left(d\ln(2\pi) + \ln(\det(C_{\mathcal{X}}))\right) + \frac{1}{2}\sum_{i=1}^{d}(C_{\mathcal{X}}^{-1}C_{\mathcal{X}})_{ii}$$
$$= \frac{1}{2}\left(d\ln(2\pi) + \ln(\det(C_{\mathcal{X}})) + d\right).$$

For this computation we used $\int p_{\mathcal{Y}}(x)\mathrm{d}x = 1$ and $\int p_{\mathcal{Y}}(x)x_i x_j \mathrm{d}x = (C_{\mathcal{X}})_{ij} = (C_{\mathcal{X}})_{ji}$. Hence we get

$$-\int p_{\mathcal{Y}}(x)\ln(\phi_{\mathcal{X}}(x))\mathrm{d}x = \frac{1}{2}\left(d\ln(2\pi) + \ln(\det(C_{\mathcal{X}})) + d\right). \tag{3.4}$$

It is noteworthy, that the expression on the right-hand side does not depend on $p_{\mathcal{Y}}$ and thus it follows $S(\phi_{\mathcal{X}}) = \frac{1}{2}\left(d\ln(2\pi) + \ln(\det(C_{\mathcal{X}})) + d\right)$. From (3.3) and (3.4) we deduce

$$-\int p_{\mathcal{Y}}(x)\ln(p_{\mathcal{Y}}(x))\mathrm{d}x \leq \frac{1}{2}\left(d\ln(2\pi) + \ln(\det(C_{\mathcal{X}})) + d\right)$$

with equality if and only if $p_{\mathcal{X}} = \phi_{\mathcal{X}}$ almost everywhere. □

Given this property one can define a distance for density functions to the Gaussian density, but this is not a distance in the strict sense.

Definition 3.8. For densities $p_{\mathcal{X}} \in \tilde{\mathbb{E}}_2^d$ we define the *negentropy* as

$$J(p_{\mathcal{X}}) = S(\phi_{\mathcal{X}}) - S(p_{\mathcal{X}}),$$

with $\phi_{\mathcal{X}}$ being the Gaussian density with the same variance and mean as $p_{\mathcal{X}}$.

Theorem 3.1. *The negentropy $J : \tilde{\mathbb{E}}_2^d \to \mathbb{R}$ has the following properties:*

i) $J(p_{\mathcal{X}}) \geq 0$ *for all* $\mathcal{X} \in \tilde{\mathbb{E}}_2^d$.

ii) $J(p_{A\mathcal{X}}) = J(p_{\mathcal{X}})$ *for all invertible matrices A, i.e. the negentropy is invariant under linear invertible changes of coordinates.*

iii) $J(p_{\mathcal{X}}) = 0$ *if and only if* $p_{\mathcal{X}} = \phi_{\mathcal{X}}$ *almost everywhere.*

Proof. Properties i) and iii) follow from Lemma 3.2. To show ii) we consider the entropy of an arbitrary density $q_{A\mathcal{X}}$. From the transformation formula for density functions follows:

$$\begin{aligned}
S(q_{A\mathcal{X}}) &= -\int q_{A\mathcal{X}}(x)\ln\left(q_{A\mathcal{X}}(x)\right)\mathrm{d}x \\
&= -\int q_{\mathcal{X}}(A^{-1}x)|\det(A^{-1})|\ln\left(q_{\mathcal{X}}(A^{-1}x)|\det(A^{-1})|\right)\mathrm{d}x \\
&= -\int q_{\mathcal{X}}(y)\ln\left(q_{\mathcal{X}}(y)|\det(A^{-1})|\right)\mathrm{d}y \\
&= -\int q_{\mathcal{X}}(y)\ln\left(q_{\mathcal{X}}(y)\right)\mathrm{d}y + \int q_{\mathcal{X}}(y)\ln\left(|\det(A)|\right)\mathrm{d}y \\
&= S(q_{\mathcal{X}}) - \ln\left(|\det(A)|\right).
\end{aligned}$$

Hence we get for the negentropy of $p_{A\mathcal{X}}$

$$\begin{aligned} J(p_{A\mathcal{X}}) &= S(\phi_{A\mathcal{X}}) - S(p_{A\mathcal{X}}) \\ &= S(\phi_{\mathcal{X}}) - \ln(|\det(A)|) - S(p_{\mathcal{X}}) + \ln(|\det(A)|) \\ &= J(p_{\mathcal{X}}). \end{aligned}$$

□

Theorem 3.2. *For the mutual information I the following*

$$I(p_{\mathcal{X}}) = J(p_{\mathcal{X}}) - \sum_{i=1}^{d} J(p_{\mathcal{X}_i}) + \frac{1}{2} \ln \left(\frac{\prod_{i=1}^{d} (C_{\mathcal{X}})_{ii}}{\det(C_{\mathcal{X}})} \right) \tag{3.5}$$

holds.

Before we start with the proof we recall the definition of *marginal density.*

Definition 3.9. The *ith marginal density function* $p_{\mathcal{X}_i}$ of a random variable $\mathcal{X} = (\mathcal{X}_1, \ldots, \mathcal{X}_d)^T$ is the probability density function associated to variable $\mathcal{X}_i$. It is related to $p_{\mathcal{X}}$ as follows

$$p_{\mathcal{X}_i}(x_i) = \int p_{\mathcal{X}}(x) \mathrm{d}x_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_d.$$

Proof of Theorem 3.2. [13]. Using Fubini's Theorem we deduce from Definition 3.9

$$\begin{aligned} -\int p_{\mathcal{X}}(x) \ln\left(p_{\mathcal{X}_i}(x_i)\right) \mathrm{d}x &= -\int \ln\left(p_{\mathcal{X}_i}(x_i)\right) \int p_{\mathcal{X}}(x) \mathrm{d}x_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_d \mathrm{d}x_i \\ &= -\int \ln\left(p_{\mathcal{X}_i}(x_i)\right) p_{\mathcal{X}_i}(x_i) \mathrm{d}x_i \\ &= S(p_{\mathcal{X}_i}), \end{aligned}$$

for all i. Thus,

$$\begin{aligned} -S(p_{\mathcal{X}}) + \sum_{i=1}^{d} S(p_{\mathcal{X}_i}) &= \int p_{\mathcal{X}}(x) \ln(p_{\mathcal{X}}(x)) \, \mathrm{d}x - \sum_{i=1}^{d} \int p_{\mathcal{X}}(x) \ln\left(p_{\mathcal{X}_i}(x_i)\right) \mathrm{d}x_i \\ &= \int p_{\mathcal{X}}(x) \ln(p_{\mathcal{X}}(x)) \, \mathrm{d}x - \int p_{\mathcal{X}}(x) \ln\left(\prod_{i=1}^{d} p_{\mathcal{X}_i}(x_i)\right) \mathrm{d}x \\ &= \int p_{\mathcal{X}}(x) \ln\left(\frac{p_{\mathcal{X}}(x)}{\prod_{i=1}^{d} p_{\mathcal{X}_i}(x_i)}\right) \mathrm{d}x \\ &= I(p_{\mathcal{X}}) \end{aligned}$$

holds and therefore using Lemma 3.2 it follows

$$J(p_{\mathcal{X}}) - \sum_{i=1}^{d} J(p_{\mathcal{X}_i}) = S(\phi_{\mathcal{X}}) - S(p_{\mathcal{X}}) - \sum_{i=1}^{d} S(\phi_{\mathcal{X}_i}) + \sum_{i=1}^{d} S(p_{\mathcal{X}_i})$$

$$
\begin{aligned}
&= I(p_{\mathcal{X}}) + \frac{1}{2}\left(d + d\ln(2\pi) + \ln\det(C_{\mathcal{X}})\right) \\
&\quad - \frac{1}{2}\sum_{i=1}^{d}\left(1 + \ln(2\pi) + \ln((C_{\mathcal{X}})_{ii})\right) \\
&= I(p_{\mathcal{X}}) - \frac{1}{2}\ln\left(\frac{\prod_{i=1}^{d}(C_{\mathcal{X}})_{ii}}{\det(C_{\mathcal{X}})}\right).
\end{aligned}
$$

Solving this equation for $I(p_{\mathcal{X}})$ leads to the required result. □

Lemma 3.3. *If $C_{\mathcal{X}}$ is invertible, the last term of* (3.5)*, namely $\frac{1}{2}\ln\left(\frac{\prod_{i=1}^{d}(C_{\mathcal{X}})_{ii}}{\det(C_{\mathcal{X}})}\right)$, is zero if and only if $C_{\mathcal{X}}$ is diagonal.*

Proof. It is obvious that the last term of (3.5) is zero if and only if $\det(C_{\mathcal{X}}) = \prod_{i=1}^{d}(C_{\mathcal{X}})_{ii}$.
'⇒': If $\det(C_{\mathcal{X}}) = \prod_{i=1}^{d}(C_{\mathcal{X}})_{ii}$ holds, we replace $C_{\mathcal{X}}$ by its Cholesky decomposition $C_{\mathcal{X}} = LL^T$, where L is a lower triangular matrix. Then the equation turns into

$$
\prod_{i=1}^{d} L_{ii}^2 = \det(L)^2 = \det(C_{\mathcal{X}}) = \prod_{i=1}^{d}(C_{\mathcal{X}})_{ii} = \prod_{i=1}^{d}\left(L_{ii}^2 + \sum_{k=1}^{i-1} L_{ik}^2\right).
$$

But this implies either

$$
\sum_{k=1}^{i-1} L_{ik}^2 = 0, \quad \text{for all } i
$$

or

$$
L_{ii}^2 + \sum_{k=1}^{i-1} L_{ik}^2 = 0 \quad \text{for at least one } k,
$$

which results in either all L_{ik} being zero or L having some zero-row. Since $C_{\mathcal{X}}$ is invertible, only the first is possible and thus $C_{\mathcal{X}}$ is diagonal.
'⇐': If $C_{\mathcal{X}}$ is diagonal, then it follows immediately that $\det(C_{\mathcal{X}}) = \prod_{i=1}^{d}(C_{\mathcal{X}})_{ii}$ holds. □

Now we are able to specify a contrast criterion.

Theorem 3.3 ([13])**.** *The function*

$$
\Psi(p_{\mathcal{X}}) = -I(p_{\tilde{\mathcal{X}}})
$$

is a contrast function over $\mathbb{E}_2^d$.

Remark 16. *Note that the definition of Ψ includes the standardization of the random vector $\mathcal{X}$. Therefore, the third term of the contrast function (see Equation* (3.5)*) cancels out due to Lemma 3.3.*

Proof of Theorem 3.3. We have to verify the three requirements of Definition 3.4. Since i) and ii) consist of special linear invertible changes of coordinates, we only have to prove that the third term of the right-hand side of (3.5) does not change under this

kind of transformations. But this term cancels because of the standardization in the definition of Ψ.
To prove iii) let us suppose $\mathcal{X}$ to have independent components, i.e. $\Psi(p_{\mathcal{X}}) = 0$. From the definition of Ψ it follows that $\Psi(p_{A\mathcal{X}}) \leq 0$ and thus $\Psi(p_{A\mathcal{X}}) \leq \Psi(p_{\mathcal{X}})$. □

If we return to the initial problem (3.1), we have to minimize the mutual information $I(p_{\tilde{\mathcal{S}}}) = I(p_{\widetilde{G\mathcal{Y}}})$ (or to maximize $-I$). It is convenient to do this in two steps. First we apply a transformation T in order to standardize $\mathcal{Y}$. The second step consists in finding an orthonormal transform Q that minimizes the second term of (3.5) while the other two remain constant. Here orthonormal refers to the orthonormality of the rows of Q. Then $\widetilde{G\mathcal{Y}}$ factorizes in $\widetilde{G\mathcal{Y}} = QT\mathcal{Y}$.

What we have figured out so far is a way to write the mutual information I such that it does not depend directly on the unknown densities $p_{\tilde{\mathcal{S}}}$ and $p_{\tilde{\mathcal{S}}_i}$ but rather on the distances $J(p_{\tilde{\mathcal{S}}})$ and $J(p_{\tilde{\mathcal{S}}_i})$ of these densities to the Gaussian density (see (3.5)). The benefits of this results may not be obvious but they become clear if we recapitulate the aim of this chapter: For a given data set Y we are looking for a $\mathcal{S}$ with almost independent components. As we have seen before this can be achieved by minimizing the distance between $p_{\tilde{\mathcal{S}}}$ and its marginal densities, which is equivalent to minimizing $I(p_{\tilde{\mathcal{S}}})$. This minimum is zero and it can be reached if $p_{\tilde{\mathcal{S}}} = \phi_{\tilde{\mathcal{S}}}$. In this case also the negentropy J reaches its minimum zero. Therefore, we would like to approximate the negentropy about zero in order to approximate I. This leads to expanding the unknown density $p_{\tilde{\mathcal{S}}}$ in the neighborhood of $\phi_{\tilde{\mathcal{S}}}$. This expansion is the core aspect of the next section.

