A Lanczos-view on Independent Component Analysis of fMRI Data

by

ERIK HANSON

Master of Science Thesis

(Applied and Computational Mathematics)

Department of Mathematics
University of Bergen

June 2010

Acknowledgements

This thesis is a result of a semi-blind exploration of an overwhelming scientific field. I'm therefore grateful for the guidance given by my supervisors prof. Hans Z. Munthe-Kaas and prof. Arvid Lundervold. They have provided me with the basic theory and support, making it possible for me to explore methods and ideas. They have also opened doors for me by introducing me to their research fields; doors that may still be open in the future. A thank you goes also to medical student (research track) Erling Westlye for help with data organization and interpretation of processing results.

I will also send a thank you to those who have helped me with the text in the final stages of my thesis. In particular I will mention Anne Hanson, Kjersti Moe, Åsmund Kjørstad and Hans-Petter Hanson.

However, most of the work on this thesis was done in the months and years prior to the final weeks. Therefore, I will thank all the students in MIL and π-happy. You have all, in different ways, contributed to the atmosphere at Matematisk institutt.

And finally a particular thank you to my lovely Trine. You have contributed in more ways then I'll ever know.

Erik Hanson
June 2010

Summary

Analysis of resting-state fMRI data is commonly done by a combination of the two signal processing methods Principal Component Analysis (PCA) and Independent Component analysis (ICA). In this thesis, a possible error caused by the combination of the two methods are pointed out. The error is described theoretically and by several examples. Furthermore a new, alternative algorithm is introduced. The new algorithm is performing the ICA by a Lanczos method on a four dimensional tensor without a PCA preprocessing step and may thereby overcome some of the possible errors. This Lanczos-based method is suited to deal with large datasets where only a limited number of components are interesting. The convergence of the method, and thereby the ordering of the independent components, are heavily dependent of the spectral properties of the data. Without prior knowledge of the eigenvalues, the Lanczos-based method may give unsatisfactory results. Nevertheless, the framework in which the Lanczos-ICA method is based, proves to be a powerful base for future ICA methods and fMRI analysis algorithms.

Contents

Introduction

Have you ever wondered how you manage to follow a conversation when you are in a room with several people talking, music playing and other noisy disturbances? Some people (often girls) are even able to follow two conversations at the same time. Our brain has a remarkable ability to sort the information contained in the sound signals. Important and coherent parts of the information are focused on, while pieces of less important information and disturbing noise are ignored.

This ability can be compared to analysis of brain fMRI data. In this case, we are positioned in a brain, listening to the communication between different neurons and trying to pay attention to what all of them are saying simultaneously. The main difference is that instead of two ears as observation points we have thousands of voxels monitoring the signals. Furthermore, we only have a simple mathematical model and a computer to perform the analysis, not a brain. The mathematical model and main analysis tool is called *Independent Component Analysis* (ICA).

ICA is a method separating mutually independent, non-Gaussian components linearly mixed in an unknown manner using fourth order statistics. The method can be compared to *Principal Component Analysis* (PCA) which is a method using second order statistics. ICA is used in wide range of applications and theoretical foundation for ICA is well described. A good introduction is given in [28].

The derivation of ICA can be done from several different viewpoints leading to different algorithms. The most known are *FastICA* by Hyvärinen [16], *Infomax* originally by Bell and Sejnowski [1] and *JADE* by Cardoso et al. [6]. In this thesis a new ICA method is introduced. The method is based on an existing framework by Cardoso [5] combined with a classical numerical method introduced by Lanczos [18]. The method is applied to medical data from an fMRI scan. Hence, the work of this thesis is tip-toeing the borderline between numerical linear algebra, statistics and medical imaging.

From a theoretical point of view, separation of independent sources is trivial. Thus in an ideal case, all the algorithms mentioned above perform well. The main problem addressed in this thesis is ICA performed on large, fMRI datasets with noise and with an unknown number of sources. This problem is commonly solved by a combination of the two techniques PCA and ICA. In this thesis a Lanczos-ICA method is introduced as an alternative solution strategy, attempting to deal with large datasets without the use of PCA.

Thesis Structure

This thesis is arranged in the following manner:
Chapter 1 gives an introduction to the mathematical framework needed to read the rest of the thesis. In Chapter 2, we define the ICA-model and get a short introduction to the theory behind general ICA. Several ICA methods are briefly introduced with focus on the common statistical properties that form the foundation of all ICA methods. ICA based on eigenvalue decomposition is the main algoritmic framework in this thesis. Chapter 3 is devoted to this subject. Here, we also find the derivation of the new Lanczos-based ICA method. Chapter 4 puts the mathematical models into a broader setting by introducing application to fMRI data analysis. A brief introduction to fMRI as data source is given and practical considerations about the analysis are discussed. In particular, the concept of PCA as a preprocessing stage for ICA is discussed. In Chapter 5, results of ICA applied to fMRI data are presented. Methods and theoretical aspects from all the previous chapters are included in the analysis. The results from the different methods are discussed and compared. An apparent paradox in the use of PCA combined with ICA is pin-pointed, and several examples thereof are given. Chapter 6 concludes the thesis and contains an outline to further work in the field of signal and image processing. In the Appendices we find a description of the visualization methods used in this thesis as well as some algebraic simplifications and proofs.

Chapter 1

Mathematical Framework and Related Theory

In this Chapter, the mathematical tools used in the thesis will be introduced. These tools are necessary to understand the ICA-model and derivation of ICA-methods discussed in later chapters. Some signal processing methods with close relation to ICA are introduced briefly. For general statistical and linear algebra tools; see e.g. [30, 19]. The experienced reader is advised to proceed to the definition of the ICA model in Chapter 2.

1.1 Projection Pursuit

Projection Pursuit (PP) is a framework of signal processing methods coping with higher dimensional data. In general, higher dimensional data are difficult to interpret and will be more easily understood in a lower dimensional subspace. The idea behind PP methods is to project higher dimensional data onto a lower dimensional space. The projection is done in a way suitable for human interpretation. Among all the possible projections, the most interesting are selected first. Hence PP methods requires an objective measure of interest. The statistical theory described later in this Chapter offers a range of different measures. The general idea behind projection pursuit can be recognized in *principal component analysis* as well as ICA. For more aspects around Projection Pursuit see [10].

1.2 Principal Component Analysis

Principal Component Analysis (PCA) describes the interesting projections in Projection Pursuit using correlation and variance. From a zero mean dataset

$$x = \begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \end{bmatrix} \tag{1.1}$$

with n observations of the time or space dependent variable x, PCA search for a representation $y = Ux$ with maximum variance. The orthogonal base of this representation can be obtained from the Eigenvalue Decomposition (EVD) of the covariance matrix $C \in \mathbb{R}^{n \times n}$:

$$C = E\{x \cdot x^T\} = U^T D U, \tag{1.2}$$

where the D is a diagonal matrix containing eigenvalues with corresponding eigenvectors in the columns of U.

A small example is given in Figure 1.1 where a set of random 2D variables are plotted (in blue). The data is slightly positive correlated. The principal components y_1 and y_2 (in red) spannes a space where the data is uncorrelated. The data has maximum variance along the direction of the first principal component y_1.

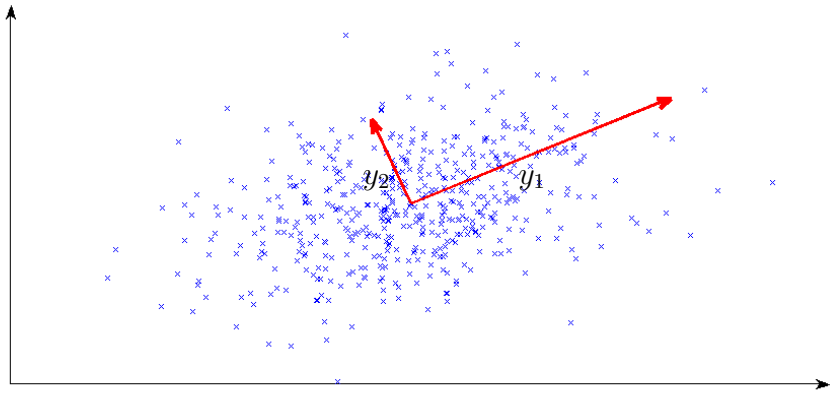


Figure 1.1: *Principal Components of a random dataset of 500 observations.*

In many applications of PCA the goal is to form a subspace containing most of the variance of x. In other words, we search for a $y \in \mathbb{R}^k$, $k < n$ representing most of the variance of the data in x. The dimensional reduction

is then performed by representing the covariance matrix by a lower rank approximation. The new covariance matrix $\hat{C}$ is constructed from the k largest eigenvalues of C. Similarly, a new $\hat{U}$ is generated from the eigenvectors corresponding to the eigenvalues in $\hat{C}$. The omitted eigenvalues of small size represents dimensions of the data spanning low variance. Hence, in terms of projection pursuit, an interesting lower dimensional estimation $y = \hat{U}x$ represents most of the variation in x. In the example in Figure 1.1 the best one-dimensional representation of the data would have been a projection onto a line in y_1-direction.

1.3 Independence

The term *independence* in independent component analysis can be described briefly as the property of having completely separate origin. Statistically, two random variables y_1 and y_2 are independent if the common probability density function can be factorized;

$$p(y_1, y_2) = p(y_1)p(y_2).$$

This is a stronger claim than zero correlation, $corr(y_1, y_2) = 0$. While zero correlation is obtained if $E\{y_1 y_2\} = E\{y_1\}E\{y_2\}$, independence also need

$$E\{g(y_1)h(y_2)\} = E\{g(y_1)\}E\{h(y_2)\}$$

where g and h are any non-linear functions. Zero correlation is a special case of independence where g and h are linear. Notice that for Gaussian variables zero correlation and independence are equivalent. For sampled data, independence can be difficult to measure, but estimates can be made using higher order statistics. This is described in Section 1.5.

1.4 Central Limit Theorem

The essence of the central limit theorem can be summarized as:

- A linear combination of k random, equally distributed, independent variables tend to Gaussian distribution as $k \to \infty$.

Even the sum of two random independent variables $a + b$ are more Gaussian then any of the variables a and b. As a simple illustration, let us consider the sum after rolling respectively one, two and three dice. Figure 1.2 shows

the histogram of the eye sums after rolling 1000 times. As only one die has a uniform histogram, it is apparent that the histogram of the sums tend to a more bell shaped form.

The assumption that the variables are equally distributed can to some extent be relaxed without loss of precision. For the most of this thesis, the central limit theorem is assumed to be valid for non-equal distributed variables as well. For further details and more formality on the central limit theorem see [30].

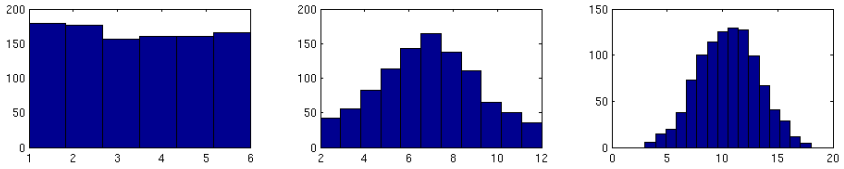


Figure 1.2: *Histograms of dice sums after rolling one, two and three dice*

1.5 Higher order statistics

Moments are used to describe properties of a variables distribution. From general statistical theory, we are familiar with a variables moments expressed using the probability density function (pdf) p_x and the mean μ of the variable:

$$m_n = E\{(x-\mu)^n\} = \int_{-\infty}^{\infty} (x-\mu)^n \, p_x(x)\mathrm{dx}$$

recognized as the Fourier coefficients of the pdf. For observed data samplings the pdf is in general unknown and estimators is used. Thus the expectation value operator $E\{\cdot\}$ is commonly evaluated as a sample mean, not based on the pdf. The moments are often expressed as central moments with $\hat{x} = x - \mu$.

Cumulants are close related to the moments. They are built of sums of products of moments, and share thereby some properties with moments. The similarities of the two concepts is apparent in their original derivation. While moments are defined as the Fourier coefficients of the pdf, Cumulants are derived as the Fourier coefficients of the natural logarithm of the pdf, rather then the pdf itself. This difference gives cumulants an extra set of nice algebraic properties not shared with the moments. These properties will be discussed later.