3.2. Edgeworth expansion

Assume for simplicity that A and G are quadratic, i.e. $\rho = d$. For our application this assumption will be no restriction, since we consider dimensionality reduced data sets. For a given standardized random vector $\tilde{\mathcal{Y}} = T\mathcal{Y}$ we are looking for an orthonormal matrix Q maximizing the contrast function

$$\Psi(p_{\tilde{\mathcal{S}}}) = -I(p_{Q\tilde{\mathcal{Y}}}), \text{ where } \tilde{\mathcal{S}} = Q\tilde{\mathcal{Y}}, \tag{3.6}$$

i.e. Q is minimizing the mutual information I. The factorization of $\tilde{\mathcal{S}}$ follows from

$$\tilde{\mathcal{S}} = \Delta^{-1}\mathcal{S} = \Delta^{-1}G\mathcal{Y} = \Delta^{-1}GT^{-1}T\mathcal{Y} = \Delta^{-1}GT^{-1}\tilde{\mathcal{Y}} = Q\tilde{\mathcal{Y}}.$$

The orthonormality of $Q = \Delta^{-1}GT^{-1}$ follows with Equation (2.2) from:

$$I_\rho = C_{\tilde{\mathcal{S}}} = C_{Q\tilde{\mathcal{Y}}} = QC_{\tilde{\mathcal{Y}}}Q^T = QQ^T,$$

where I_ρ is the unit matrix. As we have stated before, the densities $p_{\tilde{\mathcal{Y}}}$ and $p_{\tilde{\mathcal{S}}}$ are unknown. For this reason the maximization task cannot be solved directly. Since cumulants are more easily accessible, we aim to express the contrast function as a function of those.

The contrast function from Theorem 3.3 consists of three main terms: the negentropy of a standardized random vector, the marginal negentropy of each component of this random vector and an additional term involving the covariance matrix. As the random vector is standardized, the last term cancels and thus we only have to consider the two remaining ones.

First we discuss the expression of negentropy in the marginal case, i.e. the scalar case. The expression relies basically on the Edgeworth expansion of a density about its best Gaussian approximate (in our case with zero-mean and unit variance).

Definition 3.10. The *Edgeworth expansion* of a probability density function $p_{\mathcal{X}}$ of a random variable $\mathcal{X}$ having zero-mean and unit variance is given by

$$\begin{aligned}\frac{p_{\mathcal{X}}(x)}{\phi_{\mathcal{X}}(x)} &= 1 \\ &+ \frac{1}{3!}\kappa_3 h_3(x) \\ &+ \frac{1}{4!}\kappa_4 h_4(x) + \frac{10}{6!}\kappa_3^2 h_6(x) \\ &+ \frac{1}{5!}\kappa_5 h_5(x) + \frac{35}{7!}\kappa_3\kappa_4 h_7(x) + \frac{280}{9!}\kappa_3^3 h_9(x) \\ &+ \frac{1}{6!}\kappa_6 h_6(x) + \frac{56}{8!}\kappa_3\kappa_5 h_8(x) + \frac{35}{8!}\kappa_4^2 h_8(x) + \frac{2100}{10!}\kappa_3^2\kappa_4 h_{10}(x) + \frac{15400}{12!}\kappa_3^4 h_{12}(x) \\ &+ o(m^{-2}),\end{aligned}$$

where κ_i denotes the cumulant of order i of $\mathcal{X}$ and $h_i(x)$ is the Hermit polynomial of degree i (see [13]).

Remark 17. *From the Central Limit Theorem it follows that for a random variable $\mathcal{X}$, being a sum of m independent random variables with finite cumulants, the cumulant κ_i is of order $m^{(2-i)/2}$. For a derivation of the Edgeworth expansion see [31].*

Since $\tilde{\mathcal{Y}}$ is the observed variable, we have so far no information about $\tilde{\mathcal{S}}$ and thus we cannot approximate the cumulants κ_i but only the cumulants γ_i of $\tilde{\mathcal{Y}}$. But cumulants satisfy a certain multilinearity property which allows us to compute the κ_i using (3.6), see [13].

Theorem 3.4 ([13])**.** *For a standardized scalar random variable $\tilde{\mathcal{X}}$, the negentropy can be expanded as*

$$J(p_{\tilde{\mathcal{X}}}) = \frac{1}{12}\kappa_3^2 + \frac{1}{48}\kappa_4^2 + \frac{7}{48}\kappa_3^4 - \frac{1}{8}\kappa_3^2\kappa_4 + o(m^{-2}).$$

Proof. We will only discuss the idea of the proof because it ends up in solving a couple of ordinary integrals.
From the power series expansion of $\ln(1+r)$ it follows

$$(1+r)\ln(1+r) = r + \frac{r^2}{2} - \frac{r^3}{6} + \frac{r^4}{12} + o(r^4). \tag{3.7}$$

From Definition 3.10 we deduce

$$p_{\tilde{\mathcal{X}}}(x) = \phi_{\tilde{\mathcal{X}}}(x)\,(1 + r(x))$$

and from (3.4) $\int \phi_{\tilde{\mathcal{X}}}(x)(1 + r(x)) \ln\left(\phi_{\tilde{\mathcal{X}}}(x)\right) \mathrm{d}x = -S(\phi_{\tilde{\mathcal{X}}})$. Using these equalities the negentropy can be written as

$$J(p_{\tilde{\mathcal{X}}}) = \int \phi_{\tilde{\mathcal{X}}}(x)\,(1 + r(x)) \ln(1 + r(x))\mathrm{d}x$$

and we can insert (3.7) and the definition of $r(x)$. This leads to a sum of integrals which can be solved by using some properties of the Hermite polynomials. For more details see [13]. □

Theorem 3.4 provides an approximation of the marginal negentropies of $\tilde{\mathcal{S}}$. Now we can address the modification of $J(p_{\tilde{\mathcal{S}}})$. From Theorem 3.1 we know, that the negentropy is invariant under linear invertible changes of coordinates. Hence $J(p_{\tilde{\mathcal{S}}}) = J(p_{\tilde{\mathcal{Y}}})$ and thus

$$I(p_{\tilde{\mathcal{S}}}) \approx J(p_{\tilde{\mathcal{Y}}}) - \frac{1}{48} \sum_{i=1}^{d} 4\kappa(i)_3^2 + \kappa(i)_4^2 + 7\kappa(i)_3^4 - 6\kappa(i)_3^2\kappa(i)_4,$$

where $\kappa(i)_j$ denotes the jth cumulant of $\tilde{\mathcal{S}}_i$. Since $J(p_{\tilde{\mathcal{Y}}})$ does not depend on Q, the next theorem follows immediately.

Theorem 3.5. *Maximizing* (3.6) *is equivalent to maximizing*

$$\psi(Q) = \sum_{i=1}^{d} 4\kappa(i)_3^2 + \kappa(i)_4^2 + 7\kappa(i)_3^4 - 6\kappa(i)_3^2\kappa(i)_4, \tag{3.8}$$

where $\kappa(i)_j$ denotes the jth cumulant of $\tilde{\mathcal{S}}_i$.

The maximization of ψ is not a trivial task as the cumulants $\kappa(i)_j$ depend on Q. Even if we have shown that the mutual information I is a contrast function this does not imply that ψ is so, because ψ is only an approximation of I. If the cumulants $\kappa(i)_3$ are large enough it is sufficient to consider the expansion only up to order $\mathcal{O}(m^{-3/2})$, which yields $\psi(Q) = 4\sum_{i=1}^{d} \kappa(i)_3^2$. And if $\kappa(i)_3 = 0$ for all i Equation (3.8) reduces to $\psi(Q) = \sum_{i=1}^{d} \kappa(i)_4^2$. It can be shown that in these two special cases $\psi(Q)$ is a contrast function. For a proof see [13] or [12]. In the general case where the $\kappa(i)_3$ are neither all null nor large, this functions can also be used, but they are not approximating the contrast function from Theorem 3.3 any more. However, there are other criteria discussed in [9].

3.3. Algorithm

So far we have provided a guideline for solving the initial problem (3.1) of this section. For the purpose of applying these steps to a measured data set we seek an efficient

algorithm. Even though in [13] an algorithm is discussed, we use the *Joint Approximate Diagonalization of Eigenmatrices* (JADE) proposed by Cardoso in [9]. Cardoso and Souloumiac developed this algorithm in 1993 and it has been improved several times by the authors. A MATLAB implementation of JADE is available at [6].

The general proceeding of JADE is similar to what we have described in the previous sections. Recall that we are searching a matrix A such that $Y = AS$. This can be done following the steps below [9].

1. *Initialize*: Standardization of the data set $\tilde{Y} = TY$.
2. *Form statistics*: Estimation of the cumulants of the components of $\tilde{Y}$.
3. *Optimize an orthogonal contrast*: Estimation of Q.
4. *Separate*: Estimate $\tilde{A}$ as the inverse of $\tilde{G} = QT$ and/or $\tilde{S} = QTY$.

The output of the JADE algorithm is a data set $\tilde{S}$ with unit covariance. This is a consequence from a different unity condition which leads to the choice of a different representative. To get the ICA we defined in Definition 3.1 and 3.2, we need to scale and permute the columns of $\tilde{A}$.

4. ISA - Independent Subspace Analysis

In the previous section we have seen that ICA is based on the assumption that the number of sources is known and that the source signals are statistically independent. In practice this is not always the case and therefore the extraction of sources of a data set Y might be inaccurate. As a consequence it could happen that we detect more independent components as the true number of sources. In this case, two or more of the separated components pertain to the same source. Nevertheless there is a way to use the separation properties of ICA: Independent Subspace Analysis (ISA) takes advantage of this quality of ICA by extending it to a method which extracts maximally contrasting sources. These sources are combinations of the independent components obtained from a single channel mixture (see [10]). In blind signal separation ISA has been proposed from Casey and Westner [10] in 2000 and it is an upgrade of ICA which partitions the different independent components into groups, each of which is spanning a subspace. This procedure avoids the above explained problem of extracting more sources than there are. There has been previous work on ISA in the context of image processing by Hyvärinen and Hoyer [28].

Remark 18. *Usually ICA based separation requires that the given data set contains at least as many observations as there are unknown components or sources ($d \geq \rho$). In the field of blind signal separation (BSS) this leads to the constraint that there need to be at least as many sensors recording the signal as unknown sources. But in practice there are often fewer sensors or even only one, such that this constraint is very limiting. Therefore, the idea of ISA is to extend ICA not only in the sense of grouping the components into multi-component subspaces but also in the sense of the ability to handle single sensor problems. This is done considering the spectrogram of a signal (see Remark 5) instead of the signal itself. Applying the subspace analysis explained below leads to a decomposition of the spectrogram into spectrograms of the unknown sources. These sources can be obtained by back transforming each spectrogram. In this section we will only discuss the decomposition step. For the remaining steps see Chapter 7.*

The general proceeding in ISA is to first extract the independent components of a given data set Y using ICA. In a second step these components are grouped (or partitioned) into independent subspaces, each one corresponding to a source. Finally the sources are reconstructed from these multi-component subspaces. In the following we aim to discuss first the third and then the second step of this method.

4.1. Reconstruction

For a given data set $Y = (y_1^T, \ldots, y_d^T)^T \in \mathbb{R}^{d \times n}$ we suppose as before each row $y_i \in \mathbb{R}^{1 \times n}$ to be the weighted sum of ρ independent components $z_j \in \mathbb{R}^{n \times 1}$:

$$y_i^T = \sum_{j=1}^{\rho} a_{ji} z_j$$

or

$$y_i^T = Z a_i, \tag{4.1}$$

where $Z = (z_1, \ldots, z_\rho)$ and $A = (a_1, \ldots, a_d) = (a_{ji})_{j=1,\ldots,\rho, i=1,\ldots,d}$. The unknown matrices Z and A can be estimated with ICA. Due to the properties of ICA, Z has unit covariance and thus Z is an orthonormal basis of a ρ-dimensional subspace of the $\mathbb{R}^n$. In contrast to ICA at this point ISA does not assume the z_j to be the sources of the mixed signal.

The core idea of ISA is that each source is a linear combination of the z_j. Assume that we have c unknown sources and that each z_j corresponds to only one of the different sources such that the ρ-dimensional subspace U spanned by the z_j is the internal direct sum of subspaces U_k associated to the sources. Hence we get a partition of Z

$$Z = \bigcup_{j=1}^{c} Z_k, \quad Z_k \cap Z_j = \emptyset \text{ for all } 1 \leq k, j \leq c, k \neq j.$$

Each of the $Z_k = (z(k)_1, z(k)_2, \ldots, z(k)_{\rho_k})$ spans an independent subspace $U_k = \text{span}(z(k)_1, z(k)_2, \ldots, z(k)_{\rho_k})$ and thus

$$U = \oplus_{k=1}^{c} U_k. \tag{4.2}$$

Theorem 4.1. *Let Z be a set of orthonormal vectors of $\mathbb{R}^n$ which is partitioned into subsets Z_k, $k = 1, \ldots, c$, then a given data set $Y \in \mathbb{R}^{d \times n}$ with $y_i^T \in \text{span}(Z)$ can be decomposed into separate data sets Y_k formed from the subspaces $\text{span}(Z_k)$. This decomposition can be written as*

$$Y^T = \sum_{k=1}^{c} Y_k^T = \sum_{k=1}^{c} Z_k A_k,$$

where $A_k = Z_k^T Y^T$ is the matrix of coefficients.

Proof. From (4.1) follows $y_i^T \in U = \text{span}(Z)$ and thus from (4.2) there exist coefficient vectors $a(k)_i \in \mathbb{R}^{\rho_k}$ such that

$$y_i^T = \sum_{k=1}^{c} Z_k a(k)_i.$$

Extending this consideration to the whole data set Y yields

$$Y^T = \sum_{k=1}^{c} Z_k A_k, \tag{4.3}$$

where $A_k = (a(k)_1, \ldots, a(k)_d) \in \mathbb{R}^{\rho_k \times d}$.

The coefficient matrices A_k can be computed by orthogonal projection of the data set Y^T on the kth subspace. This projection is given by multiplying 4.3 from the left with Z_k^T as

$$A_k = Z_k^T Y^T.$$

□

From Theorem 4.1 we have learned that, given the subspaces spanned by Z_k, we can decompose the data set Y^T into a sum where each subspace appears as a weighted sum of its basis vectors.