In ICA theory it is common to assume that the mean $\mu = 0$ and for moments of higher order than two the variance $\sigma^2 = 1$. Under this assumption, moments, central moments and cumulants are equivalent up to the third order. The first two cumulants, mean and variance, are intuitive and frequently used in statistical analysis. The third order cumulant is called *skewness* and is a measure of asymmetry in a distribution. Theory around the third cumulant will not be discussed further. The forth order cumulant is of special interest when studying ICA, thus further details follows:

In the scalar case, the fourth cumulant is known as *kurtosis*,

$$kurt(y) = E\left\{y^4\right\} - 3(E\left\{y^2\right\})^2. \tag{1.3}$$

Kurtosis can be interpreted as a measure of a tail thickness in a variables pdf. When z is Gaussian $|kurt(z)| = 0$ and kurtosis can thereby serve as a Gaussianity-estimator. The evaluation of kurtosis can be sensitive to outliers and in some applications other estimators are preferred.

In the multivariate case, all cumulants are arranged in a way suitable for tensor analysis. Still assuming $\mu = 0$, the second cross-cumulant is recognized as the covariance matrix, while the forth cross-cumulant:

$$\begin{aligned} \text{cum}(x_i, x_j, x_k, x_l) = {} & E\{x_i x_j x_k x_l\} - E\{x_i x_j\}E\{x_k x_l\} - \\ & E\{x_i x_k\}E\{x_j x_l\} - E\{x_i x_l\}E\{x_j x_k\}, \end{aligned} \tag{1.4}$$

is a tensor in $\mathbb{R}^{n\times n\times n\times n}$ containing all fourth order statistical information of x. As mentioned earlier, the cumulant tensors are constructed to ensure three essential properties:

1. Linearity, such that if $a = b + c$, $cum(a) = cum(b) + cum(c)$.
2. Gaussian minimum, such that the cumulant of a Gaussian distributed variable is zero.
3. Diagonality with independent data, such that the cross-cumulant of independent variables is diagonal.

Further details about cumulants and a more detailed list of properties can be found in [2]

In many applications one assume that the data is Gaussian and the first two moments, or cumulants, are sufficient. Since a Gaussian distribution can be described uniquely by the mean and variance, higher order moments will be irrelevant. When dealing with non-Gaussian data some of the information about the data may lie in the higher order moments. This particular concept will be paid further attention later in this thesis.

1.5.1 Negentropy

The properties of random variables and pdfs can be described by its moments, but a more general description can be made using information theoretical concepts. When dealing with ICA *negentropy* is one of the most essential information theoretic terms. Before defining negentropy it is first necessary to define *entropy*.

Entropy is a measure of how structured a random variable is. This is: to what extent is the variable random? If a variable has a tendency to take a certain value, the variable is less random then an other variable without this tendency and will thereby have a lower entropy value. Mathematically the entropy H of the random variable x is defined as:

$$H(x) \quad = \quad -\int p_x(\eta) \log p_x(\eta) \, \mathrm{d}\eta. \tag{1.5}$$

In the continuous case H is sometimes named differential entropy. Using this definition it can be shown that given a fixed variance the Gaussian distribution has maximum entropy. This fact forms the basis for the term *negentropy*.

The *negentropy* J of a variable x is expressed using the entropy in function (1.5) as

$$J(x) = H(x_g) - H(x) \tag{1.6}$$

where x_g is a random Gaussian variable with same variance as x. Hence negentropy is a measure of how far a variable is from Gaussian distribution. As the Gaussian distribution in terms of entropy is the least structured, negentropy also serves as a measure of how structured or random a variable is and as a perfect theoretical measure of non-Gaussianity.

Due to the pdf dependence in function (1.5), direct measurement of entropy or negentropy is impossible for general data samplings. Using estimators based on higher order cumulants, an approximate negentropy value is obtainable, but negentropy will for the rest of this thesis mostly serve as a theoretical tool.

1.6 Operations on Tensors

The generalisation of the correlation matrix into higher order cumulant tensors must be accompanied with a framework of suitable tensor tools. In the

second order (matrix) case, familiar linear algebra concepts as eigenvalues and eigenvectors can be used in analysis of data variation. In order to generalize these concepts we need products and norms in tensor spaces as well as a notion of symmetry in a tensor. The arrangement of tensors opens a number of possible products and operators in tensor spaces. For deeper mathematical insight in tensor operators in the framework of blind source separation see [23].

For the purpose of this thesis, a linear cumulant operator on a matrix space $F : \mathbb{R}^{n\times n} \mapsto \mathbb{R}^{n\times n}$, is given by

$$F \cdot Y = F(Y), \quad F_{ij}(Y) = \sum_{kl} y_{k,l} T_{i,j,k,l}, \tag{1.7}$$

with $T \in \mathbb{R}^{n\times n\times n\times n}$ like in Equation (1.4) and $Y \in \mathbb{R}^{n\times n}$. The operator (1.7) can be seen as the tensor analogy to a standard linear matrix operator from one vector space to another, $Ax = b$, $A : \mathbb{R}^n \mapsto \mathbb{R}^n$.

We equip our matrix space with the *Frobenius inner product*

$$\langle A, B\rangle = \sqrt{trace\,(A^T B)} \tag{1.8}$$

and state that the operator F is hermitian:

$$\langle F(M), X\rangle \;\; = \;\; \langle M, F(X)\rangle. \tag{1.9}$$

Hence, by the spectral theorem, F has an eigenvalue decomposition [11]. This is not necessary true if the hermitian condition 1.9 fails. For general tensor decompressions more relaxed decomposition criteria applies [23].

Chapter 2

The ICA Model

In this Chapter, a formal definition of Independent Component Analysis is stated. An illustrative example is presented and general ICA methods are briefly derived at the end of the Chapter.

2.1 Blind Source Separation Problem

Consider $x(t) \in \mathbb{R}^n$ as an observed set of signals generated by a mix of the original source signals $s(t) \in \mathbb{R}^n$. The mixture of s is described by the unknown mixing matrix $A \in \mathbb{R}^{n \times n}$. This gives the basic blind source mixing equation.

$$x(t) = As(t)$$

, where both A and s are unknown. For simplicity and since the model also can be used on other data then time series, we will from now on omit the t dependence in the notation:

$$x = As \tag{2.1}$$

and name this the *Blind Source Equation*.

The goal of the blind source separation is to estimate the unknown s without any knowledge of A or A's structure. The under-determination in the Blind Source Equation (2.1) may force certain weak conditions upon s in order to make the problem solvable.

2.2 Definition of ICA

Independent Component Analysis (ICA) is a special case of the more general Blind Source Separation Problem. ICA determines the components of the

source signal s up to a scalar factor using only the observation x and the two assumptions:

- The elements in s must be mutually independent.
- The elements in s must be non-Gaussian.

We will return to the motivation behind these assumptions later in this Chapter. In practical situations the estimation is done by first estimating the other unknown parameter A, or rather $W^T = A^{-1}$, and then calculate s from

$$s = W^T x. \tag{2.2}$$

A is often named *mixing matrix* and W *unmixing matrix*.

2.3 Example with uniformly distributed variables

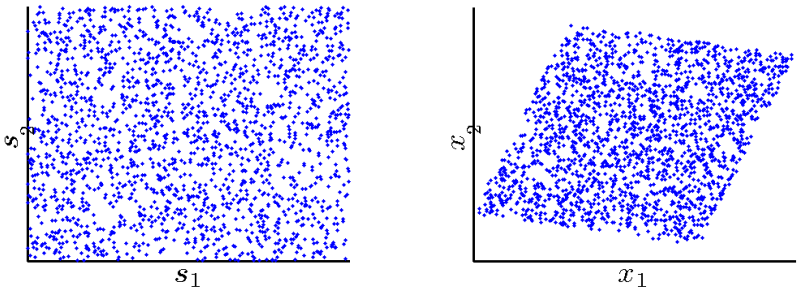
The features of ICA are easily illustrated with a simple example. Assume s_1 and s_2 are two uniformly distributed independent vectors. Due to the independence, prior information about s_1 will not give away anything about s_2 and vice versa; see Figure 2.1(a). The mixes x_1 and x_2 are represented by a linear transformation A:

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = A \begin{bmatrix} s_1 \\ s_2 \end{bmatrix}.$$

If A is orthogonal, the transformation is a rotation. We will in later sections see that such an orthogonality-assumption is reasonable. The transformation cancels the independence of the variables as seen in Figure 2.1(b). If we for instance set x_2 to its maximum value, we also know the value of x_1. This is not the case for the source signals s. ICA tries to reconstruct s by transforming, or in the orthogonal case, rotating x back to independence. This simple example is also valid in higher dimensional spaces and with more general distributions.

2.3.1 Limitations of ICA

The example above also illustrates ICA's ambiguities. As the mixed variables x are rotated until independence, an extra 90° rotation will still fulfil



(a) Uniformly distributed source signals

(b) Mixed, observed signals

Figure 2.1: *Independent variables (a) are made dependent (b) under a linear transformation*

the independence criteria and thereby give a different solution. The two solutions are equivalent when it comes to independence but the ordering of the components is changed. Also notice that the scaling of the data is not related to the independence criteria. Hence s_i can be scaled by a scalar without changing the degree of independence. ICA can thereby only estimate the independent components up to a scalar factor (positive or negative) and can not say anything about the original ordering of the components.

2.3.2 ICA and Gaussian Data

The example with the two uniformly distributed variables can also be used to illustrate why ICA on Gaussian data is impossible. The rotation to independence done in Figure 2.1 is possible because of the higher order moment properties of the uniform distribution. Most distributions have similar properties. However, the Gaussian distribution is the only distribution uniquely determined by the two first moments. In the normalized case this gives a rotation invariant distribution. This can be seen in Figure 2.2. By normalization we mean fixing the standard deviation to one. Further argumentation on normalization of data and why Gaussian variables can not be used will be pointed out in later sections. For a general argument see [15].

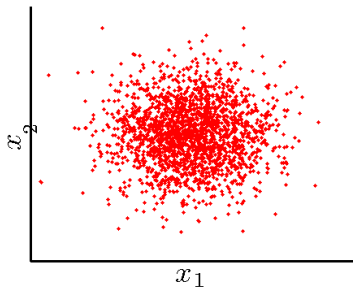


Figure 2.2: *Normalized Gaussian variables are rotation invariant.*

2.4 ICA Machinery - Maximization of Non-Gaussianity

The example in Section 2.3 points out the following strategy for ICA estimation:

- Given a suitable measure of independence, use a simple optimization routine to maximize this measure.

Hence ICA-methods can be derived from different viewpoints motivated by different independence measures and different optimisation routines. However, the statistical structures that make the methods possible are mostly the same.

The Central Limit Theorem from Section 1.4 states that a mixture of two components is more Gaussian then any of the two components themselves. An independent component is thereby most likely as far from Gaussian as possible. Under this assumption ICA becomes non-Gaussianity maximization. Using the Blind Source Equation (2.1) this is: Find a w_i satiafying

$$\max_{w} \quad [\text{non-Gaussianity}(w^T x)] \tag{2.3}$$

An estimation of one independent component is thereby given by Equation (2.2) as

$$\hat{s}_i = w_i^T x.$$

We remember from Chapter 1, that Gaussianity and non-Gaussianity can be measured in several ways.

Using negentropy or a kurtosis based negentropy estimator as objective function in a fixed point optimization is described by Hyvärinen et.al. and

named *FastICA* [16]. Alternatively an optimization of the *Maximum Likelihood* (ML) estimation of the Blind Source Equation (2.1) can be done. In ML framework assumptions about the pdfs of the components must be given. This is dealt with in a elegant manner in the *Bell-Sejnowski* algorithm. This algorithm can be also be motivated from an information theoretic point and is then known as the Infomax algorithm [1, 6, 22]. Finally, statistical information about non-Gaussianity can also be obtained from the fourth order cross-cumulant, giving the framework of tensor based ICA-methods. This framework will be further pursued in this thesis and an in-depth definition is given in the next Chapter.