Remark 19. *From Theorem 4.1 we get a decomposition of each data point* $y_i^T = \sum_{k=1}^{c} Z_k a(k)_k$ *into parts each corresponding to a source.*

In this section we analyzed how to reconstruct the data sets Y_k from Y, where Y_k represents the kth source, provided an adequate partition of Z has been found. In the next section we present a method to compute such a partition.

4.2. Grouping

The main difficulty in the concept of ISA is to identify the components z_j that belong to the same multi-component subspace. This can be done by some type of grouping. Here, we like to discuss the grouping method introduced by Casey and Westner in [10]. This method is based on calculating the similarities of the independent components z_j and sorting them by using their pairwise dissimilarities. To understand this concept we return once more to the world of stochastics. In the next section we will define a similarity measure which is the basis of the clustering done in the last section.

4.2.1. Similarity measure

A similarity measure quantifies the similarity of two objects. If we consider for example the Euclidean distance between two points, we state that the smaller the distance the more similar the points are in terms of location. In Section 3.1 we have seen that each z_j can be interpreted as n realizations of a random variable $\mathcal{Z}_j$. If we search a similarity measure for the z_j it seems reasonable to take a distance measure for density functions in order to compare $p_{\mathcal{Z}_i}$ and $p_{\mathcal{Z}_j}$. As an essential property of this measure we require its symmetry. In Definition 3.6 we introduced the Kullback-Leibler divergence. This distance measure is not symmetric so that it has to be modified.

It may seem absurd to use the dissimilarities of the z_j as they are computed to be as independent as possible. However, in practice they are usually not completely independent since unit covariance does not imply this property. Therefore, there are still similarities to detect.

Definition 4.1. The *symmetric Kullback-Leibler divergence* of two probability density functions $p_{\mathcal{X}}$ and $p_{\mathcal{Y}}$ is defined as

$$\delta_{sym}(p_{\mathcal{X}}, p_{\mathcal{Y}}) = \frac{1}{2}\int p_{\mathcal{X}}(x)\ln\left(\frac{p_{\mathcal{X}}(x)}{p_{\mathcal{Y}}(x)}\right)\mathrm{d}x + \frac{1}{2}\int p_{\mathcal{Y}}(x)\ln\left(\frac{p_{\mathcal{Y}}(x)}{p_{\mathcal{X}}(x)}\right)\mathrm{d}x.$$

Lemma 4.1. *The symmetric Kullback-Leibler divergence is symmetric and positive definite if the integrals exist.*

Remark 20. *It turns out that the symmetric Kullback-Leibler divergence is not a proper distance measure since the triangular inequality does not hold. As a counterexample consider the densities $p_{\mathcal{X}}(x) = 2x$, $p_{\mathcal{Y}}(x) = 1$ and $p_{\mathcal{Z}}(x) = 2 - 2x$ for $x \in [0,1]$ and $p_{\mathcal{X}}(x) = p_{\mathcal{Y}}(x) = p_{\mathcal{Z}}(x) = 0$ for $x \notin [0,1]$. Then integration leads to $\delta_{sym}(p_{\mathcal{X}}, p_{\mathcal{Y}}) = \delta_{sym}(p_{\mathcal{Y}}, p_{\mathcal{Z}}) = \frac{1}{4}$ and $\delta_{sym}(p_{\mathcal{X}}, p_{\mathcal{Z}}) = 1$.*

Proof of Lemma 4.1. We observe that δ_{sym} can be written as the weighted sum of Kullback-Leibler divergences:

$$\delta_{sym}(p_{\mathcal{X}}, p_{\mathcal{Y}}) = \frac{1}{2}\left(\delta(p_{\mathcal{X}}, p_{\mathcal{Y}}) + \delta(p_{\mathcal{Y}}, p_{\mathcal{X}})\right) \tag{4.4}$$

Using (4.4) the positivity of δ_{sym} is a direct consequence of Lemma 3.1. And since $\delta_{sym}(p_{\mathcal{X}}, p_{\mathcal{Y}}) = 0$ holds if and only if both, $\delta(p_{\mathcal{X}}, p_{\mathcal{Y}})$ and $\delta(p_{\mathcal{Y}}, p_{\mathcal{X}})$, are zero we have positive definiteness. The symmetry follows also directly from (4.4). □

It is important to keep in mind, that the probability densities of the $\mathcal{Z}_j$ are unknown but we do, however, have the data z_j from which we can estimate the underlying densities. Recall that this can be done by the Edgeworth expansion (see Definition 3.10):

$$\begin{aligned} p_{\mathcal{X}}(x) = \phi_{\mathcal{X}}(x)\Big(1 &+ \frac{1}{3!}\kappa_3 h_3(x) \\ &+ \frac{1}{4!}\kappa_4 h_4(x) + \frac{10}{6!}\kappa_3^2 h_6(x) \\ &+ \frac{1}{5!}\kappa_5 h_5(x) + \frac{35}{7!}\kappa_3\kappa_4 h_7(x) + \frac{280}{9!}\kappa_3^3 h_9(x) \\ &+ \frac{1}{6!}\kappa_6 h_6(x) + \frac{56}{8!}\kappa_3\kappa_5 h_8(x) + \frac{35}{8!}\kappa_4^2 h_8(x) \\ &+ \frac{2100}{10!}\kappa_3^2\kappa_4 h_{10}(x) + \frac{15400}{12!}\kappa_3^4 h_{12}(x) \\ &+ o(m^{-2})\Big) \end{aligned}$$

Using this expansion the density $p_{\mathcal{Z}_j}$ is written as a function of the cumulants of $\mathcal{Z}_j$. These cumulants can be expressed using the central moments of $\mathcal{Z}_j$. In our case the central moments are equal to the general moments $m(j)_k$ because we assume the

variables to have zero mean. The moments $m(j)_k$ of the random variable $\mathcal{Z}_j$ can be estimated using the realizations of the variable:

$$m(j)_k = \frac{1}{n}\sum_{i=1}^{n} z_{j,i}^k,$$

where $z_{j,i}$ denotes the ith component of z_j.

Remark 21. *The first order cumulant of a random vector is its mean and the second its variance.*

The kth order cumulant can be expressed by a term involving the cumulants of order 1 to $k-1$ and the moments of order 1 to k:

$$\kappa(j)_k = m(j)_k - \sum_{i=1}^{k-1} \binom{k-1}{i-1} \kappa(j)_i m(j)_{k-i}.$$

For the cumulants $\kappa(j)_k$ of $\mathcal{Z}_j$ up to order k this yields

$$\begin{aligned}
\kappa(j)_1 &= m(j)_1 \\
\kappa(j)_2 &= m(j)_2 \\
\kappa(j)_3 &= m(j)_3 \\
\kappa(j)_4 &= m(j)_4 - 3m(j)_2^2 \\
\kappa(j)_5 &= m(j)_5 - 10m(j)_3 m(j)_2 \\
\kappa(j)_6 &= m(j)_6 - 15m(j)_4 m(j)_2 - 10m(j)_3^2 + 30m(j)_2^3.
\end{aligned}$$

See for example [31] for detailed information. So far we have seen how the z_j can be used to calculate probability densities which can be compared using the symmetric Kullback-Leibler divergence. Other similarity measures for probability density functions are known (see [11], [37]), but since we use the standard Kullback-Leibler divergence for ICA, it is reasonable to use a similar measure in this situation. In the next section we will see how to use this similarity measure to classify the orthogonal vectors z_j.

4.2.2. Clustering

The similarity of the vectors z_j will be represented in a so called *ixegram*, the independent component cross-entropy matrix (see [10]). This self-similarity matrix contains the pairwise dissimilarities of the vectors z_j.

Definition 4.2. The *ixegram* of a set of vectors $z_j \in \mathbb{R}^n$ is the matrix

$$D = (D_{ij})_{i,j=1,\dots,\rho} = \begin{pmatrix}
\delta_{sym}(p_{\mathcal{Z}_1}, p_{\mathcal{Z}_1}) & \delta_{sym}(p_{\mathcal{Z}_1}, p_{\mathcal{Z}_2}) & \cdots & \delta_{sym}(p_{\mathcal{Z}_1}, p_{\mathcal{Z}_\rho}) \\
\delta_{sym}(p_{\mathcal{Z}_2}, p_{\mathcal{Z}_1}) & \delta_{sym}(p_{\mathcal{Z}_2}, p_{\mathcal{Z}_2}) & \cdots & \delta_{sym}(p_{\mathcal{Z}_2}, p_{\mathcal{Z}_\rho}) \\
\vdots & \vdots & \ddots & \vdots \\
\delta_{sym}(p_{\mathcal{Z}_\rho}, p_{\mathcal{Z}_1}) & \delta_{sym}(p_{\mathcal{Z}_\rho}, p_{\mathcal{Z}_2}) & \cdots & \delta_{sym}(p_{\mathcal{Z}_\rho}, p_{\mathcal{Z}_\rho})
\end{pmatrix} \in \mathbb{R}^{\rho\times\rho}.$$

Lemma 4.2. *The ixegram D is symmetric and has non-negative entries. Its diagonal entries are all zero.*

Proof. The properties are all a direct consequence of Lemma 4.1. □

The task is to identify c groups of vectors by means of the ixegram without knowing how many vectors belong to each group. The idea is to assign those vectors to the same group which are similar to each other in such a way that vectors from different groups are as dissimilar as possible. This can be done by a pairwise clustering algorithm.

Clustering is the segmentation of a set of objects into subsets called clusters such that objects in the same cluster are similar. There are several pairwise clustering methods (see [40] and [41]). The one we use is the same as Casey and Westner proposed by Hofmann and Buhmann [27] in 1997. This method estimates an assignment matrix $M \in \mathbb{R}^{\rho \times c}$

$$M = (M_{jk})_{j=1,\dots,\rho,k=1,\dots,c} = \begin{pmatrix} P(z_1|C_1) & P(z_1|C_2) & \cdots & P(z_1|C_c) \\ P(z_2|C_1) & P(z_2|C_2) & \cdots & P(z_2|C_c) \\ \vdots & \vdots & \ddots & \vdots \\ P(z_\rho|C_1) & P(z_\rho|C_2) & \cdots & P(z_\rho|C_c) \end{pmatrix},$$

where $P(z_j|C_k)$ is the probability of assigning vector z_j to cluster (or group) C_k. The sum of each row of the matrix M should be one because it represents the probability that a certain z_j is assigned to any cluster. This property leads to ρ constraints.

The matrix M is computed so that it minimizes the following cost function proposed in [27]:

$$h(M, D) = \frac{1}{2} \sum_{i=1}^{\rho} \sum_{j=1}^{\rho} \frac{D_{ij}}{\rho} \left(\sum_{k=1}^{c} \frac{M_{ik} M_{jk}}{p_k} - 1 \right), \tag{4.5}$$

where $p_k = \frac{1}{\rho} \sum_{l=1}^{\rho} M_{lk}$ is the probability with which any component is assigned to cluster C_k. This normalization is necessary to compensate the different numbers of z_j in the different clusters. One can see if z_i and z_j are not very similar, they contribute a high cost if they are assigned to the same cluster.

Theorem 4.2. *Minimizing the cost function $h(M, D)$ is equivalent to minimizing*

$$H(M, D) = \sum_{k=1}^{c} \frac{1}{\sum_{l=1}^{\rho} M_{lk}} \sum_{i=1}^{\rho} \sum_{j=i+1}^{\rho} M_{ik} M_{jk} D_{ij}.$$

Proof. Resorting to (4.5) leads to

$$\begin{aligned} h(M, D) &= \frac{1}{2} \sum_{i=1}^{\rho} \sum_{j=1}^{\rho} \frac{D_{ij}}{\rho} \left(\sum_{k=1}^{c} \frac{M_{ik} M_{jk}}{p_k} - 1 \right) \\ &= \frac{1}{2\rho} \sum_{k=1}^{c} \frac{1}{p_k} \sum_{i=1}^{\rho} \sum_{j=1}^{\rho} M_{ik} M_{jk} D_{ij} - \frac{1}{2\rho} \sum_{i=1}^{\rho} \sum_{j=1}^{\rho} D_{ij}. \end{aligned}$$

The minimizer M does not depend on scaling changes, as the multiplication with $\frac{1}{\rho}$, or the addition of constant terms, as $\frac{1}{2\rho}\sum_{i=1}^{\rho}\sum_{j=1}^{\rho} D_{ij}$. Thus, the minimization of $h(M, D)$ is equivalent to minimizing

$$H(M, D) = \frac{1}{2}\sum_{k=1}^{c}\frac{1}{p_k}\sum_{i=1}^{\rho}\sum_{j=1}^{\rho} M_{ik}M_{jk}D_{ij}. \tag{4.6}$$

The term $M_{ik}M_{jk}D_{ij}$ can be interpreted as the entry (i, j) of a matrix A. The matrix A is symmetric since D is symmetric and has diagonal zero entries due to the fact that D has so. Therefore, the sum of all entries of A can be split into two sums of the same value by summing only the upper triangular entries and the lower triangular entries:

$$\begin{aligned}\sum_{i=1}^{\rho}\sum_{j=1}^{\rho} M_{ik}M_{jk}D_{ij} &= \sum_{i=1}^{\rho}\sum_{j=1}^{i-1} M_{ik}M_{jk}D_{ij} + \sum_{i=1}^{\rho}\sum_{j=i+1}^{\rho} M_{ik}M_{jk}D_{ij} \\ &= 2\sum_{i=1}^{\rho}\sum_{j=i+1}^{\rho} M_{ik}M_{jk}D_{ij}.\end{aligned}$$

In combination with (4.6) this yields

$$H(M, D) = \sum_{k=1}^{c}\frac{1}{p_k}\sum_{i=1}^{\rho}\sum_{j=i+1}^{\rho} M_{ik}M_{jk}D_{ij},$$

where $p_k = \frac{1}{\rho}\sum_{l=1}^{\rho} M_{lk}$. □

Of course the computational cost depends on the number of independent components, i.e. on the dimension ρ. Therefore, it is essential to keep this number low. So even if increasing the number of components would cause a more accurate separation this seems to be not a good idea.