Chapter 3

ICA by Eigenvalue Decomposition

Tensor decomposition is a classical framework for ICA. Several successful algorithms are based on tensor algebra. The best known algorithm is probably the *Joint Approximate Diagonalization of Eigenmatrices* (JADE) introduced by Cardoso and Soloumiac in 1993 [8, 9]. Both JADE and other tensor algorithms are based on the existence of *eigenmatrices*. Hence these methods has an analogy to the eigenvector solution of the PCA-problem in Section 1.2. In this Chapter, different methods for eigenmatrix extraction are discussed and a *Krylov Space* approach to ICA is derived.

3.1 The EVD-ICA Strategy

As we saw in Chapter 1, the variance maximization in the Principal Component Analysis (PCA) problem can be done by a Eigenvalue Decomposition (EVD) of the covariance matrix; shown in expression 1.2. A similar strategy can be applied to the ICA problem, but this time using the forth order cumulant tensor as a subject for decomposition. In Chapter 1, a matrix space operator (1.7) is introduced. We could compare this operator to the regular matrix-vector product used in the EVD in the PCA-problem. In other words: a generalized PCA can be done on the basis of this operator rather then the regular matrix operator represented by the covariance matrix. Such a generalized PCA is motivated by the higher order statistics of the fourth order cumulant.

Introducing the cumulant tensor from Equation (1.4) to the operator from Equation (1.7), we get an operator F containing all fourth order information

of a variable $x \in \mathbb{R}^n$ $F : \mathbb{R}^{n \times n} \mapsto \mathbb{R}^{n \times n}$:

$$F_{ij}(Y, x) = \sum_{kl} y_{kl} cum(x_i, x_j, x_k, x_l). \tag{3.1}$$

As stated in Equation (1.9), the operator F is hermitian. Hence an EVD of F exist and F can be spanned by a set of orthogonal *eigenmatrices* M_i satisfying:

$$F(M_i) = \lambda_i M_i, \tag{3.2}$$

for any given data set x. We know that a decomposition of F is related to the diagonalization of the cumulant tensor and this again relates to independence among the sources. In the independent case, the kurtosis values of the sources are on the diagonal of the cumulant tensor [15]. Hence the eigenvalue λ_n corresponding to each eigenmatrix M_n is the kurtosis value of one of the components. The analogy to PCA with decomposition of the covariance matrix is still valid. In PCA, the *variance* of the different components are diagonal elements in the covariance matrix when the sources are uncorrelated. Furthermore the maximization of variance is obtained by a decomposition making the covariance matrix diagonal.

In this framework, an independent component can be extracted by a three step routine due to Cardoso [5]. The routine is given here, and the details of the steps are further discussed in the following sections.

1. **Whitening**

   First the data is *whitened*. This makes the mixing matrix orthogonal and the cumulant tensor symmetric. Whitening is a fundamental step when dealing with the rest of the theory in this thesis. In other approaches to ICA, this whitening step can be omitted [1]. Furthermore Cardoso also introduce other EVD-ICA methods without this step [7]. In such applications whitening is often viewed as a preprocessing step used to encourage faster convergence. The concept of whitening is derived in detail in Section 3.2.

2. **Cumulant eigenmatrix extraction**

   From the whitened data, an eigenmatrix M, $F(M) = \lambda M$ is obtained. This can be done by e.g. a *power method*, a *Lanczos routine* or a simplified solver only searching for solutions in the space $\mathcal{W} = \{W \in \mathbb{R}\ ,\ W = w\ \cdot w^T\}$. These methods are discussed further in Section 3.3.

3. **Mixing pattern extraction from eigenmatrix**

   The eigenmatrix number i is assumed to be on the form $M_i = w_i \cdot w_i^T$, [5, 15]. A proof of this property is outlined in Appendix B. The vector w_i represents the unmixing pattern of a single independent component and is therefore the indirect goal of our routine. A full set of unmixing vectors makes out the columns of the unmixing matrix in Equation 2.2. The unmixing pattern w_i is the eigenvector corresponding to the dominant eigenvalue of M_i and can be extracted from M_i by a regular EVD. Thereby, the independent component estimate $\hat{s}$ can be expressed $\hat{s}_i = w_i^T x$. Hence a full set of independent components can be obtained from a full set of eigenmatrices M_i.

3.2 Whitening and Centering

The estimation of cumulant values involves a various of terms dependent of the data mean and standard deviation. A preprocessing stage which normalizes all data to mean *zero* and standard deviation *one* will therefore simplify cumulant estimation a great deal. In most ICA theory, the mean is assumed to be zero. If this not is the case a simple centering operation can be performed:

$$x \leftarrow x - E\{x\}.$$

This assumption is used in e.g the fourth order cumulant in Equation (1.4). Without a zero mean assumption, Equation (1.4) would have been extended drastically with terms involving the mean.

Assuming centred data, the standard deviation normalization is called *whitening* and can be done by a linear operator.

$$z = Vx,$$

giving $\sigma_{z_i} = 1$ and $cov(z, z) = E\{zz^T\} = I$.

The linear operator V can be derived in several ways. One example using an eigenvalue decomposition is:

$$V = D^{-\frac{1}{2}} E^T \tag{3.3}$$

where the diagonal matrix D and the eigenvector matrix E is the same as in the EVD of the covariance matrix in Equation (1.2). The matrix exponent in Equation (3.3) is a component-vice exponent. Normalization of the variables will not interfere with the independence of the components. In the blind

source model, only the mixing matrix is changed. This is shown in the next section.

The cumulant based operator F from Equation (3.1) will be significantly simplified when working with whitened data. The expression reduces to:

$$F(M, z) = E\{(z^T M z) z z^T\} - 2M - \text{trace}(M) I. \tag{3.4}$$

A proof for the simplification is given in Appendix B. Without this simplification, tensor based methods suffer from very high computational complexity. For further theory on different whitening methods see [15].

3.2.1 Whitening Gives Orthogonality

As a result of the whitening we get a new mixing matrix $U = VA$. This new mixing matrix will give the whitened ICA equation:

$$z = Vx = VAs = Us. \tag{3.5}$$

We can further see that:

$$I = E\{zz^T\} = E\{Us(Us)^T\} = UE\{ss^T\}U^T = UU^T, \tag{3.6}$$

thus the new mixing matrix is expected to be orthogonal as the independence among the sources s implies no expected correlation in s. An orthogonal mixing has limited degrees of freedom and is thereby simpler to estimate than a general matrix.

3.3 Methods for Eigenvalue Decomposition

Most methods for eigenvalue decomposition are known from their applications to matrix operators. In this section, we will look at some of the methods in a setting with a fourth order tensor operator.

3.3.1 Power Iteration

The power method is a classical method used to reveal eigenvectors of matrices. It also applies to our more general case with the linear operator F, shown in Algorithm 1. The method finds the eigenmatrix corresponding to the highest eigenvalue, hence the highest kurtosis value for any source.

Several components can by found by restarting the method on a subspace not containing the projection of the first component. This is possible due to the orthogonality of the whitened mixing matrix. In this manner the power iteration is a Projection Pursuit method and uses the fourth order statistics in the cumulant operator as a measure of an interesting projection.

Algorithm 1 Power(F)

Initialize: $M_0 \leftarrow$ random matrix.
  for $n \leftarrow 1, 2, 3, ...$ **do**
    $M_n \leftarrow F(M_{n-1})$
    $M_n \leftarrow \dfrac{M_n}{\|M_n\|}$
  end for
  return M_n

Simplified Power Iteration - FastICA

As described by Hyvärinen [15], the power method for finding an eigenmatrix can be simplified to only search in the space $\mathcal{W} = \{W \in \mathbb{R}, W = w \cdot w^T\}$ by using the operator $G : \mathbb{R}^n \mapsto \mathbb{R}^n$

$$G(w) = w^T \cdot F(w \cdot w^T). \tag{3.7}$$

Exploiting the algebraic properties of the cumulant tensor, this will reduce to the well known FastICA iteration:

$$w \leftarrow E\{z \sum_i w_i z_i\} - 3w. \tag{3.8}$$

Hence, cumulant based power iteration and FastICA are essentially the same when it comes to measuring independence. FastICA has the advantage that it can operate in a vector space rather then a matrix space and is thereby a computationally efficient method.

3.3.2 Lanczos Method

Assuming the existence of different eigenmatrices corresponding to F, *Lanczos method* will, using *Krylov subspaces*, simultaneously reveal several of these eigenmatices [18]. The full derivation of the Lanczos method is not given here, but some ideas are outlined. For further derivation and theory see e.g. [14, 29].

The idea behind Krylov subspace methods is to take several steps of the power method into account at each iteration. Hence a Krylov space is given as $\mathcal{K}_n = \{F(M), F(F(M))...\}$. By forming an orthogonal basis for $\mathcal{K}_n$ the Krylov space methods may be used to solve several types of numerical problems. The general eigenvalue solver using $\mathcal{K}_n$ is called an Arnoldi method and uses recursively all dimensions of $\mathcal{K}_n$ at each iteration. The symmetric edition of Arnoldi is called Lanczos method and exploit the symmetry of the operator to only use a three term recursion. This makes Lanczos methods suited to deal with large datasets in a fast manner. Recall that the fourth order cumulant operators from Equation (3.1) and Equation (3.4) are symmetric, hence Lanczos method can be applied.

The further idea behind Lanczos is to reduce the linear operator to tridiagonal form. The reduction to tridiagonal form is done in order to use simpler methods to solve the reduced problem. In the case of the two dimensional matrix operator, the reduction is done by orthogonal vectors. This is described briefly by Trefethen in [29]. When dealing with higher dimensional operators such as our cumulant function F, the same theory can be applied. A series of n orthogonal matrices $Q_1, Q_2 \dots Q_n$ reduce F to a tridiagonal matrix operator T_n,

$$T_n = \begin{bmatrix} \alpha_1 & \beta_1 & 0 & 0 & 0 \\ \beta_1 & \alpha_2 & \beta_2 & 0 & 0 \\ 0 & \beta_2 & \alpha_3 & \ddots & 0 \\ 0 & 0 & \ddots & \ddots & \beta_{n-1} \\ 0 & 0 & 0 & \beta_{n-1} & \alpha_n \end{bmatrix}.$$

According to the theory of Krylov methods, the eigenvalues of T (called Ritz-values) converge to the eigenvalues of F [14, 29]. However, the eigenmatrices M_l is a result of a slightly more involved calculation and will be revealed by a linear combination of the matrices Q_i:

$$M_l = \sum_j Q_j \xi_l(j) \tag{3.9}$$

where ξ_l is one of the eigenvectors corresponding to T. This is shown in the next subsection.

According to Krylov theory, the Lanczos method will, at each iteration, give a lower dimensional approximation of the eigenspace of F. This property can be made useful in an ICA method since an approximation of the eigenspace of F also can be used as an approximation of the independent components in the ICA model.

Algorithm 2 is performing a Lanczos routine on F. The algorithm will in each iteration obtain the best possible lower dimensional approximation of the full eigenmatrix set M_i. The approximated eigenmatrices are orthogonal under the Frobenius inner product in Equation (1.8).