Part II.

Applications

5. Outline

Many audio related applications take advantage of the ability to separate sources from a mixture without a prior knowledge about the mixing process. Thus, the analysis and separation of audio signals into their source components is an important tool for the extraction of metadata from audio data as for example separating musical instruments from a polyphonic ensemble, music restoration or extracting speech from a noisy background. Data obtained from a single-channel recording are often characterized by their high dimensionality. Therefore, the application of dimensionality reduction tools in this field is justified.

In this part we discuss how exactly dimensionality reduction methods can be used in Independent Subspace Analysis (ISA) for signal detection. In this context, signal detection is about identifying the time locations at which a certain source signal is active. In the field of signal processing, this is an important aspect, since it provides relevant information about a mixture of signals. This information can be used for further analysis as for example signal separation. In fact, provided the time locations where a certain source is active are known, separation algorithms could concentrate on these regions and perform the source extraction with higher resolution, but this is not the objective of this work.

The objective is to evaluate the usage of two dimensionality reduction methods (PCA and LE) in signal detection and separation algorithms. The combination of these methods is not a new concept (see [18], [50]). But to improve these strategies, a better mathematical understanding of these procedures supported by empirical tests is needed. In particular, we focus on the signal detection problem in a complex mixture of transitory acoustic sounds.

Assume a given band-limited signal $f \in L^2(\mathbb{R})$ to be the sum of c unknown source signals f_i:

$$f(t) = \sum_{i=1}^{c} f_i(t).$$

In the following we will consider the discretized problem, obtained by sampling the original signal f respecting the Nyquist-Shannon Sampling Theorem 1.2. This leads to a discrete signal $\mathfrak{s} = (f(t_l))_{l=1}^{N}$, which we assume to be the sum of c unknown discrete source signals $\mathfrak{s}_i \in \ell^\infty(\mathbb{Z}_N) = \mathbb{R}^N$:

$$\mathfrak{s} = \sum_{i=1}^{c} \mathfrak{s}_i.$$

The source signals $\mathfrak{s}_i$ are not necessary active during all time, i.e. there might exist t_l such that $\mathfrak{s}_{i,l} = f_i(t_l) = 0$ for some $i \in \mathbb{Z}_c$. The knowledge about the time steps t_l when

this happens is a crucial preprocessing step in order to extract the unknown source signals $\mathfrak{s}_i$. In the following, we describe how this detection can be done combining ISA and dimensionality reduction techniques.

In the previous chapters we have seen that ICA (and thus ISA) needs a data set $X \in \mathbb{R}^{d \times n}$ of different observations as an input in order to estimate $\rho \leq d$ sources of length n. This means that we need at least as many observations as sources. In practice, this is usually not the case. In fact, in many cases we have fewer sensors than sources and it is very common to consider even single sensor problems.

In this context, the observed signal $\mathfrak{s}$ is only a single-channel recording and we cannot directly apply ICA to the measured data. Therefore, in a further preprocessing step, by multiplication with a window function we split the observed signal $\mathfrak{s}$ into vectors $\mathfrak{s}^k$, $1 \leq k \leq n$ of length D, such that we get a data set $X_\mathfrak{s} = \{\mathfrak{s}^k\}_{k=1,\ldots,n} \subset \mathbb{R}^D$. This data set $X_\mathfrak{s}$ lies in $\mathcal{M}_\mathfrak{s}$, a low-dimensional space or manifold embedded in the high-dimensional space $\mathbb{R}^D$. One of the reasons justifying this assumption is that all data points are segments from the same signal and thus they are similar.

Since each source can be characterized by the frequencies it is containing, we need some information about the frequencies occurring in the signal at each time step in order to detect the time steps where a source signal is active. As the Fourier transform provides only information about the frequencies of a signal during its whole time period, we have to consider the spectrogram of the signal $\mathfrak{s}$. Therefore, we perform a D-points discrete Fourier transform on each of the segments $\mathfrak{s}^k \in X_\mathfrak{s}$ of the windowed signal. Each of this Fourier transforms is assumed to represent the frequency range of $\mathfrak{s}$ at one time step. This procedure can be written as $X = \mathcal{F}_D(X_\mathfrak{s}) \in \mathbb{R}^{D \times n}$, i.e. we switch from time-amplitude space to time-frequency space by Fourier analysis. Effectively, we apply the short-time Fourier transform on the signal $\mathfrak{s}$

$$X = \mathcal{F}_{\varphi,D}\mathfrak{s},$$

in order to compute the spectrogram. This step is necessary to make the hidden information in the signal accessible, so that we can use it for the detection. The manifold $\mathcal{M}_\mathfrak{s}$ is also transformed by $\mathcal{F}_D$ into another manifold $\mathcal{M} = \mathcal{F}_D(\mathcal{M}_\mathfrak{s})$. The value D depends on the frequency range of the signal which is usually large.

To determine te independent components of X, we use the JADE algorithm introduced in Section 3.3. This algorithm computes a mixing matrix A, which would be of size $D \times D$. Since D is huge, performing an ICA is computationally very expensive. Thus, we can take advantage of the assumption that the data lies on a low-dimensional manifold and apply a dimensionality reduction method. This step reduces the dimensionality of the data from D to d.

At this point the estimation of the intrinsic dimensionality of the data set is a crucial task, whose influence is a matter for future studies. However there are different strategies available for estimating this dimensionality (see [36], [39]).

After the reduction of the dimensionality and the computation of the d independent components, we perform a grouping in which each component is assigned to one of the c source signals. It is important to know how many source signals we look for in order to determine c, the number of clusters.

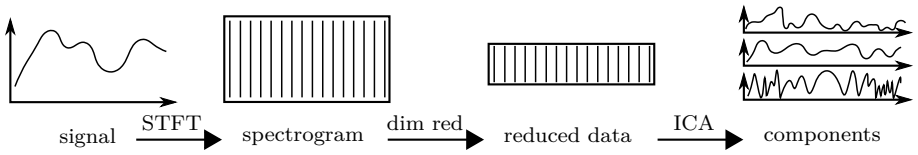


Figure 5.1.: General proceeding of ISA with dimensionality reduction.

By means of the components the time locations where the different sources are active are estimable. A complete separation and computation of the source signals is only possible if a back transform for the dimensionality reduced data is known. This is not always provided since the underlying models are often not fully respected or the methods are highly non-linear. A schematic overview of the above explained procedure is shown in Figure 5.1 and for a better understanding of the involved mathematical objects see the following diagram.

$$
\begin{array}{rl}
\mathfrak{s} = \sum \mathfrak{s}_i & \subset \mathbb{R}^N \\
\downarrow \text{windowing} & \\
X_{\mathfrak{s}} & \subset \mathcal{M}_{\mathfrak{s}} \subset \mathbb{R}^D \\
\downarrow \text{FFT} & \\
X = \mathcal{F}_D(X_{\mathfrak{s}}) & \subset \mathcal{M} \subset \mathbb{R}^D \\
\downarrow \text{dimensionality reduction } P & \\
Y = P(X) & \subset \Omega \subset \mathbb{R}^d \\
\downarrow \text{ICA} & \\
\{\bar{s}_i\} & \subset \mathbb{R}^n \\
\downarrow \text{grouping} & \\
\{s_i\} & \subset \mathbb{R}^n
\end{array}
$$

6. Independent Subspace Analysis: An illustrative example

Now we discuss an illustrative example in order to elucidate the above explained algorithm and to compare the different dimensionality reduction methods PCA and LE. In the next sections we present the results, obtained by the different steps of a MATLAB implementation of the algorithm, following roughly Part I. The signal $\mathfrak{s}$ is the mixture of two transitory acoustic sounds

$$\mathfrak{s} = \mathfrak{s}_1 + \mathfrak{s}_2,$$

where $\mathfrak{s}_1$ is the signal of a cymbal and $\mathfrak{s}_2$ the signal of castanets (see Figure 6.1) both of length $N = 100000$ samples. In this setting we know the source signals of $\mathfrak{s}$ so that we are able to compare the detection results with the real sources.

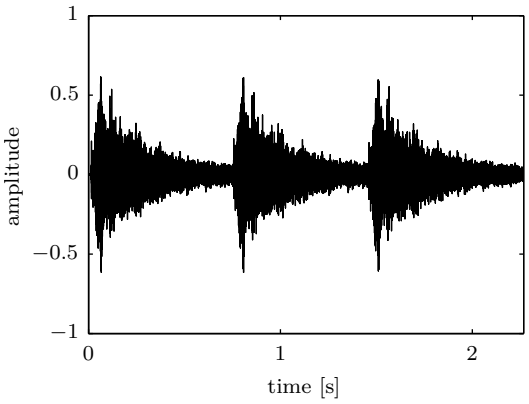
The signal of the castanets $\mathfrak{s}_2$ is active only at few time steps beside some transient effects (Figure 6.1c). Regarding the mixed signal $\mathfrak{s}$ (Figure 6.1e) we observe that the signal $\mathfrak{s}_2$ is quite difficult to distinguish from the mixture $\mathfrak{s}$. This illustrates once more, why detecting the time locations where $\mathfrak{s}_2$ is active is important. Provided these time steps are known, a separation algorithm with high resolution in these regions could do a good job without considering the whole time period of $\mathfrak{s}$.

6.1. Time-Frequency Analysis

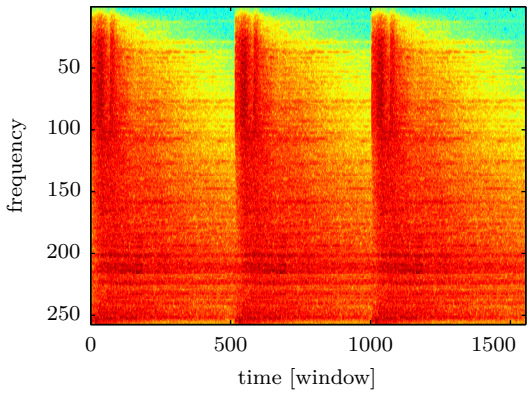
The next step is to compute the spectrogram $X = \mathcal{F}_{\varphi,D}\mathfrak{s}$ of $\mathfrak{s}$. The spectrogram plots the frequency range of the signal at each time step against the time. While in the time-amplitude plot only one information per time step is available, namely the amplitude of the signal, the spectrogram provides much more information per time step. Time-frequency analysis somehow opens the signal to make this information accessible. Figure 6.1 illustrates this benefit of time-frequency transforms as STFT.

We use a STFT based on the discrete Hann window defined in Section 1.2 with window length $D = 512$, i.e. each of the segments $\mathfrak{s}^k$ is made of 512 samples. For the frequency range at each time step a 512-point FFT is applied on the windowed signal. The hop size is chosen as $h = 64$, which means that the segments $\mathfrak{s}^k$ have a distance of 64 sample points, and thus the segments overlap. As a consequence the interval between two time steps in the spectrogram consists of 64 sample points. Since $N = 100000$ this yields $n = 1555$ time steps.

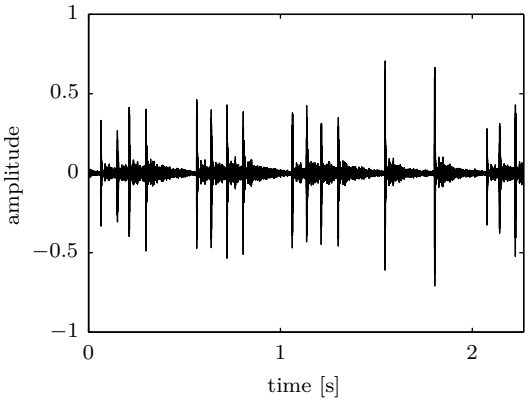
Figure 6.1 shows the spectrograms of $\mathfrak{s}$, $\mathfrak{s}_1$ and $\mathfrak{s}_2$. As well as in the time-amplitude plot the signal $\mathfrak{s}_2$ is hidden in the signal $\mathfrak{s}_1$, with the result that the spectrograms of $\mathfrak{s}$ and $\mathfrak{s}_1$ are hardly distinguishable.



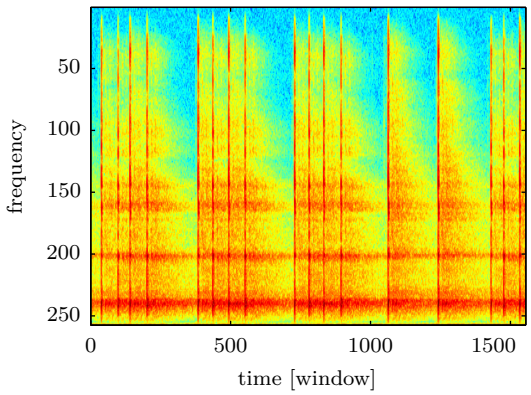
(a) Cymbal $\mathfrak{s}_1$.



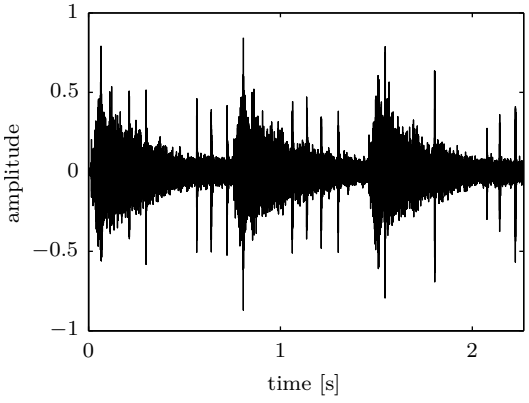
(b) Spectrogram of the cymbal $\mathcal{F}_{\varphi,D}\mathfrak{s}_1$.



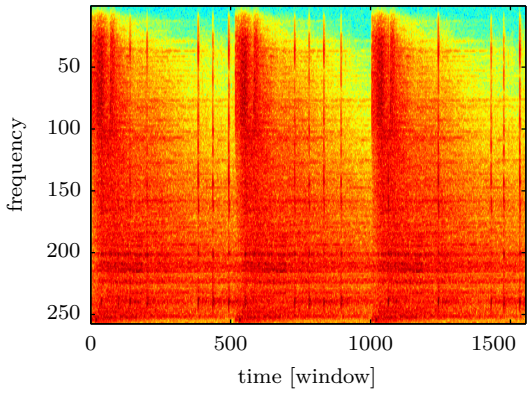
(c) Castanets $\mathfrak{s}_2$.



(d) Spectrogram of the castanets $\mathcal{F}_{\varphi,D}\mathfrak{s}_2$.



(e) Mixture $\mathfrak{s} = \mathfrak{s}_1 + \mathfrak{s}_2$.



(f) Spectrogram of the mixture $X = \mathcal{F}_{\varphi,D}\mathfrak{s}$.

Figure 6.1.: The mixed signal $\mathfrak{s}$ is the sum of $\mathfrak{s}_1$ and $\mathfrak{s}_2$.

6.2. Dimensionality Reduction

If we interpret the spectrogram of $\mathfrak{s}$ (Figure 6.1f) as a 512×1555 matrix, each row of the spectrogram corresponds to one frequency and each column to a time step. Thus the entry (i, j) of the spectrogram provides the portion of the frequency i at time j. Various rows of the spectrogram resemble each other, such that the assumption, that the data set X lies in a low-dimensional space or manifold is plausible. Reducing the dimensionality from $D = 512$ to $d = 10$ leads to the reduced data sets $Y_{PCA} = P_{PCA}(X)$ and $Y_{LE} = P_{LE}(X)$. A plot of each row of the reduced data set is presented in Figure 6.2 and 6.3. Even though some of the rows are quite diffuse, we can already conjecture that for example 6.2a or 6.3a correspond to the cymbal and 6.2b or 6.3c to the castanets.

We take $d = 10$ as the intrinsic dimensionality of the data set X because this gives the best result. This choice is consistent with [50] where a range from 10 up to 30 dimensions is proposed. In the case of LE we use r-ary neighborhoods (cf. Section 2.3.1) with $r = 100$. For the implementation the dimensionality reduction toolbox for MATLAB provided by van der Maaten [38] is employed.

6.3. Independent Component Analysis

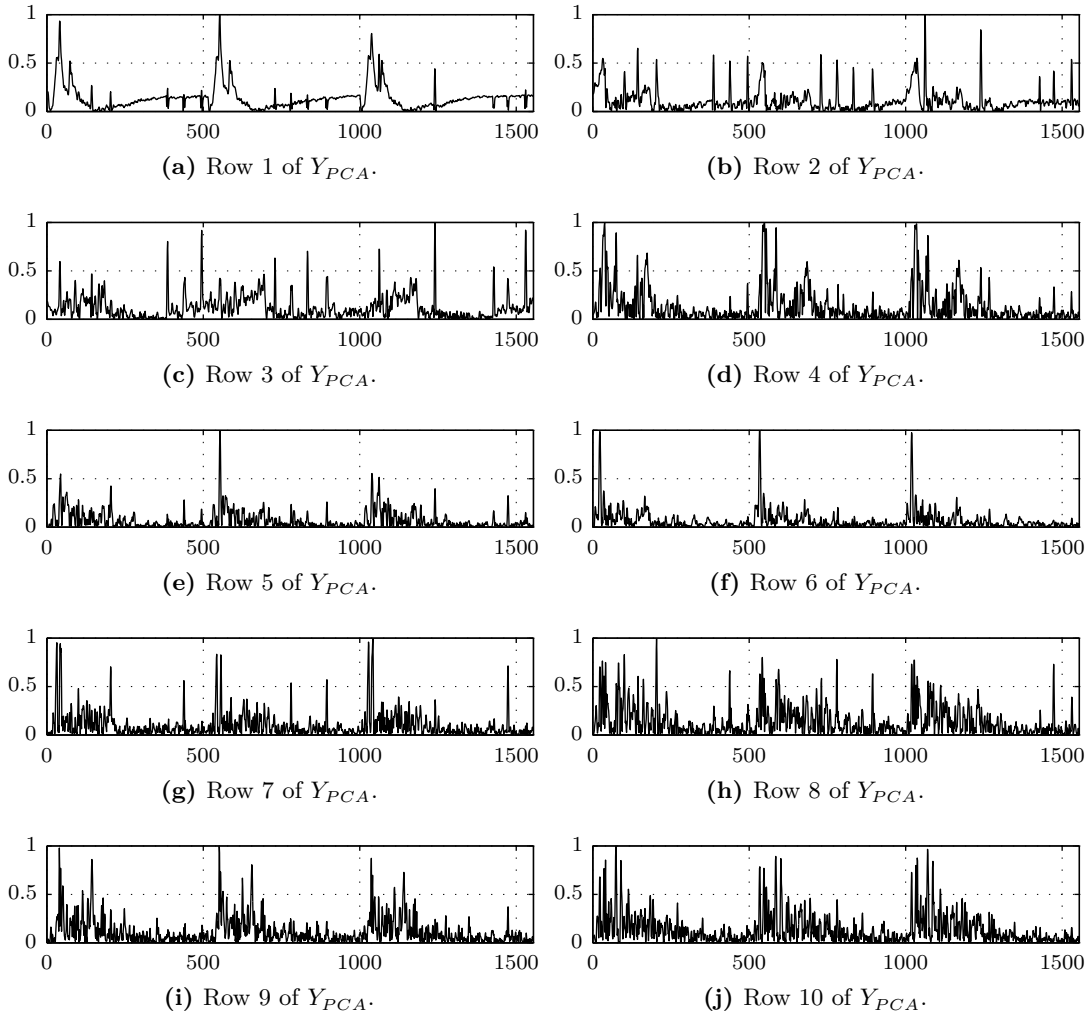
Application of Independent Component Analysis on the data sets Y_{PCA} and Y_{LE} leads to two sets of independent components Z_{PCA} and Z_{LE} shown in Figure 6.4 and 6.5. Later on, these components are partitioned into two sets each of which is corresponding to one of the source signals. This clustering can be done by a grouping algorithm.

Nevertheless, at this point we can already identify manually some components belonging to each of the sources. For example in the case of PCA the components 6.4b and 6.4g should be assigned to the castanets and the component 6.5c in the case of LE. Furthermore, the components 6.4d and 6.4j, 6.5d and 6.5h respectively, match the cymbal. This classification is heuristic and only possible since we know the sources.

However, some of the components are not easily assignable. But in general the separation property of ICA improves the detection quality, as can be seen comparing Figure 6.2 and 6.4, 6.3 and 6.5 respectively. More precisely, the peaks of the castanets emerge noticeably after the application of ICA and some background signals are suppressed.

Regarding the results of ICA with LE, component 6.5c is noteworthy, because it reflects almost the complete pattern of the castanets (compare Figure 6.1c). We conclude that ICA combined with the non-linear dimensionality reduction technique leads to a slightly better detection of the transitory castanets.

Note that the independent components are not a filtered version of the reduced data set, but rather a basis, i.e. the rows of the data matrix Y are linear combinations of the independent components. To compute the mixing matrix A we used the JADE algorithm implemented by Cardoso [6] which we introduced in Section 3.3.



(a) Row 1 of Y_{PCA}. **(b)** Row 2 of Y_{PCA}.

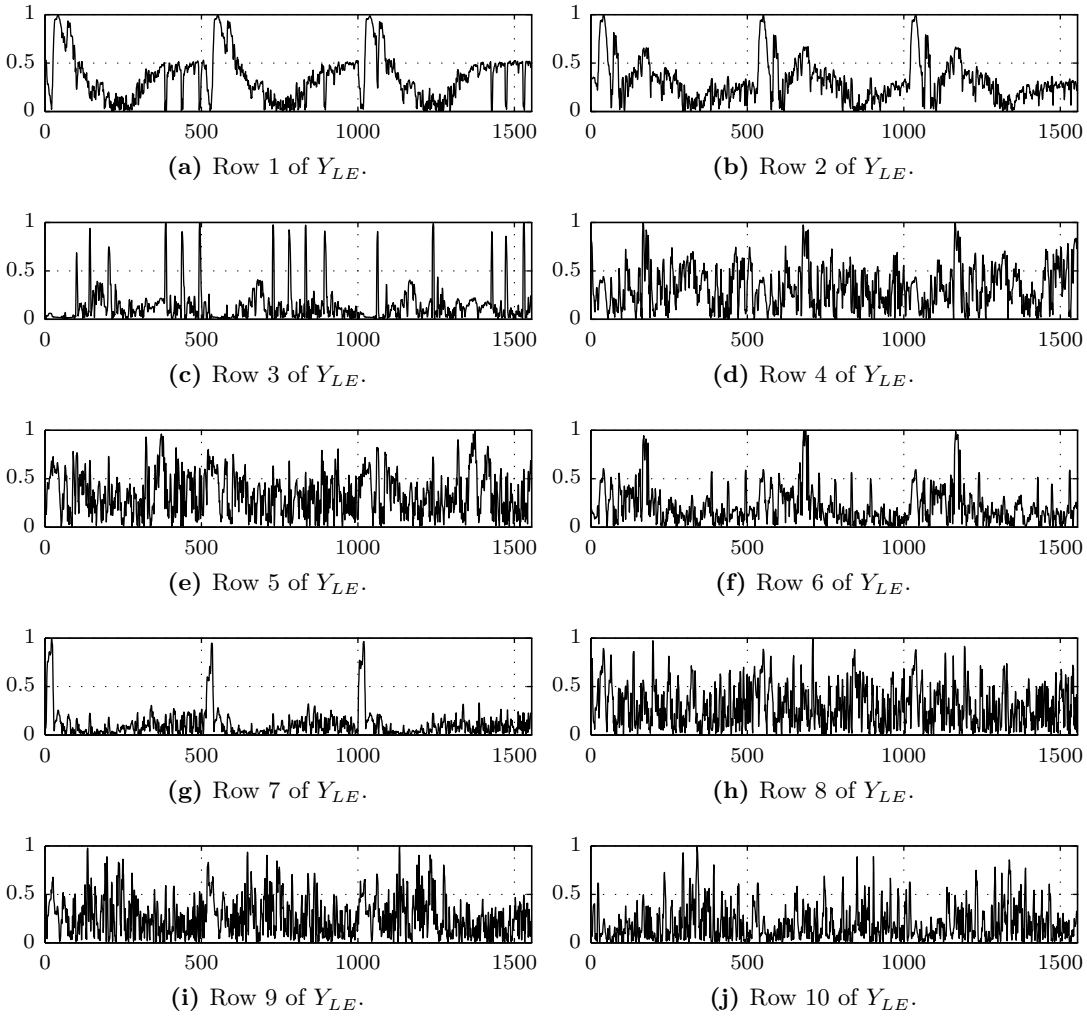
(c) Row 3 of Y_{PCA}. **(d)** Row 4 of Y_{PCA}.

(e) Row 5 of Y_{PCA}. **(f)** Row 6 of Y_{PCA}.

(g) Row 7 of Y_{PCA}. **(h)** Row 8 of Y_{PCA}.

(i) Row 9 of Y_{PCA}. **(j)** Row 10 of Y_{PCA}.

Figure 6.2.: Results after applying Principal Component Analysis to the data set X. The reduced data matrix Y_{PCA} has $d = 10$ rows.



(a) Row 1 of Y_{LE}.

(b) Row 2 of Y_{LE}.

(c) Row 3 of Y_{LE}.

(d) Row 4 of Y_{LE}.

(e) Row 5 of Y_{LE}.

(f) Row 6 of Y_{LE}.