Algorithm 2 Lanczos(F)

Initialize: $\beta_0 \leftarrow 0$, $Q_0 \leftarrow \mathbf{0}$, $B \leftarrow$ random matrix, $Q_1 \leftarrow \frac{B}{\|B\|}$
  for $n \leftarrow 1, 2, 3, ...$ **do**
    $V \leftarrow F(Q_n)$
    $\alpha_n \leftarrow \langle Q_n, V \rangle$ [Inner product (1.8).]
    $V \leftarrow V - \beta_{n-1} Q_{n-1} - \alpha_n Q_n$ [Three term recursion.]
    $\beta_n \leftarrow \|V\|$
    $Q_{n+1} \leftarrow \dfrac{V}{\beta_n}$
  end for
  $T \leftarrow tridiag(\alpha, \beta)$
  $\Xi, \lambda \leftarrow eig(T)$
  for $l \leftarrow 1$ to n **do**
    $M_l \leftarrow \sum_{j=1}^{n} Q_j \xi_l(j)$
  end for
  return M

Constructing Eigenmatrices from Q and T

We will now show that the matrix M_l expressed in Equation (3.9) actually is an eigenmatrix. Hence we will show that a M_l given by $M_l = \sum_j Q_j \xi_l(j)$ is satisfying

$$F(M_l) = \lambda_l M_l. \tag{3.10}$$

This can be verified by considering an expansion of the eigenmatrix condition (3.10)

$$F(M_l) = F(\sum_j Q_j \xi_l(j)) = \lambda_l M_l = \lambda_l \sum_j Q_j \xi_l(j). \tag{3.11}$$

By first examining the left hand side of (3.11) using the linearity of F we get

$$F(M_l) = F(\sum_j Q_j \xi_l(j)) = \sum_j \xi_l(j) F(Q_j).$$

Further we need to use the information contained in T from the Lanczos routine. First of all, T can be expressed using Q: $T_{i,j} = \langle Q_i, F(Q_j)\rangle$. The EVD of T will then give:

$$T\Xi = \Xi\Lambda \quad \Rightarrow \quad T_{j,j} = \sum_k \xi_k(j)\Lambda_{k,k}\xi_j(k), \tag{3.12}$$

where Ξ is a matrix containing all the eigenvectors of T and Λ is a diagonal matrix of eigenvalues. Notice that according to Krylov theory, the eigenvalues of T converge to the eigenvalues of F.

Still considering the left hand side of (3.11), we introduce Q_j and the EVD of T:

$$Q_j F(M_l) = \sum_j \xi_l(j)\langle Q_j, F(Q_j)\rangle = \sum_j \xi_l(j)T_{j,j} = \sum_j \xi_l(j)\sum_k \xi_k(j)\Lambda_{k,k}\xi_j(k) = \sum_{j,k}\lambda_k\xi_l(j)\xi_j(k)\xi_k(j) = \sum_{j,k}\lambda_k\delta_{l,k}\xi_k(j) = \lambda_l\sum_j \xi_l(j). \tag{3.13}$$

While also introducing Q_j to the right hand side of Equation (3.11) we get:

$$\lambda_l Q_j \sum_j Q_j\xi_l(j) = \lambda_l \sum_j \langle Q_j, Q_j\rangle\xi_l(j) = \lambda_l\sum_j \xi_l(j). \tag{3.14}$$

This verifies Equation (3.10) and thereby the eigenmatrix assumption on M_l since the left hand side (3.13) equals the right hand side (3.14)

Simplified Lanczos?

It may be tempting to use a simplified formulation of F in the Lanczos method, such as G in Equation (3.7) in the simplification of the power method. The Krylov-like space corresponding to G is of lower dimension than the one corresponding to F, but G is no longer a linear operator, and further examination of the theory on non-linear operators is needed to develop such a method. However it can be shown that a non-linear method will not have the same nice properties as the FastICA algorithm. Thus, we can not change all matrix operations to vector operations as is done in FastICA.

3.4 Practical considerations

Not all theoretical aspects can fully be reproduced in real life examples. In this section some practical considerations will be pointed out in order to put the theory in a different perspective.

3.4.1 Independence of sampled data

One of the basic assumptions in ICA is that the initial sources are independent. This condition is used widely in the derivation of the ICA theory. However, sampled data rarely have this property in practice. This can not necessarily be interpreted as if all sampled data are dependent, but may be a result of the insufficient representation of the data due to the sampling, some unexpected dependence in the data, or the properties of the expectation operator $E\{\cdot\}$.

Due to the mentioned uncertainties, the covariance matrix may differ from identity even though the original sources are expected to be independent:

$$ss^T \neq E\{ss^T\} = I \tag{3.15}$$

This may have impact on the theory behind the estimations in several ways.

Loss of orthogonality

From Equation (3.6) in the Section about whitening, we have that

$$UE\{ss^T\}U^T = I, \tag{3.16}$$

where U was the new mixing matrix after whitening. We see that due to the inequality of Equation (3.15) one can no longer argue that U is orthogonal.

Loss of Kurtosis-Eigenvalue Relation

In Section 1.5 on higher order statistics, one of the requirements for a diagonal cumulant tensor is independence among the sources. However, we notice that a eigen-decomposition of the operator F in Equation (3.1) always is possible due to the hermitian properties of the operator, but the expected relation between eigenvalues and kurtosis values of the sources can no-longer be obtained.

Full-Rank Eigenmatrices

The expected loss of orthogonality in the mixing matrix is apparent in the eigenmatrices obtained from the Lanczos routine. These matrices can no longer be expected to be rank one. Thus they can not be represented as a outer-product ww^T. We also find that they not satisfy the eigenmatrix criteria from Equation (3.10).

3.5 Lanczos-ICA Routine

Inspired by the practical considerations discussed above and the Cardoso EVD framework, we will now introduce a extended version of the Lanczos method (Algorithm 2) suited to deal practical ICA problems. The main idea behind the method is to neglect the fact that the cumulant tensor is decomposed in an undesired manner and force the variable to stay as close as possible to the theoretical model. Letting the Lanczos process run as previous described, the eigenmatrices generated by Algorithm 2 are not necessary rank one as expected. The routine should furthermore go on as planned an pick the dominant eigenvalue of M (if any) to represent the unmixing vector w. The full set of unmixing vectors form the columns of the unmixing matrix W. This is the same matrix as defined in the definition of the ICA model in Equation (2.2). Due to the mentioned exceptions from the theoretical model W can not be expected to be orthogonal. Thus the method can be further modified by orthogonalizing the W by e.g. a QR-routine as a post-processing step. This last step is optional and may not give a better unmixing result. For implementations of QR-methods see [29].

The full Lanczos-ICA routine is presented as Algorithm 3 and takes a mixed signal x and an iteration number i as input, and returns estimates of i independent components. This algorithm serves as a complete ICA algorithm. Due to the properties of the Lanczos method, it is suited to deal with large datasets when the number of desired independent components is less then the number of observations. The method is tested on fMRI data. Results of the test are given in Chapter 5.

3.6 Convergence of Lanczos Methods

The convergence of the Lanczos method is dependent of the spectre of the eigenvalues of F. That is: how the eigenvalues are spread on the line of real numbers. The typical case is when the spectre has some outliers. Then the

Algorithm 3 Lanczos-ICA(x,i)

[i independent components are extracted from the mixed observations in x]
$z \leftarrow Vx$ [Whitening using Equation (3.3).]
Define: $F(M) = F(M, z)$ [From Equation (3.4).]
$M \leftarrow Lanczos(F)$ [Set of eigenmatrices obtained from Algorithm 2.]
for $n \leftarrow 1, ...i$ **do**
$\xi \leftarrow \text{eig}(M_n)$ [Truncated eigenvalue decomposition.]
$w_n \leftarrow \xi$ [first eigenvector of M_n]
end for
$W \leftarrow \text{qr}(W)$ [Optional orthogonalization]
$\hat{s} \leftarrow W^T z$
return $\hat{s}$

convergence will be rapid towards these values. The convergence of the rest of the values follows and clustered values converge slowly. Hence Lanczos method is ideal in situations when only a limited number of eigenvalues are required and the outliers are interesting.

For further details and convergence theorems, the textbooks by Golub and Trefethen are recommended reading [14, 29]. As a short summary we note that the convergence is described using the Chebyshev points; that is the roots of the Chebyshev polynomials (displayed below). One result is that the worst case convergence is found when the eigenvalues lie in the Chebyshev points. This is considered unlikely in most cases, also when dealing with the fMRI data in this thesis.

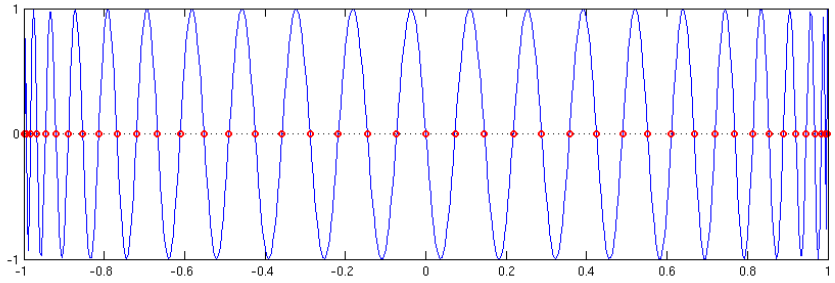


Figure 3.1: *Chebyshev polynomial of degree 43 with corresponding Chebyshev points.*

Chapter 4

fMRI Data

In clinical and cognitive neuroscience research it is useful to record and visualize the neural activity in the human brain. Using different measurement methods, brain activity patterns can be obtained while the patients are performing specific tasks. In recent studies resting-state activity maps have also been used [13]. Among the different imaging methods we will in this thesis only look at *fMRI*.

Functional Magnetic Resonance Imaging (fMRI) is an indirect measure of neural activity. The measured quantities are blood-oxygenation-level-dependent (BOLD) contrast based on the different magnetic properties of oxygenated and deoxygenated blood. An activation of a neuron population in the brain causes local oxygination of blood and thereby a measurable change in magnetic resonance. The recording is repeated over time creating a time-series of assumed electrical activity in the brain. This time variation gives rise to the term *functional*, as a contrast to the snapshot information given in a single (structural) MR image. A brief introduction to the subject is given in [26].

The deidentified image data[1] used in this thesis is a fMRI dataset from a resting-state study reported in [21, 32].

4.1 Data Interpretation

Each recorded BOLD time sample is assumed to contain a mixture of signals from different parts of the brain. The goal of the data interpretation is to isolate different origins of signals, hence find spatial isolated regions in the

[1] Courtesy of the "Cognitive Aging Project / Bergen" (PI: prof. Astri J.Lundervold)

brain associated with different time-courses. These regions accompanied with their times-courses can again be associated to isolated brain controlled tasks.

The anatomy of the human brain is widely studied. Models for brain activity can be made and the BOLD signals can be fitted to different models. This is useful when a particular task-related region is studied. In such studies, information about the time-course is known as the patient is instructed to do certain tasks or is inflicted by certain stimuli at given time intervals. When less is known about the time-courses such a methodology is unsatisfactory. This is the case in resting-state studies where the patient is instructed to lie still with closed eyes while scanned [13]. In such settings so called data driven methods are used. ICA is on of them.

4.2 fMRI Analysis Using ICA

Basically, fMRI allows capturing of a $k \times l \times m$ 3D-spatial brain activity map over n time samples resulting in a 4-D array of data. Prior to any analysis, the 4-D array is reshaped into a large matrix,

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \end{bmatrix}$$

where $x_i \in \mathbb{R}^{k \cdot l \cdot m}$, $i \in \{1, n\}$. Hence, in this representation, the spatial data points are stretched into a single vector for each time sample. This is done to prepare statistical analysis in the spatial domain. The notation from the Blind Source Equation (2.1) is still valid yielding:

$$X = AS. \tag{4.1}$$

The dataset used in this thesis consists of a $64 \times 64 \times 25$ voxel grid observed in 256 temporal samplings. The spatial data dimension is thereby $64 \times 64 \times 25 = 102400$ and the data matrix X has dimensions 256×102400.

ICA on fMRI data is discussed in [3]. Here, group studies of several individuals are also introduced. Since the ordering and scaling of the components are undefined after ICA it is difficult to compare results from different analysis processes on different subjects. In [3] this is solved by doing ICA on the entire group contemporaneously. This is done by including all data from all the subjects into a large data matrix. Data reduction methods is required to make the problem practically solvable. Data reduction is paid further attention to in Section 4.3.

In general, ICA can be performed in the temporal domain of the fMRI data, maximizing the temporal independence of the data, or in the spatial domain, maximizing the spatial independence of the data. In most ICA algorithms computational complexity is strongly dominated by the number of data observation points and not the number of samplings at each observation point. The temporal approach to ICA uses each spatial voxel as an observation (102400 in our case) causing a large number of observations. Therefore it is common to prefer the spatial approach to ICA on fMRI data where the number of observations is limited to the number of temporal sampling points (256 in our case) and thereby more suited for computationally demanding ICA algorithms.

Notice that for all ICA methods both the sources and the mixing matrix are obtained. For spatial ICA the columns of the mixing matrix represent the time-courses of the independent components. However, independence is achieved in the temporal domain hence the time-courses are not necessary independent. An illustration of the spatial approach to ICA on fMRI is shown in Figure 4.1.