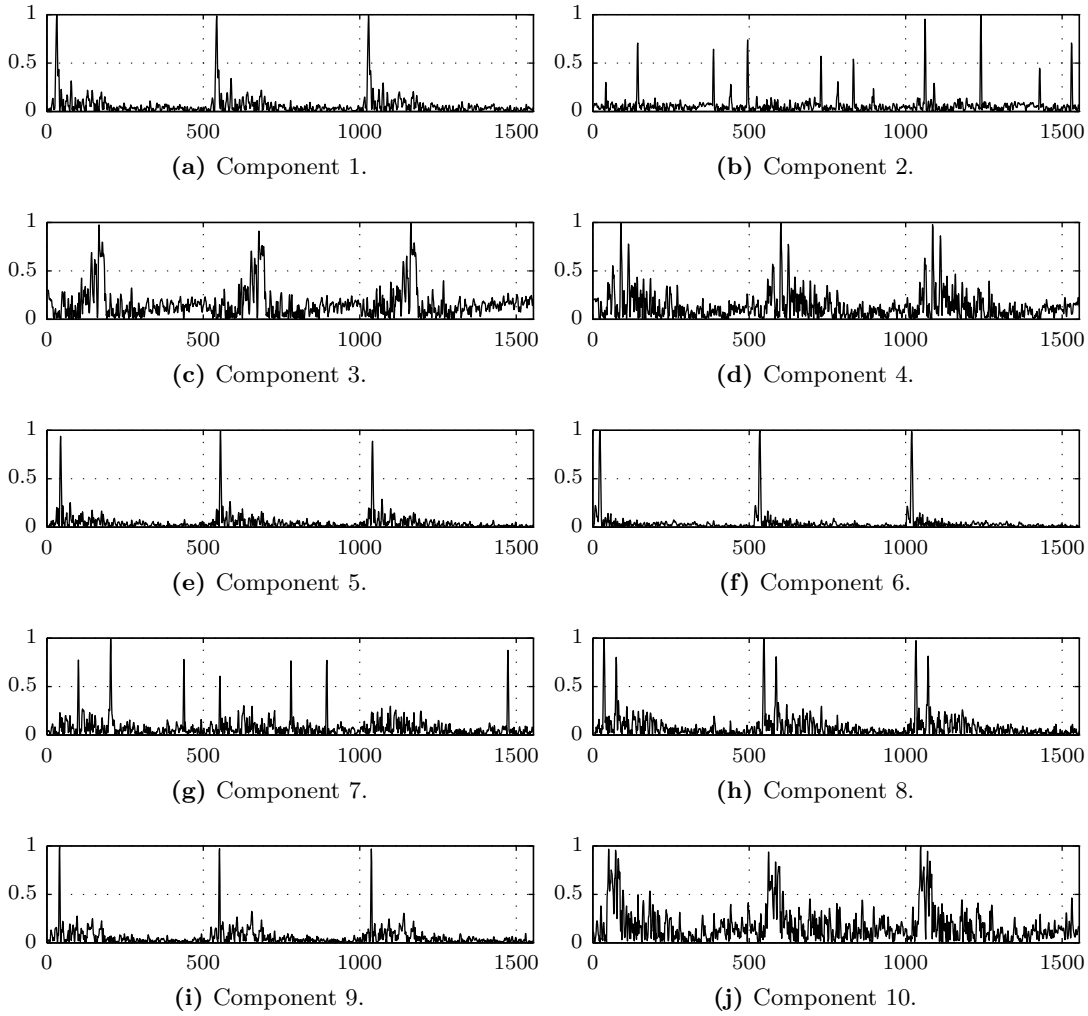
(g) Row 7 of Y_{LE}.

(h) Row 8 of Y_{LE}.

(i) Row 9 of Y_{LE}.

(j) Row 10 of Y_{LE}.

Figure 6.3.: Results after applying Laplacian Eigenmaps to the data set X. The reduced data matrix Y_{LE} has $d = 10$ rows.



(a) Component 1. **(b)** Component 2.

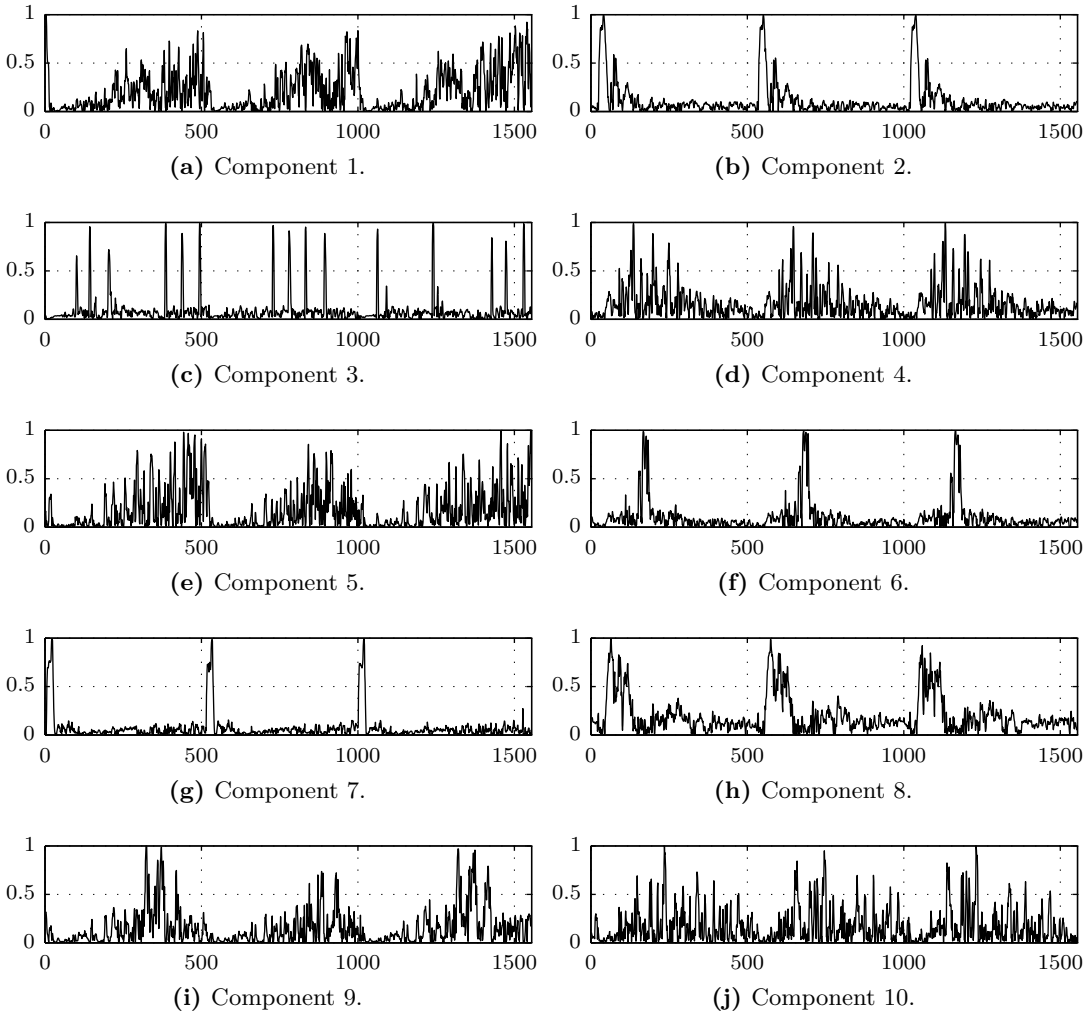
(c) Component 3. **(d)** Component 4.

(e) Component 5. **(f)** Component 6.

(g) Component 7. **(h)** Component 8.

(i) Component 9. **(j)** Component 10.

Figure 6.4.: Applying Independent Component Analysis to the reduced data set Y_{PCA} leads to a set of 10 components Z_{PCA}.



(a) Component 1. **(b)** Component 2.

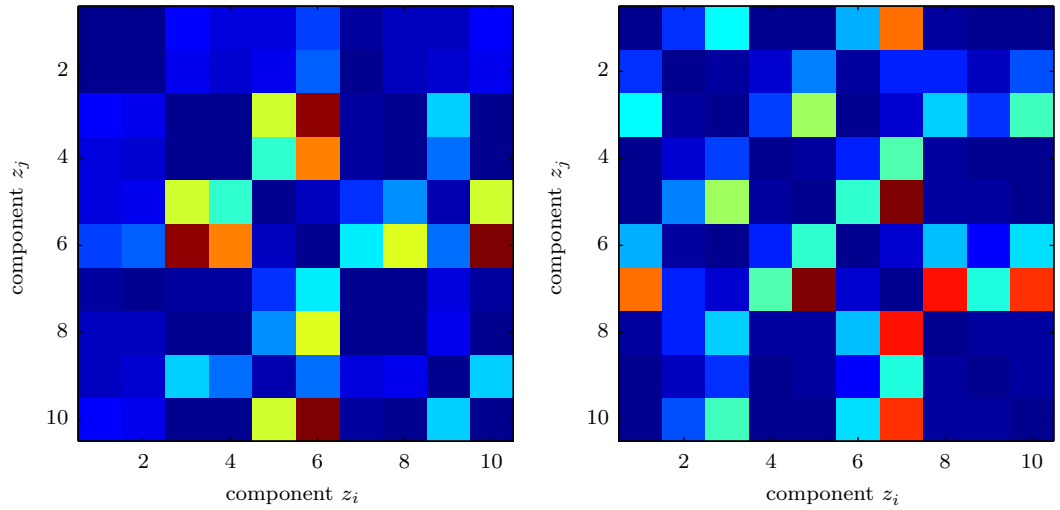
(c) Component 3. **(d)** Component 4.

(e) Component 5. **(f)** Component 6.

(g) Component 7. **(h)** Component 8.

(i) Component 9. **(j)** Component 10.

Figure 6.5.: Applying Independent Component Analysis to the reduced data set Y_{LE} leads to a set of 10 components Z_{LE}.



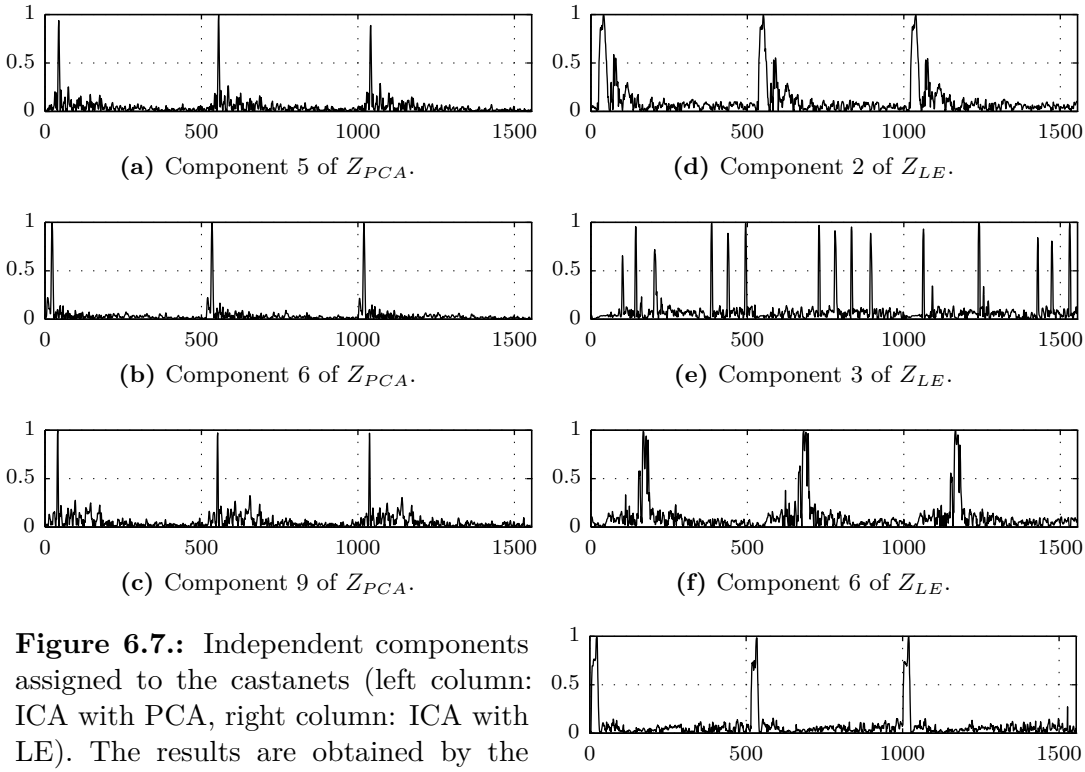
(a) Ixegram D_{PCA} of Z_{PCA}. **(b)** Ixegram D_{LE} of Z_{LE}.

Figure 6.6.: Ixegram, matrix of dissimilarities. The bluer a cell, the more similar the components are.

6.4. Grouping

Regarding the independent components in Z_{PCA} or Z_{LE}, we are already able to classify some of them but not all. This assignment is only possible since we know the sources, which is usually not the case. Due to this fact, the clustering should be done automatically. One possibility is to sort the components by their similarity to each other as discussed in Section 4.2. This is done by an evaluation of the ixegram, a matrix containing the pairwise dissimilarities of the components (see Figure 6.6). The more similar the components z_i and z_j are, the smaller the entry (i, j) of the ixegram is. The proposed clustering algorithm computes an assignment matrix M using the ixegram D. This computation is based on the minimization of the functional $H(M, D)$. The assignment matrices obtained by this procedure are

$$M_{PCA} = \begin{pmatrix} 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad M_{LE} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}.$$



(a) Component 5 of Z_{PCA}.

(d) Component 2 of Z_{LE}.

(b) Component 6 of Z_{PCA}.

(e) Component 3 of Z_{LE}.

(c) Component 9 of Z_{PCA}.

(f) Component 6 of Z_{LE}.

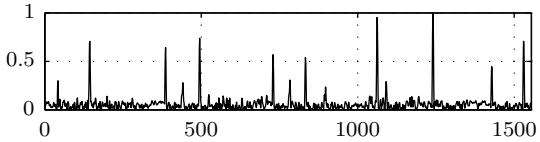
Figure 6.7.: Independent components assigned to the castanets (left column: ICA with PCA, right column: ICA with LE). The results are obtained by the grouping algorithm based on similarity measurements.

(g) Component 7 of Z_{LE}.

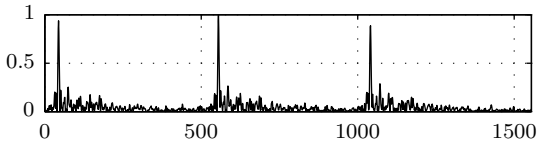
The entry (i, k) of the assignment matrix has to be interpreted as the probability with which component z_i belongs to the kth source, for $i = 1, \ldots, 10$ and $k \in \{1, 2\}$. This clustering leads to a classification of the components. The components assigned to the signal of the castanets are shown in Figure 6.7.