4.2.1 Noise

fMRI data have, as all other sampled data, some level of noise. This will cause problems for data analysis and interpretation. However, ICA may perform well on some noisy dataset. Thus the performance of ICA in a noisy environment is dependent of the structure of the noise.

In a general setting it is common to assume that noise is Gaussian and additive on the observation. This yields the noisy ICA model:

$$x = As + n \tag{4.2}$$

where n is Gaussian noise. Due to the Gaussian properties of the noise, it will not interfere with the relative nongaussianity measurements in the ICA algorithms. In other words: The noise will make all components more Gaussian and thereby flatten the optimization landscape but not move the minima and maxima. Hence the algorithms obtain the same independent components but they are as noisy as the observations and the optimization will be more inaccurate.

From an application point of view, noise in fMRI data can be interpreted as signal without neural origin. Hence disturbance caused by blood flow in larger vessels, movement of the brain under scanning etc. is considered as noise. This kind of noise differs from the Gaussian noise as it will be

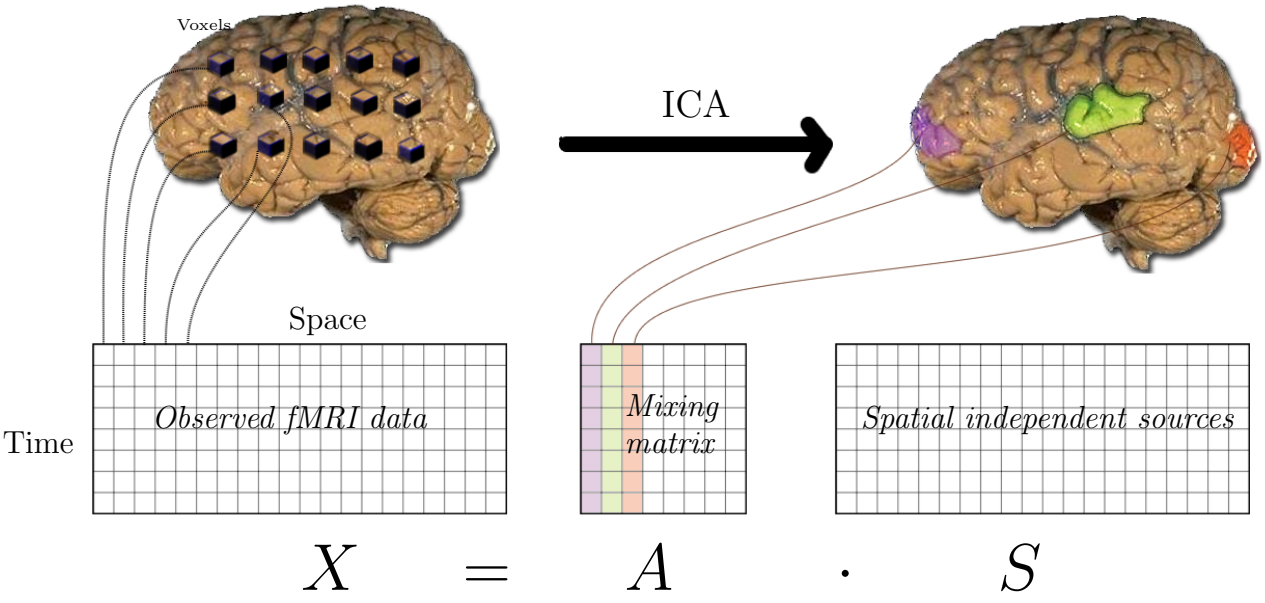


Figure 4.1: *ICA on fMRI data represented by the Blind Source Equation (4.1). The matrix X consist of one column for each voxel in the BOLD recording. The independent sources S and the mixing matrix A can be estimated using ICA. The columns of the mixing matrix are time-courses for the spatial independent components.*

included in the ICA model and appear as separate components in processed data.Thus this kind of noise can manually be removed as a post-processing routine. Notice that from a mathematical point of view these components are as real as the components of neural origin. The noise components are of special interest as they often have other statistical properties than the rest of the component set.

4.3 Coping with the Data - Preprocessing

We are faced with two problems when doing ICA on fMRI data. First of all, the data is noisy and even if components can be extracted from a noisy dataset (as mentioned in the previous section), the result may be hard to interpret. The second problem is the data size. The large number of sampling points obtained in a 3D scan of the brain results in high memory usage. Combined with the complexity of the ICA algorithms the processing time

for such large datasets will be unsatisfactory.

4.3.1 Time Filtering

A common way to remove noise is by low-pass filtering [2]. This is done under the assumption that the noise has higher frequency than the signal. In the case of fMRI studies, this is a plausible assumption in the temporal domain since the BOLD recordings will have relative low frequency. However low-pass filtering will smoothen the data and make it more Gaussian. Thus as a preprocessing to ICA, filtering is not recommendable in the same domain as nongaussianity is measured.

The data used in this thesis is low-pass filtered in the temporal domain.

4.3.2 Dimensional Reduction

In a 4-D fMRI data set each voxel contains a time course and thereby a potential component. Only a few of these components are of interest. The large amount of sources represent a huge practical challenge. It is therefore common to reduce the number of sources using regular PCA on the original data. This will result in a lower order data set suited to be analysed by ICA. A short introduction to data reduction by PCA was given in Section 1.2.

Dimensional reduction is of even greater importance when dealing with group studies such as in [3] and [32]. The combination of PCA and ICA will from now on be referred to as PCA-ICA and further discussed in later chapters.

4.4 Lanczos and Projection Pursuit vs. PCA reduction

Lanczos method is designed to deal with large systems. The Lanczos-ICA method described in Section 3.5 can therefore be applied to fMRI-data with a limited level of preprocessing. Letting the Lanczos routine run for a as many iterations as the desired number of independent components, the results can be compared to *PCA-ICA*. An alternative approach is to use FastICA as a Projection Pursuit method, and stopping the process after a desired amount of components are obtained. This will from now on be referred to as *PP-ICA*.

4.5 Selecting the Number of Components

Reducing the dimension of the dataset raises an additional question. How many components should be estimated? This question can be answered both from a mathematical and a neuroimaging point of view. From a neuroimaging view it turns out that only a handful of components are interesting hence the number of estimated components should be low. However, it is difficult to make the ICA algorithm choose the correct components, thus many components are often estimated and the selection of components must be done manually.

From a mathematical point of view the number of components should be selected in a way including as much information as possible without making the ICA methods impractical when it comes to computational time. Moreover, practical experiments indicate that over-determined data sets with noise may cause errors in the analysis. An over-determined data set is obtained if the real number of sources is less then the number of observations. In these cases the number of components are normally selected as the assumed number of original sources. However, in a blind and noisy situation, it may be difficult to distinguish between noise and sources. A method for automatic selection of data dimension in suggested in [20]. This method refers to the strategy behind PCA-ICA and takes both second and fourth order statistical information into account when estimating the number of original sources. Note that the method only gives a desired number of components, not the reduced components themselves. When studying a single subject using GIFT [4] the number of components used is 43. GIFT has a built in implementation of the method in [20].

Choosing number of components under Lanczos-ICA and PP-ICA involves considerations about the convergence properties of the methods. This will be discussed in the next chapter.

Chapter 5

Observations and Remarks

In this chapter, we find results after fMRI data analysis. The data used are from a single subject in the fMRI dataset from Chapter 4. The analysis is done by the three ICA-methods: PP-ICA, PCA-ICA and Lanczos-ICA presented and discussed throughout Chapter 3 and 4. The ICA step of the two first methods is done by FastICA. All algorithms are initialized to find 43 independent components as noted in Chapter 4.3. Information about the visualization of the obtained components is found in Appendix A.

5.1 General observations

Compared briefly to similar work and other models, the ICA gives reasonable results for all three methods. They agree on parts of the result. Moreover, there are some differences and trends among the algorithms and some errors are expected. These trends are summarized in this section and possible explanations are given here and later in the chapter.

5.1.1 The PP-ICA algorithm

PP-ICA has a tendency to find temporal small but high intensity regions. An example is found in Figure 5.7(a). FastICA searches for components by maximizing kurtosis or kurtosis-based estimators. Thus the result is not surprising since the observed regions can be considered as spiky regions or outliers with very high kurtosis values. The result is to some extent also explained by the concept of *over-learning* discussed in Section 5.2. Furthermore, the arise of spiky components are not unique for the PP-ICA method and the other two methods also finds some components with this property.

5.1.2 The PCA-ICA algorithm

Given the mentioned data, PCA-ICA has a tendency to find larger connected regions and double components. By double components we mean component containing two ore more unconnected spatial regions of high intensity. An example is given in Figure 5.7(b). Some of these components might be mathematically false. This is discussed further in Section 5.5 about the PCA-ICA paradox. Signal-to-noise-ratio is high in all obtained components using PCA-ICA. The outcome of the PCA-ICA analysis is similar to the outcome of a analysis by GIFT [4]. Such a similarity is expected since algorithms used in GIFT are the same as used in PCA-ICA in this thesis.

5.1.3 The Lanczos-ICA algorithm

The Lanczos-ICA algorithm (Algorithm 3 in Chapter 3) gives a result with similarities to both PP-ICA and PCA-ICA. One notable difference is the duplication of several components. By duplication we mean that two ore more components have close-to-identical spatial appearance. Hence after 43 iterations, only 10 different components are clearly found. In addition several more weak components appear. It is not clear whether these weak components have converged properly. They consist of numerous unconnected regions and are thereby assumed to be a mix of several components. An example is given in Figure 5.8(a). The duplication of the components may be caused by several factors. The most likely explanation is by so-called *Ghost Eigenvalues* in the Lanczos eigen-decomposition. This is a result of numerical errors. The ghost components are further discussed in Section 5.4. An other plausible explanation of the duplicated eigenvalues is that the dataset has several components with similar kurtosis values. This is not expected but can be solved by numerical techniques not discussed here. For further reference; see [8].

The development at each iteration of the Lanczos-ICA for the fMRI analysis is displayed in Figure 5.1. The rapid convergence among some of the components can clearly be spotted. The figure is further discussed in the next sections.

5.2 Over-learning

If the number of components is large and the number of sampling points is relatively small, the estimation may suffer from *over-learning*. The analysis

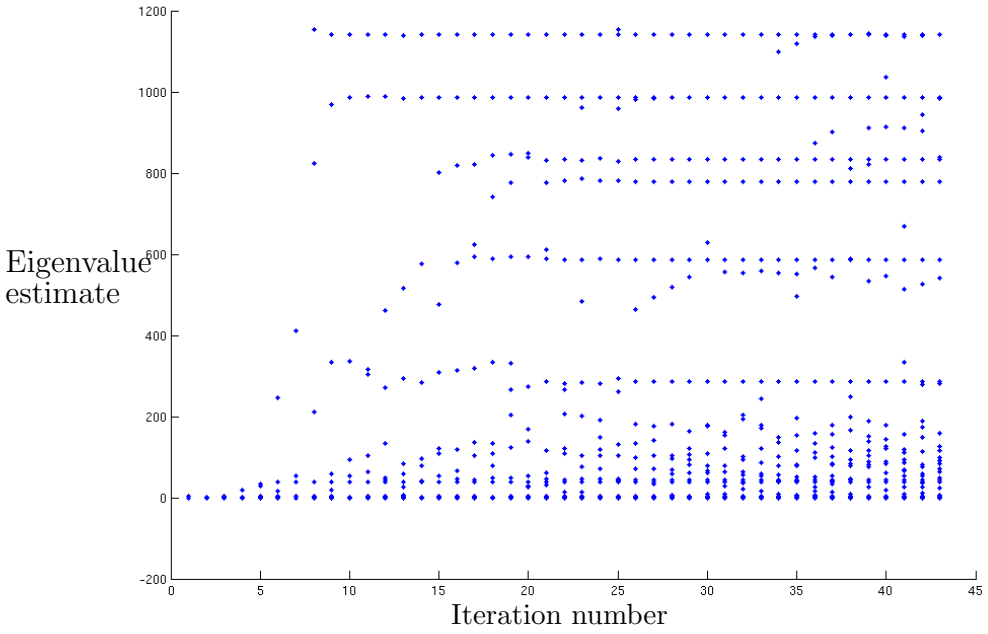
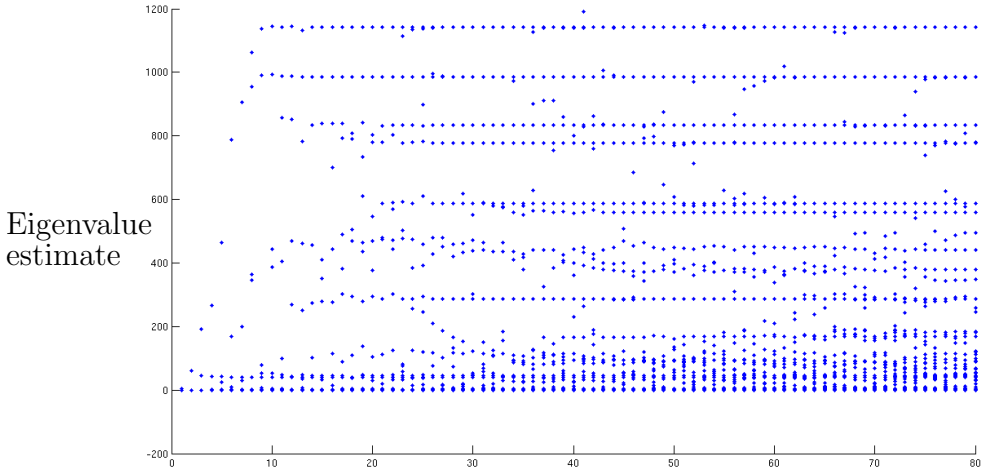