Unfortunately, the selected components in the case of PCA are not at all representing the pattern of the castanets. Therefore, it is a natural proceeding to compare these results with a manual grouping in order to see if better results can be obtained. A possible assignment matrix corresponding to a manual choice would be

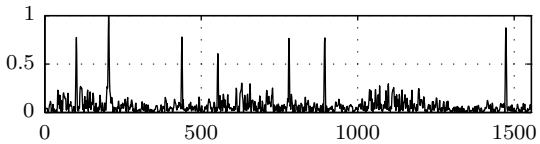
$$M_{PCA}^{hand} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}.$$



(a) Component 2 of Z_{PCA}.



(b) Component 5 of Z_{PCA}.



(c) Component 7 of Z_{PCA}.

Figure 6.8.: These three components computed by ICA with PCA are assigned manually to the castanets.

The components assigned to the castanets by this manual grouping are shown in Figure 6.8. This result is much better than the previous one (see Figure 6.7a - 6.7c). This raises the conjecture that the grouping method proposed by Casey and Westner [10] in not adequate in our situation. This could be due to the transience of the source signals.

Running the algorithm explained in this chapter needs only a few seconds on a standard personal computer. This is an important progress caused by the inclusion of dimensionality reduction to the concept of ISA. This modification reduces the dimensions of the mixing matrix A from 512×512 to 10×10.

7. Separation in the case of PCA

So far we have seen how dimensionality reduction and ISA can be used for signal detection. To give a perspective for further applications we shortly discuss the separation in the special case of PCA.

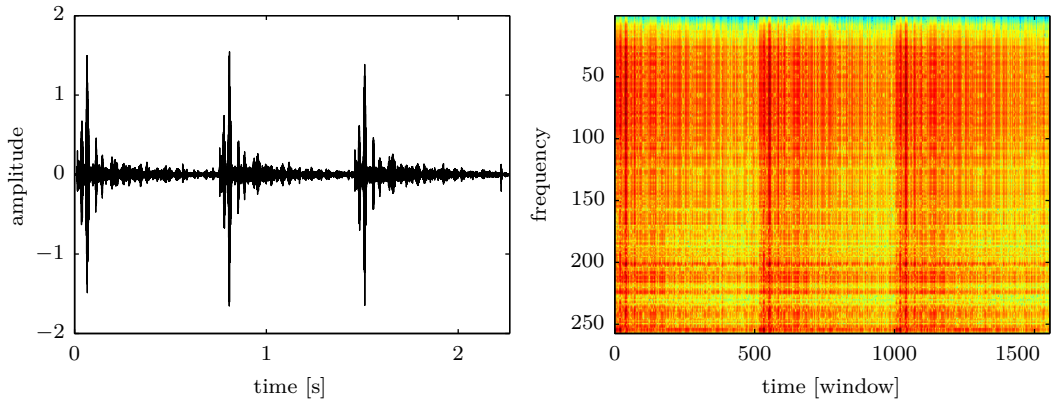
Referring to the separation of signals, we have already stated that a computation of the source signals is only possible if we are able to back transform the dimensionality reduced data into the original high-dimensional space. In the case of PCA, for instance, a back transform is given by the matrix W. Provided this back transform, theoretically nothing prevents us from a complete separation of the signal, since inversion formulas for the other steps of the algorithm, i.e. STFT and ICA, are well studied. But in practice, even if the back transform is known, this does not necessarily imply that an adequate reconstruction of the source signals is possible. This is due to the fact that the underlying mixing model might not bee completely satisfied.

Returning to the signal detection in which PCA was involved, we can continue with a reconstruction step. The low-dimensional data set Y_{PCA} can be decomposed by ISA in two data sets Y_1 and Y_2 corresponding to the two sources (compare Theorem 4.1). If the spectrogram $X = \mathcal{F}_{\varphi,D}\mathfrak{s}$ and the reduced data Y fulfill the PCA model $X = WY$, each of the data sets can be back transformed by W. This leads to the spectrograms $X_1 = WY_1$ and $X_2 = WY_2$ of the source signals.

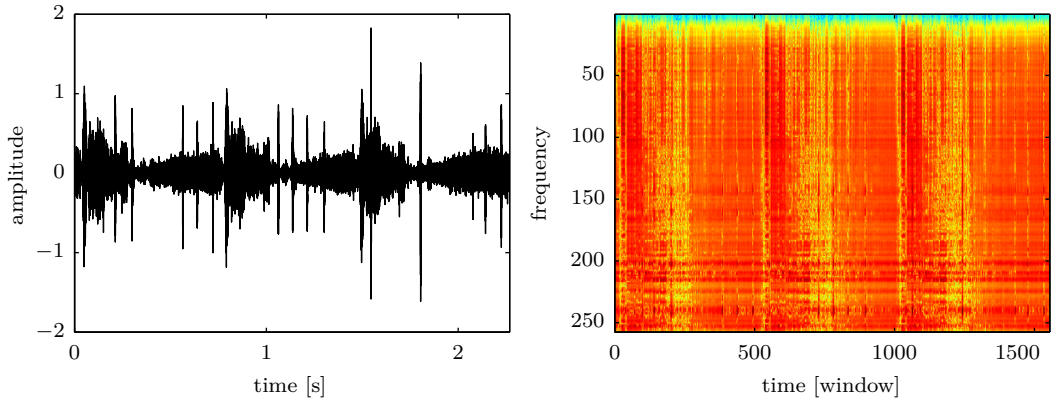
The inverse STFT cannot be directly applied to the so obtained spectrograms, since the phase information is missing. Therefore, in a naive ansatz we assume that the phase angles of the sources are the same as of the original signal because this information is accessible.

If we take a look at the results shown in Figure 7.1, we observe that the extracted sources are different from the input signals $\mathfrak{s}_1$ and $\mathfrak{s}_2$ (cf. Figure 6.1). The poor separation can be explained partly by the non-conformity of the data and the PCA model $X = WY$. The gravity of this mismatching becomes apparent if we compare the back transformed mixture $\mathfrak{s}$ with the original one in Figure 6.1e. In the ideal case these two signal should be the same. The spectrograms depict the fact that a lot of information was lost during the process of PCA because not only small eigenvalues of the covariance matrix XX^T were neglected. This increases the mismatching.

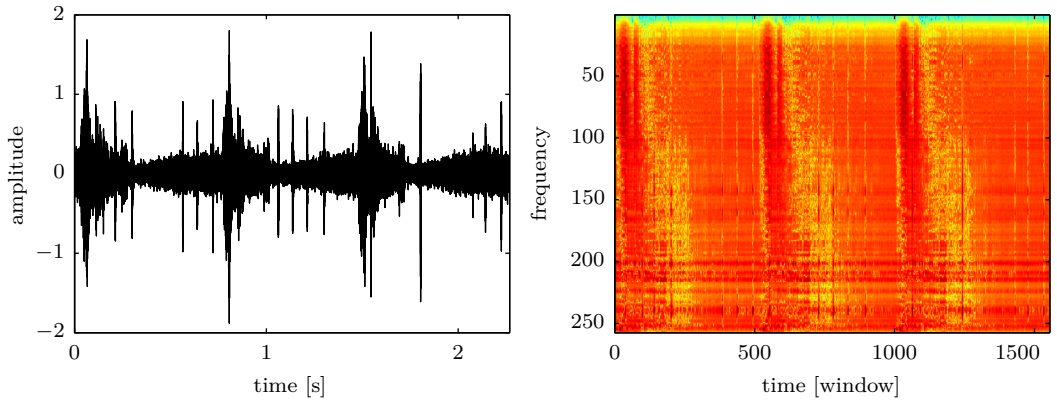
Further reasons for the poor separation could be the fact that the ICA mixing model is not completely respected and the use of an inappropriate clustering algorithm. For the purpose of comparison we also reconstruct the source signals based on the manually composed assignment matrix M_{PCA}^{hand}. The results are shown in Figure 7.2. For the grouping by hand the obtained results are much better than before, which confirms the impression that the clustering algorithm has to be improved. The signal of the castanets is quite well extracted but further work has to be done for entirely satisfying results.



(a) Source 1. **(b)** Spectrogram of source 1.

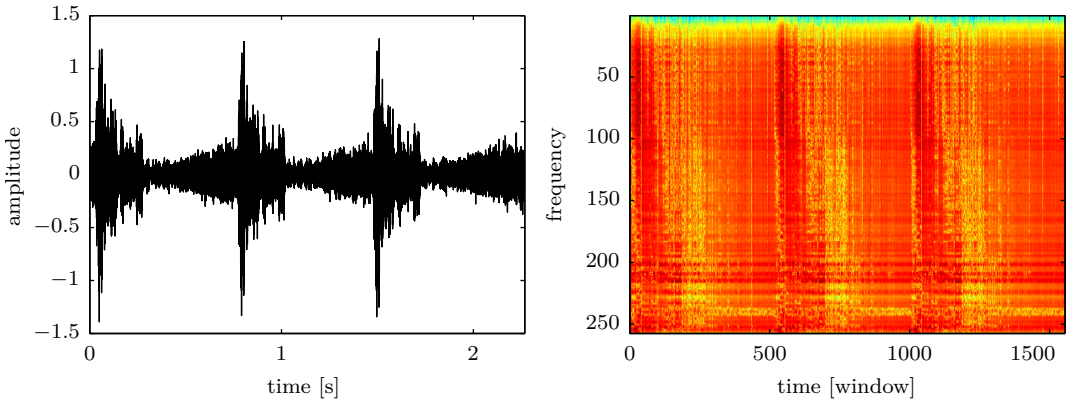


(c) Source 2. **(d)** Spectrogram of source 2.

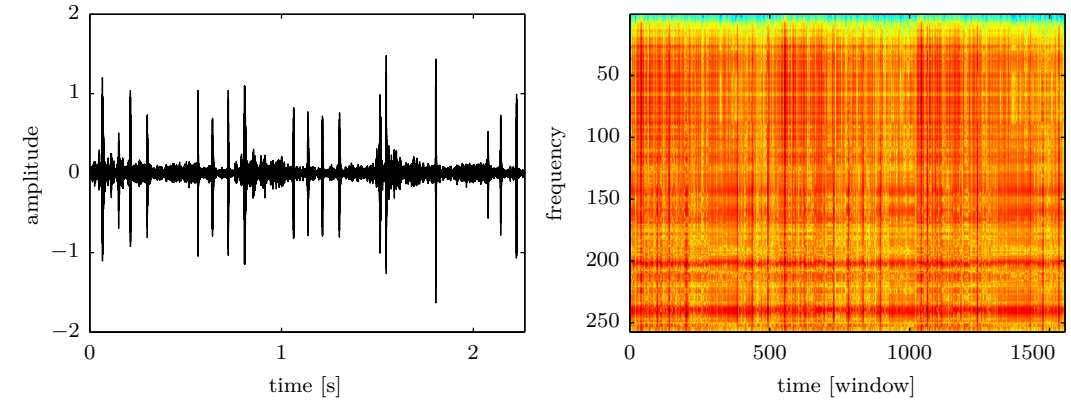


(e) Reconstructed mixture. **(f)** Spectrogram of the reconstructed mixture.

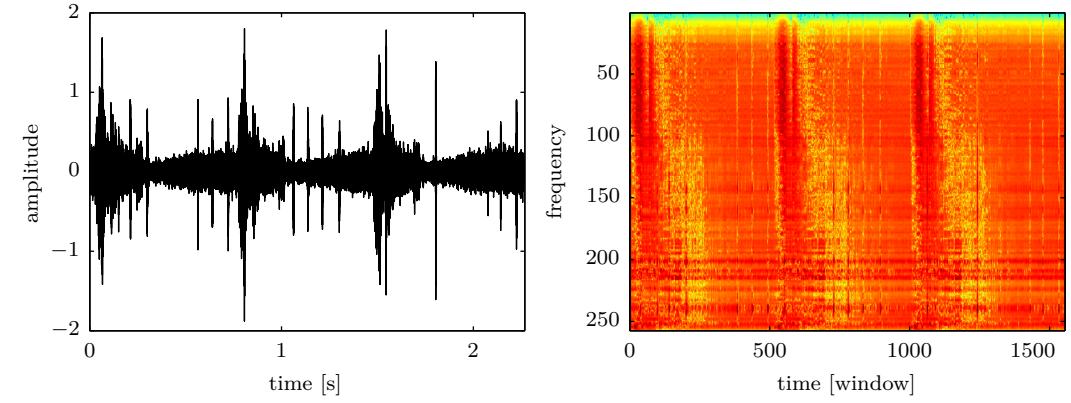
Figure 7.1.: In the case of PCA, a back transform is possible. The so obtained separation depends on the assignment matrix M_{PCA}.



(a) Source 1. **(b)** Spectrogram of source 1.



(c) Source 2. **(d)** Spectrogram of Source 2.



(e) Reconstructed mixture. **(f)** Spectrogram of the reconstructed mixture.

Figure 7.2.: The back transform based on the manual assignment of the independent components leads to better results.

8. Conclusion

In many application fields signal separation and, as a consequence, signal detection is a crucial task. Even though there are some well-understood methods, the development depends on experimental results supported by a correct understanding of the mathematical background. In recent years, the research has focused on several extensions of Independent Component Analysis (ICA) (see [8], [10], [29]) among them Independent Subspace Analysis (ISA). ISA is a combination of the classical ICA method and time-frequency analysis. Since time-frequency data have usually a high dimensionality, there have been several approaches to involve dimensionality reduction in the concept of ISA (see [17], [18], [50]).

The objective of this work was to evaluate the usage of Principal Component Analysis (PCA) and Laplacian Eigenmaps (LE) in signal detection. Therefore, we have introduced the concepts of short-time Fourier transform (STFT), PCA, LE, ICA and ISA in the first part of this work and in the second we discussed a concrete example to illustrate these methods and to see how they act on a complex mixture of transitory signals.

We have seen that sources can still be detected if the ICA is applied to reduced spectrogram. PCA and LE improve this detection in various aspects. On the one hand, it speeds up the separation into statistically independent components since the dimension of the unmixing matrix G decreases immensely. On the other hand, the computational cost of the assignment matrix M, depending on the number of independent components is reduced. It is surprising that the linear reduction method PCA gives relatively good results, but the results obtained with the non-linear method LE are even more sophisticated.

Furthermore, we introduced a clustering algorithm to assign the independent components to the different sources. The resulting clustering did not meet our expectations since a manual assignment leads, at least in the case of PCA, to much better results.

In conclusion, the application of ICA improves the quality of the detection in both cases (PCA and LE).

Of course, we do not claim that this work includes all of the diverse aspects related to the topic of dimensionality reduction in ISA for signal detection, but we tried to give an overview of this field. Although we have achieved relatively good results, further improvement of the detection scheme is necessary. There are various aspects whose influence on the detection quality of the scheme could be studied. Among these are the usage of different time-frequency analysis methods as for example the wavelet transform, the comparison of new dimensionality reduction methods, other grouping algorithms and alternatives to ICA. In the next paragraphs we like to highlight some of these.

Dimensionality reduction methods: As we have seen in the previous chapters the quality of the detection and separation depends on the conformity of the data and

the dimensionality reduction model. Due to the fact that dimensionality reduction methods are only able to approximate the manifold where the data set lies on, the model is typically not fully respected. Especially in the case of complex manifolds this leads to an inexact detection. Since the family of dimensionality reduction methods is huge, there is a lot more work to do. Auditory signals are complicated structures which might be extremely complex, such that the usage of non-linear techniques becomes necessary. In particular in the last decades, there have been developed many innovative dimensionality reduction algorithms as for example ISOMAP as an extension of Multidimensional Scaling (MDS), Local Tangent Space Alignment (LTSA), Whitney embedding based methods or Riemannian Normal Coordinates (RNC). It might be interesting to compare the different methods in interaction with ICA.

An additional task is to study the possibility of back transforming the reduced data into the original high-dimensional space. This would enable the extraction of source signals and thus a complete separation of the mixed signals could be obtained.

Grouping: Another aspect which could be analyzed is the grouping method. In Section 6.4 we have seen that clustering the components in not an easy task since for some of the components it is not clear at all to which source they belong. We used the clustering algorithm proposed in [10] to identify the subspaces where the different sources are lying in. But the results presented in the second part have raised the conjecture that this algorithm was not adequate. Thus, a further task could be the evaluation of other clustering algorithms which probably lead to more sophisticated results. While the presented method uses pairwise dissimilarities of the components, there are other methods which are not based on statistical tools as the measure for the similarity of density functions. For instance, the clustering algorithm proposed in [50] relying on different features of a signal, namely percussiveness, noise-likeness and spectral dissonance, and another algorithm, based on the envelopes of the subspaces corresponding to the sources (see [52]), are very promising approaches in this context.

Non-Negative Matrix Factorization: In order to decompose the reduced spectrogram Y we have modelled the data as a random process, i.e. we have interpreted the data matrix Y as n realizations of a random vector. This is an assumption about which one can argue, because the origin of many signals (especially music) is deterministic. Therefore, it could be interesting to see how non-statistic matrix decompositions for a data set behave in combination with dimensionality reduction methods. One of these decompositions is non-negative matrix factorization which has already been used for single-channel problems (compare [46] and [51]).

Parameter tuning: Some of the involved parameters, as for example the intrinsic dimensionality of the spectrogram or the number of source signals, were chosen manually. This choice should be automated for an extension of the application to other signals. Also a study concerning the dependence of the result on the parameters could be interesting.

Summarizing we can say that the discussed dimensionality reduction techniques are able to detect the source signal in our example, but further investigations are necessary to dispose the algorithm for practical use.

Bibliography

[1] D. Barry, E. Coyle, D. Fitzgerald and R. Lawlor. *Single Channel Source Separation Using Short-Time Independent Component Analysis.* In *Audio Engineering Society Convention 119*. 2005.

[2] D. Barry, E. Coyle and B. Lawlor. *Sound Source Separation: Azimuth Discrimination and Resynthesis.* In *Proceedings of the 7th International Conference on Digital Audio Effects (DAFX-04)*. Naples, Italy, 2004.

[3] M. Belkin and P. Niyogi. *Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering.* In *Advances in Neural Information Processing Systems*, volume 14, pp. 585–591. MIT Press, 2001.

[4] M. Belkin and P. Niyogi. *Laplacian Eigenmaps for Dimensionality Reduction and Data Representation.* In Neural Computation, **15**(6), pp. 1373–1396, 2003.

[5] R. Blackman and J. Tukey. *The measurement of power spectra: from the point of view of communications engineering.* Dover books on engineering and engineering physics. Dover Publications, Dover, 1959.

[6] J.-F. Cardoso. `perso.telecom-paristech.fr/~cardoso/guidesepsou.html`.

[7] J.-F. Cardoso. *Blind signal separation: statistical principles.* In *Proceedings of The IEEE*, volume 86, pp. 2009–2025. 1998.

[8] J.-F. Cardoso. *Multidimensional Independent Component Analysis.* In *Proceedings of the International Workshop on Higher-Order Statistics*, pp. 111–120. 1998.

[9] J.-F. Cardoso. *High-order contrasts for independent component analysis.* In Neural Computation, **11**, pp. 157–192, 1999.

[10] M. A. Casey and A. Westner. *Separation of Mixed Audio Sources by Independent Subspace Analysis.* In *Proceedings of the International Computer Music Conference.* Berlin, 2000.

[11] S.-H. Cha. *Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions.* In International Journal of Mathematical Models and Methods in Applied Sciences, 2007.

[12] P. Comon. *Analyse en Composantes Indépendantes et Identification Aveugle.* In Traitement du Signal, **7**(5), pp. 435–450, 1990.

[13] P. Comon. *Independent component analysis, A new concept?* In Signal Processing, **36**(3), pp. 287–314, 1994.

[14] J. W. Cooley and J. W. Tukey. *An Algorithm for the Machine Calculation of Complex Fourier Series.* In Mathematics of Computation, **19**(90), pp. 297–301, 1965.

[15] I. Daubechies. *The Wavelet Transform, Time-Frequency Localisation and Signal Analysis.* In IEEE Transactions on Information Theory, **36**(5), pp. 961–1005, 1990.

[16] R. Der. *Blind Signal Separation.* `www-mmsp.ece.mcgill.ca/documents/reports/2001/derr2001.pdf`, 2001.

[17] D. Fitzgerald, E. Coyle and B. Lawlor. *Sub-band Independent Subspace Analysis for Drum Transcription.* In *Proceedings of the 5th International Conference on Digital Audio Effects (DAFX-02).* Hamburg, Germany, 2002.

[18] D. Fitzgerald, E. Coyle and B. Lawlor. *Independent subspace analysis using locally linear embedding.* In *Proceedings of the 6th International Conference on Digital Audio Effects (DAFX-03)*, pp. 13–17. London, UK, 2003.

[19] I. Fodor. *A Survey of Dimension Reduction Techniques.* Technical report, Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, 2002.

[20] B. Forster and P. Massopust. *Four Short Courses on Harmonic Analysis: Wavelets, Frames, Time-Frequency Methods, and Applications to Signal and Image Analysis.* Applied and Numerical Harmonic Analysis. Birkhäuser, 2009.

[21] J. B. J. Fourier. *Théorie analytique de la chaleur.* Firmin Didot Père et Fils, Paris, 1822.

[22] D. Gabor. *Theory of communication. Part 3: Frequency compression and expansion.* In Journal of the Institution of Electrical Engineers - Part III: Radio and Communication Engineering, **93**(26), pp. 445 –457, 1946.

[23] C. Geiger and C. Kanzow. *Theorie und Numerik restringierter Optimierungsaufgaben.* Springer, Berlin, 2002.

[24] M. Guillemard, A. Iske and S. Krause-Solberg. *Dimensionality Reduction Methods in Independent Subspace Analysis for Signal Detection.* In *Proceedings of the 9th International Conference on SamplingTheory and Applications (SampTA2011).* Singapore, 2011.

[25] M. Guillemard, A. Iske and U. Zölzer. *Clifford Algebras and Dimensionality Reduction for Signal Separation and Classification.* In Hamburger Beiträge zur Angewandten Mathematik, **4**, 2010.

[26] J. Herault and C. Jutten. *Space or time adaptive signal processing by neural network models.* In *AIP Conference Proceedings 151 on Neural Networks for Computing*, pp. 206–211. New York, 1987.

[27] T. Hofmann and J. M. Buhmann. *Pairwise data clustering by deterministic annealing.* In IEEE Transactions on Pattern Analysis and Machine Intelligence, **19**(1), pp. 1–14, 1997.

[28] A. Hyvärinen and P. Hoyer. *Independent subspace analysis shows emergence of phase and shift invariant features from natural images.* In *Proceedings of the International Joint Conference on Neural Networks.* 1999.

[29] A. Hyvärinen, P. O. Hoyer and M. Inki. *Topographic Independent Component Analysis.* In Neural Computation, **13**(7), pp. 1527–1558, 2001.

[30] A. Hyvärinen and E. Oja. *Independent component analysis: algorithms and applications.* In Neural Networks, **13**, pp. 411–430, 2000.

[31] M. Kendall and A. Stuart. *The Advanced Theory of Statistics.* Charles Griffin & Company Limited, 1977.

[32] S. Kullback and R. A. Leibler. *On Information and Sufficiency.* In Annals of Mathematical Statistics, **22**(1), pp. 79–86, 1951.

[33] R. Lasser. *Fourier Analysis, an introduction*, volume 199 of *Monographs and textbooks in pure and applied mathematics.* Marcel Dekker Inc., New York, 1996.

[34] J. Lee. *Introduction to smooth manifolds*, volume 218 of *Graduate texts in mathematics.* Springer, New York, 2003.

[35] J. Lee and M. Verleysen. *Nonlinear Dimensionality Reduction.* Information Science and Statistics Series. Springer, London, 2010.

[36] E. Levina and P. J. Bickel. *Maximum Likelihood Estimation of Intrinsic Dimension.* In *Advances in Neural Information Processing Systems*, volume 17, pp. 777–784. MIT Press, 2005.

[37] S. López Rosa. *Information-theoretic measures of atomic and molecular systems.* Ph.D. thesis, Universidad de Granada, 2010.

[38] L. van der Maaten. `homepage.tudelft.nl/19j49`.

[39] L. van der Maaten, E. O. Postma and H. J. van den Herik. *Dimensionality Reduction: A Comparative Review.* Technical report, Tilburg University, 2009.

[40] M. Meila and J. Shi. *Learning Segmentation by Random Walks.* In *Advances in Neural Information Processing Systems*, volume 14, pp. 873–879. MIT Press, 2001.

[41] A. Y. Ng, M. I. Jordan and Y. Weiss. *On Spectral Clustering: Analysis and an algorithm.* In *Advances in Neural Information Processing Systems*, volume 14, pp. 849–856. MIT Press, 2001.

[42] K. Pearson. *On lines and planes of closest fit to systems of points in space.* In Philosophical Magazine, **2**(6), pp. 559–572, 1901.

[43] I. Potamitis and A. Ozerov. *Single channel source separation using static and dynamic features in the power domain.* In Image Processing, 2008.

[44] L. R. Rabiner and R. W. Schafer. *Introduction to digital speech processing.* In Foundations and Trends in Signal Processing, **1**, pp. 1–194, 2007.

[45] W. Rudin. *Functional analysis.* International Series in Pure and Applied Mathematics. McGraw-Hill Inc., New York, second edition, 1991.

[46] P. Smaragdis and J. C. Brown. *Non-Negative Matrix Factorization for Polyphonic Music Transcription.* In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 177–180. 2003.

[47] M. Solazzi, F. Piazza and A. Uncini. *Nonlinear blind source separation by spline neural networks.* In *IEEE International Conference on Acoustics, Speech, and Signal Processing.* 2001.

[48] E. M. Stein and R. Shakarchi. *Fourier Analysis, an introduction.* Princeton Lectures in Analysis. Princeton University Press, Princeton, 2003.

[49] J. B. Tenenbaum, V. Silva and J. C. Langford. *A Global Geometric Framework for Nonlinear Dimensionality Reduction.* In Science, **290**, pp. 2319–2323, 2000.

[50] C. Uhle, C. Dittmar and T. Sporer. *Extraction of drum tracks from polyphonic music using Independent Subspace Analysis.* In *Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, pp. 843–848. Nara, Japan, 2003.

[51] T. Virtanen. *Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria.* In IEEE Transactions on Audio, Speech, and Language Processing, **15**(3), pp. 1066–1074, 2007.

[52] J. Wellhausen. *Audio Signal Separation Using Independent Subspace Analysis and Improved Subspace Grouping.* In *Proceedings of Nordic Signal Processing Symposium NORSIG '06*, pp. 310–313. Rejkjavik, Iceland, 2006.

[53] M. W. Wong. *Discrete Fourier Analysis.* Birkhäuser, Basel, 2011.

[54] M. Zaunschirm, J. Reiss and A. Klapuri. *A high quality sub-band approach to musical transient modification.* Technical report, QMUL, London, 2010.

[55] A. I. Zayed. *Advances in Shannon's Sampling Theory.* CRC Press, Boca Raton, 1993.

Eidesstattliche Erklärung

Ich versichere an Eides statt, dass ich die Arbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen benutzt habe.

Hamburg, 29. August 2011