Figure 5.1: *Eigenvalue estimates (Ritz values) at each iteration of the Lanczos method. The observations on this plot follows the observations from the visualized independent BOLD components. Some Ritz values converge fast. These values correspond to the clear components such as in Figure 5.8(b). Other Ritz values have convergence difficulties. These values corresponds to the unclear components. The duplication of components can be recognized as duplication of some fast convergent Ritz values.*

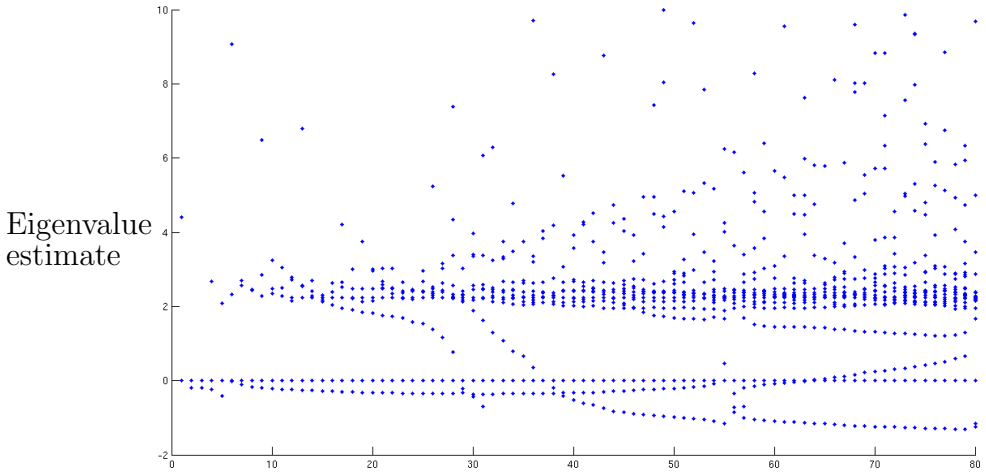
is then heavily dependent of the particular sampled values rather then the process that generated the data. The concept of over-learning is adapted from Neural-Networks theory. ICA in i Neural-Network setting is described in [15], were over-learning is discussed in more detail.

Over-learning may give components with small spiky regions. This is because the optimization in the ICA-algorithm will not be sufficiently controlled by the data, hence the components can freely be modified to obtain a maximum. For kurtosis based algorithms a spike gives maximum, and small high-intensity regions can be expected.

Spatial ICA on fMRI data should apparently not suffer from over-learning since no more then 256 observations are used and each observation has 102400 spatial samples. (Described in Chapter 4.2). Nevertheless, numerical experiments show that the the risk of over-learning may increase under certain



(a)



(b)

Figure 5.2: *Eigenvalue estimates at each iteration of the Lanczos method. 5.2(a) shows the full spectrum of Ritz values and 5.2(b) shows the spectrum in the interval* $[-2, 4]$*. The Figure can be compared to Figure 5.1, but in this plot 80 iterations are done.*

conditions [17]. These conditions are: Failure to fulfil the independence criteria, hence a small dependence among the sources, and high amount of components containing mostly noise. Both these conditions are to some extent valid for the fMRI data, hence over-learning may be expected. The

noise condition plays a particular role when the number of observations is higher then the actual number of sources. In the spatial ICA in this thesis the number of observations is 256 and the assumed number of sources is 43.

5.3 Component Clustering

An analogy for the over-learning problem can be found in the Lanczos-ICA procedure. The same conditions that caused a risk for over-learning, dependence and noise, have impact on the Lanczos algorithm. In Section 3.5 we discussed problems around not fulfilling the eigenvalue criteria. This is typically a result of not fulfilling the independence criteria. In Figure 5.1 and Figure 5.2(b) we see a dense cloud of Ritz values in the area close to zero. These values may correspond to the components containing mostly noise. We also see that extreme values among the Ritz values converge rapidly. These values correspond to the spiky components. Figure 5.3 displays the estimated spectrum with 43 components. According to the convergence theory from Chapter 3.6, the Lanczos-ICA method will be suited to fast find a smaller number of components with high kurtosis value, but can not be expected to find components with kurtosis values in the cloud closer to zero.

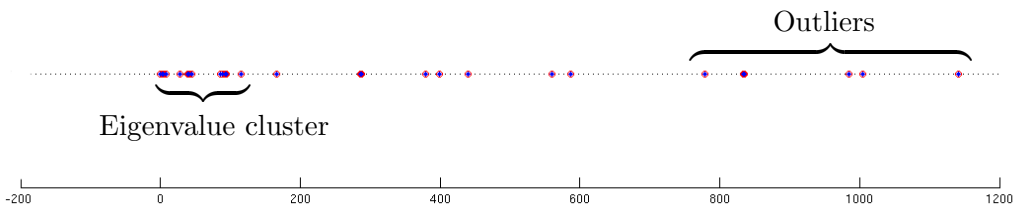


Figure 5.3: *Estimated eigenvalue spectrum. The convergence is assumed to be good for the outlier values while the values in the cluster close to zero is assumed to have slow convergence.*

5.4 Ghost Components

One of the more undesirable properties of the Lanczos method is the effect of round-off errors. This comes to the surface as false duplication of some eigenvalue estimates. The phenomenon is called Ghost Eigenvalues and is a result of the tree-term recursion in the orthogonalization part of the algorithm. Round-off errors in the orthogonalization make the algorithm forget some of

the early discovered Ritz-values and rediscover them at a later iteration step. This effect is further described in [25, 29].

The raise of Ghost Components in the Lanczos-ICA routine on fMRI data may be observed on the convergence plot in Figure 5.2(a). Only strong fast-converging components are duplicated, hence the ghost components may be seen as a guarantee that a certain component has converged properly and represent a genuine result.

5.5 The PCA-ICA Paradox

From a cumulant viewpoint there is an apparent analogy between PCA and ICA: ICA is the fourth order cumulant generalization of the second order cumulant factorization in PCA. This close relation is one of the reasons why PCA is viewed as a good preprocesser for PCA. Hyvärinen gives an example where over-learning ruins the ICA without a PCA preprocessing step [15]. Nevertheless, we will in this section look at other examples where the PCA preprocessing stage can be of more harm then use.

5.5.1 Loss of Fourth order Information

As mentioned in Section 4.3, the higher order information exploited in ICA algorithms introduces large calculations and reduces ICAs ability to operate on large datasets such as a group fMRI study. It is therefore common to reduce the data dimension to a lower-dimensional subspace containing the waste majority of the data variation using PCA. The data reduction is, in most applications, only based on second order cumulant information. The fourth order cumulant information considered in ICA is completely ignored. The result may be that important forth order cumulant properties are lost in the preprocessing stage.

PCA sorts the data according to standard deviation (variance), while ICA sorts according to kurtosis. Both methods leave out data corresponding to the lowest values. Fourth order statistical information can then be lost if the data is arranged in a different way when sorting according to kurtosis compared to sorting according to standard deviation. Hence a variable with low standard deviation, but high kurtosis value is in risk of being left out. We find this particular property among e.g. t-distributed variables. Looking to observed samples, this phenomenon is apparent when a close-to-constant observation has a few outliers. The outliers boost the kurtosis, but not the standard deviation. In the framework of fMRI images, this may be a small

temporal, but high intensity component in a image with few other artefacts and little noise.

5.5.2 1D Example

The most simple example is given by set of close-to-independent sources. Consider a set of 7 sine wave sources. The separate sources have periods given by different prime numbers with a small random parameter. A mix of these sources can be decomposed in a good manner with FastICA.

When only considering the second order cumulant information the data variation can be described close to perfectly in a reduced 5-dimensional subspace. In light of data reduction, it can be tempting to reduce the data dimension by two, as little information apparently is lost. In Figure 5.4 the result after both reconstructing the sources with a full and a reduced dataset is displayed. We also see the dominance of the first 5 eigenvalues of the second order cumulant, the covariance matrix, which indicates that the 5-dimensional subspace used in the last ICA calculation represents most of the data variation. From the figure it is apparent that the dimension reduction results in a low quality ICA result. The obtained components are still a mixture of the sources.

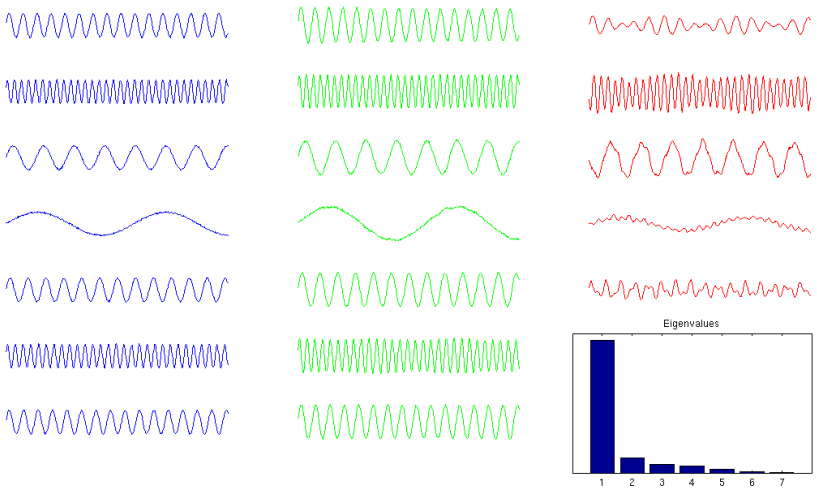


Figure 5.4: *Original sine wave sources (left) are mixed and reconstructed by FastICA (centre) and PCA-FastICA (Right). The amount of variance represented by each of the covariance eigenvalues are displayed in the eigenvalue bar chart.*

5.5.3 2D Example

A 2D example can be generated by making a set of 2D images with a random sized and placed square or disk. Such a dataset with 12 different sources is shown in Figure 5.5. When each of the spacial images are assigned to a time-course sampled with n sampling-points, the data can be mixed, and spatial ICA can be done based on the n observations of the mixed images. This is analogous to the spatial ICA done on fMRI data.

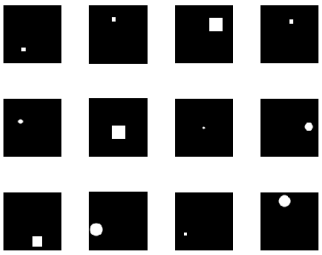
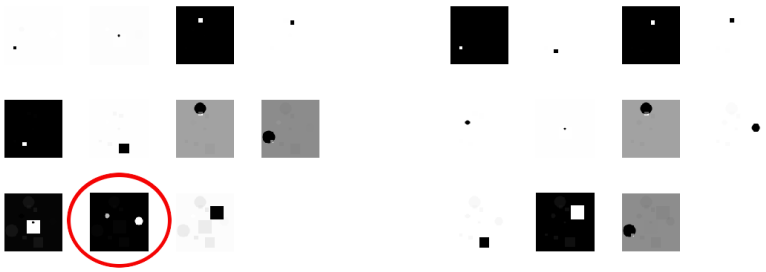


Figure 5.5: *Sources from synthetic random dataset. Each source image is associated to a time-course.*

This example uses FastICA in order to extract 11 images in two different ways. In 5.6(a) the dimension of the data is reduced to 11 by PCA before FastICA is applied (PCA-ICA). In 5.6(b) FastICA is applied to the full dataset and stopped after 11 components are obtained (PP-ICA). As expected, none of the methods can give information of the scaling of the components. Combined with prior information of the scaling, both methods give rather good reconstructions of the images up to unit sign. However, this example also illustrates one major drawback with the PCA-ICA procedure: As indicated earlier, the preprocessing PCA step disturbs the original ICA model. This can be observed as one of the originally independent components are combined to one single. The ringed out component in the bottom line in Figure 5.6(a) is a false component representing two independent sources from Figure 5.5. We also observe that both methods experience problems with spatial overlapping regions as described by Daubechies et.al. in [12].

5.5.4 fMRI Example

The last example is made by using a real fMRI dataset from a single subject. Figure 5.9 shows an independent component from a PCA-FastICA computation with a PCA step reducing the data dimension to 43. Figure 5.10(a) and



(a) PCA-FastICA (b) Projection-Pursuit-FastICA

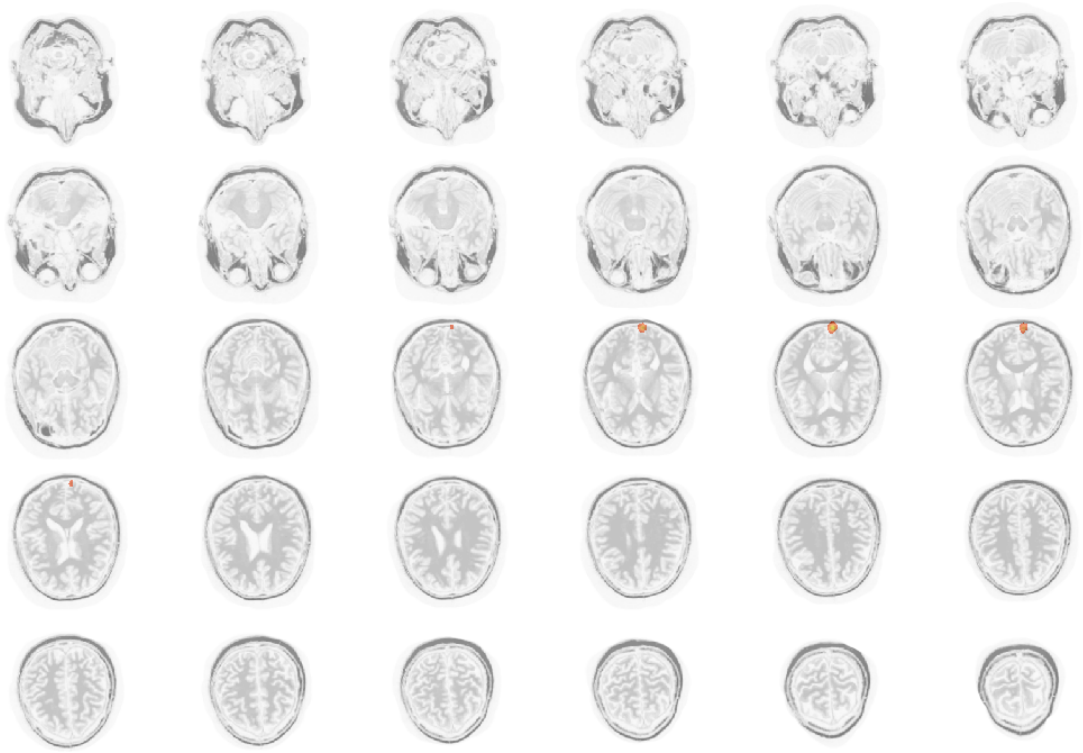
Figure 5.6: *Dimension reduction with PCA prior to FastICA (a) can cause false components (ringed out). Projection Pursuit approach to same problem never introduce false components (b)*

.

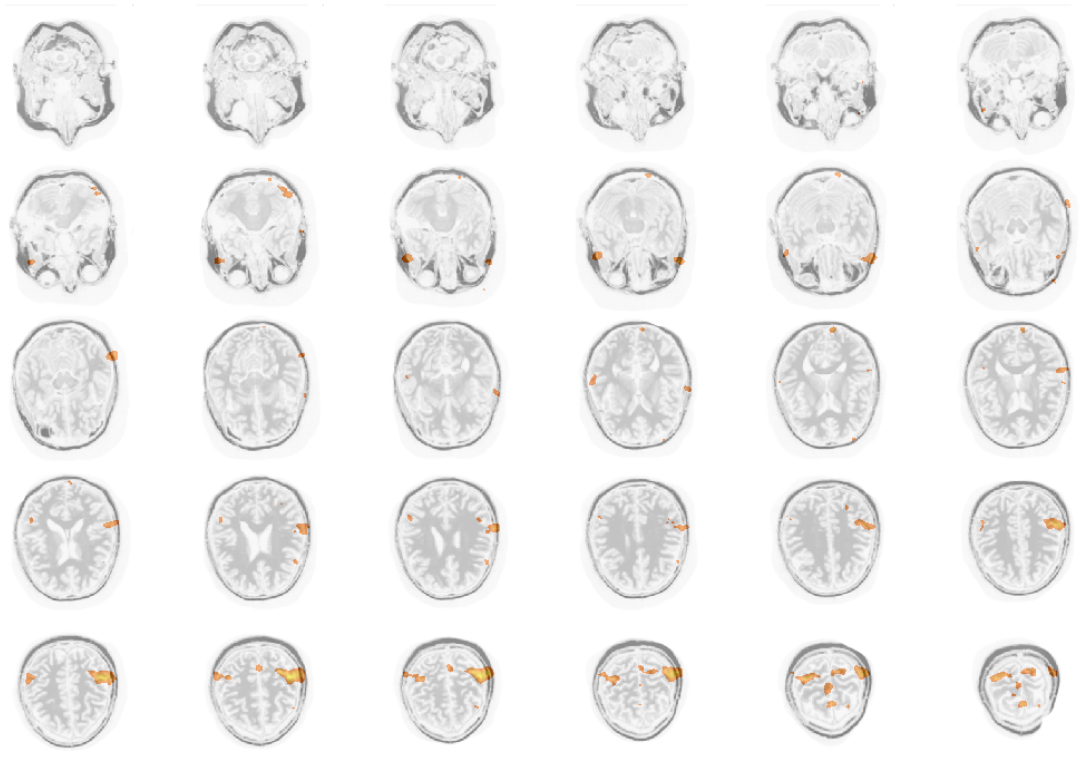
5.10(b) shows two components from a Projection Pursuit FastICA (PP-ICA) computation over the entire dataset. The component in Figure 5.9 can be expressed as a linear combination of the to components in Figure 5.10(a) and 5.10(b). Since the data is from a real dataset, we do not know the original sources. A comparison to the example above with synthetic data will indicate that the PCA-ICA result in Figure 5.9 is a false double component.

5.6 Components

Visualization of some of the obtained components follows...

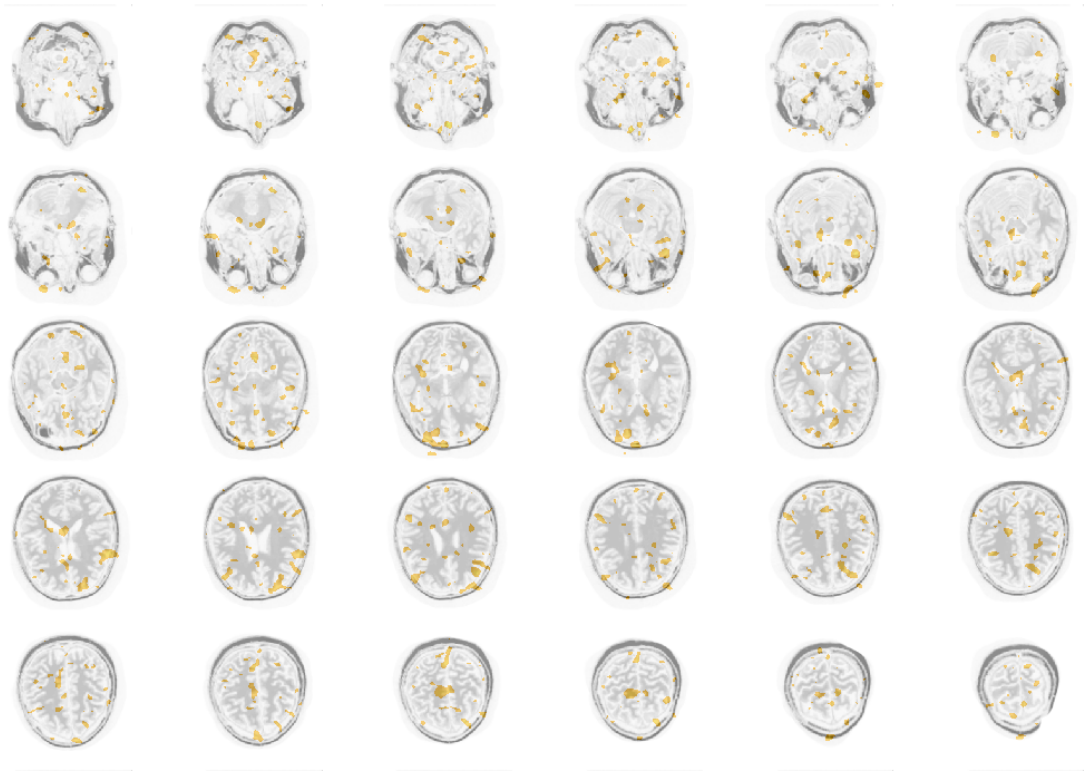


(a) Component extracted using PP-ICA

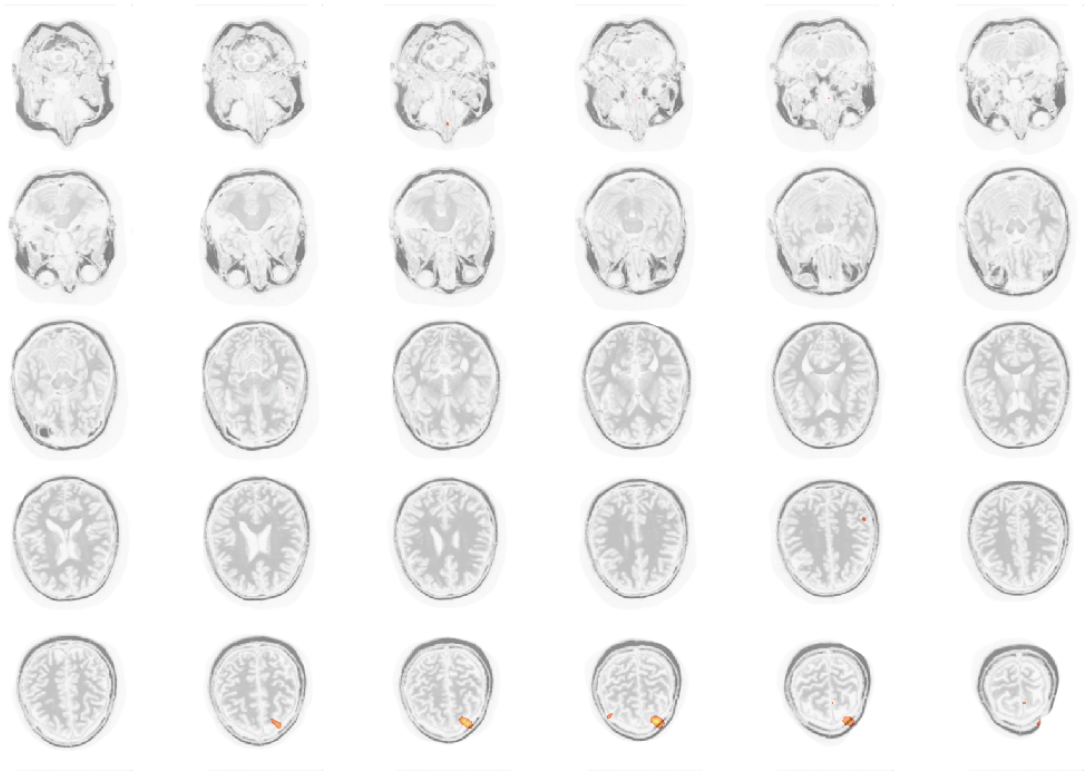


(b) Component extracted using PCA-ICA

Figure 5.7: *Example of component extracted with PP-ICA (a) and PCA-ICA (b). The PP-ICA component has low temporal extent and high value. The PCA-ICA component consists of different temporal connected regions.*



(a)



(b)

Figure 5.8: *Two example components extract with Lanczos-ICA. Component (a) has not converged while (b) has converged and appear as well spatially isolated.*

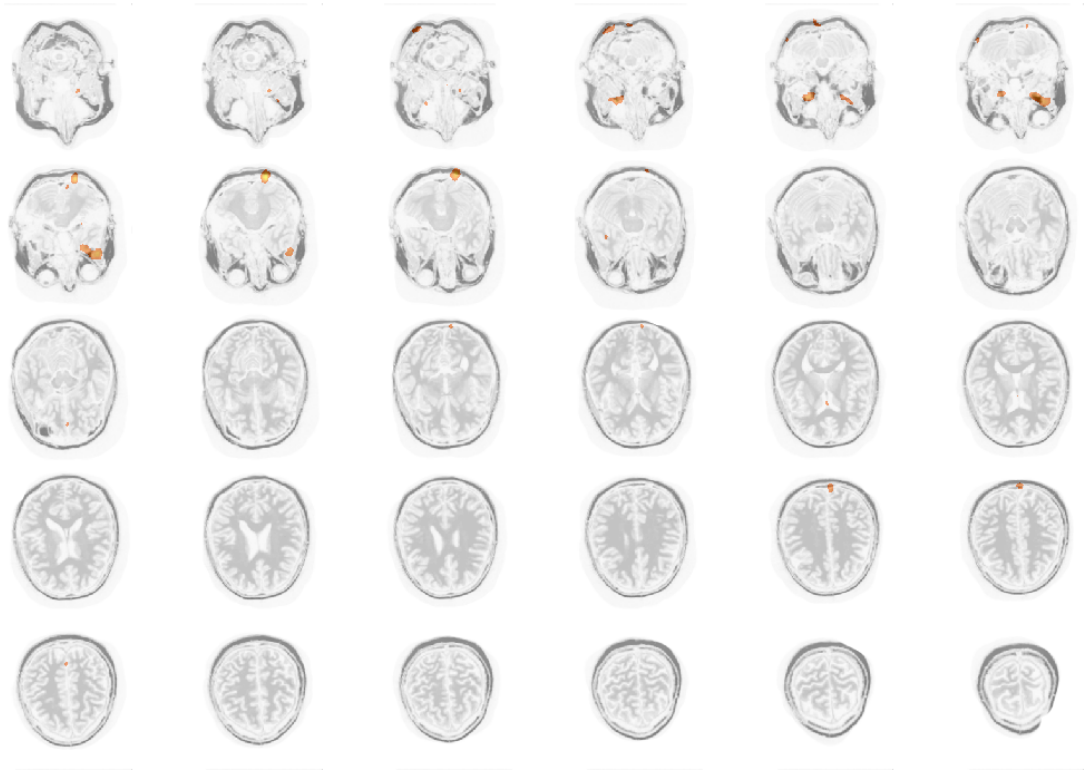
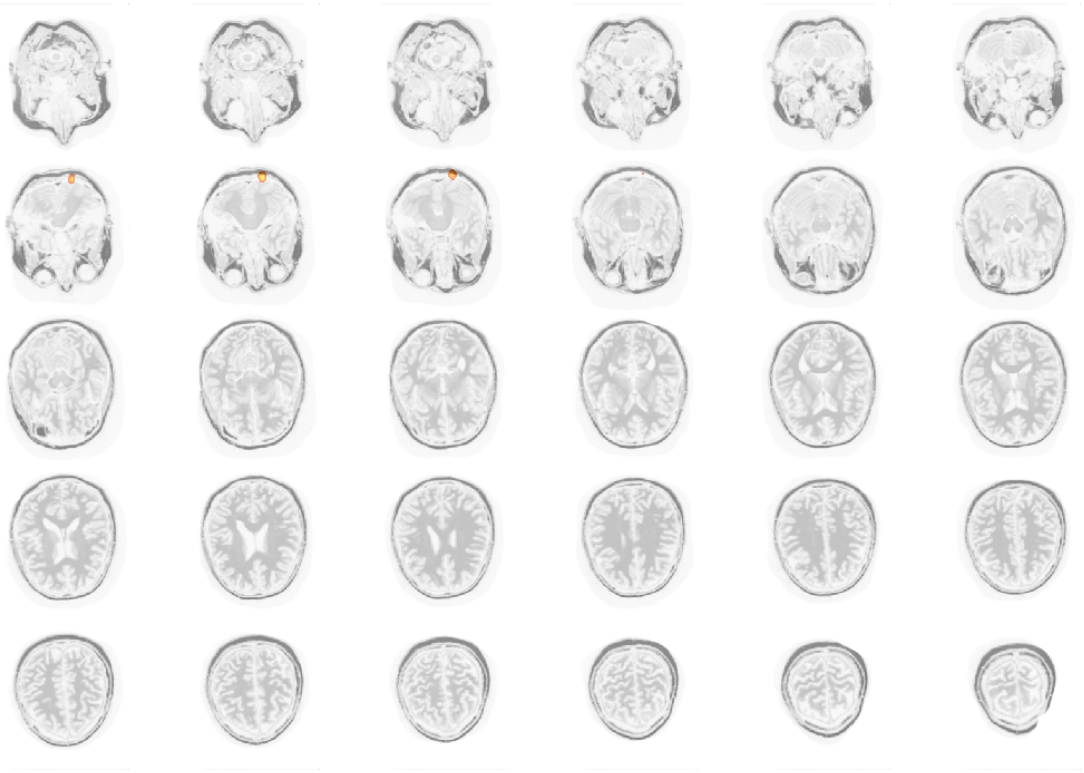
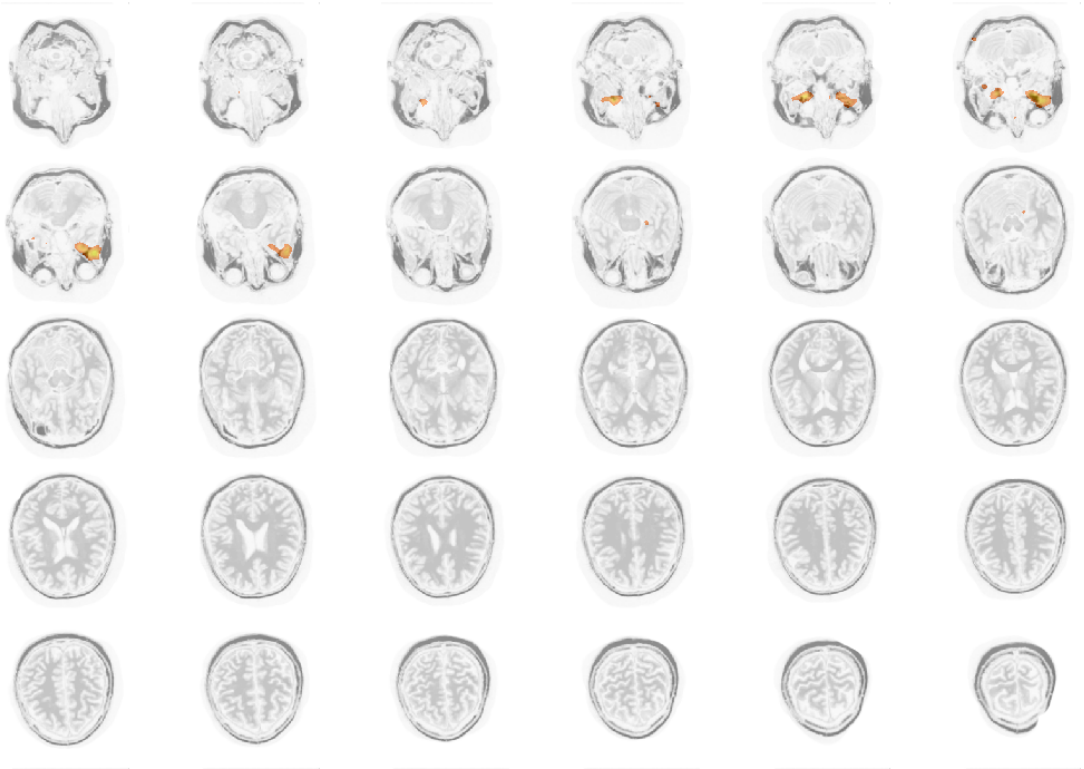


Figure 5.9: *Double component extracted by PCA-ICA. This component should be compared to the components in Figure 5.10(a) and Figure 5.10(b).*

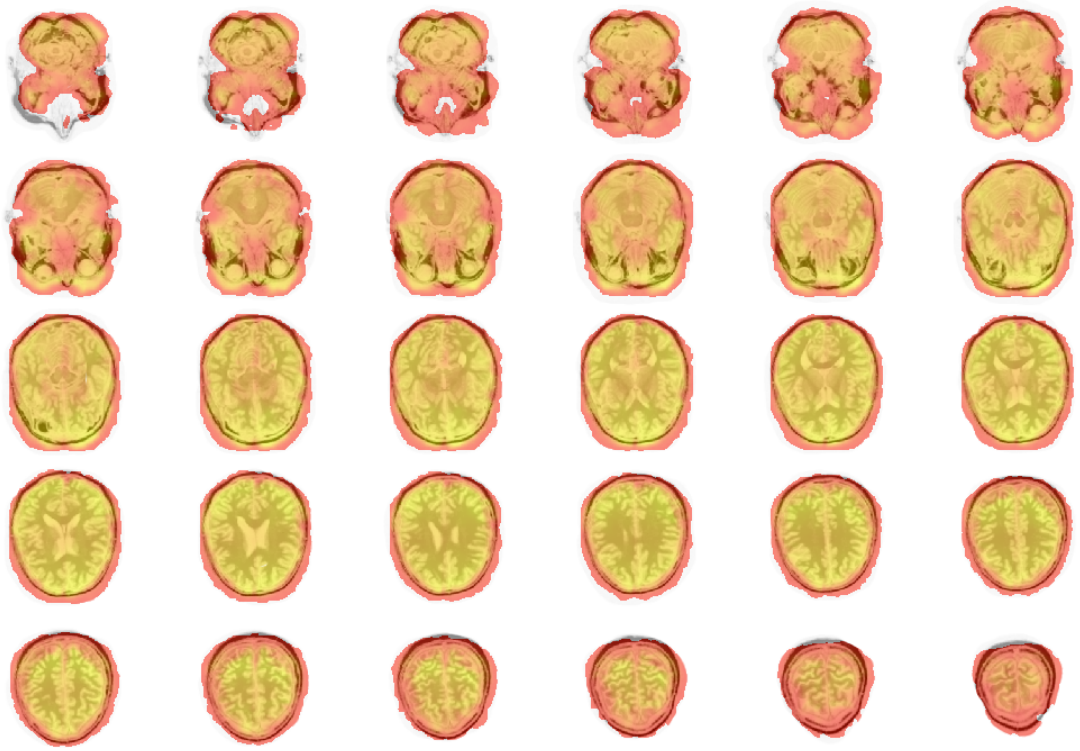


(a)

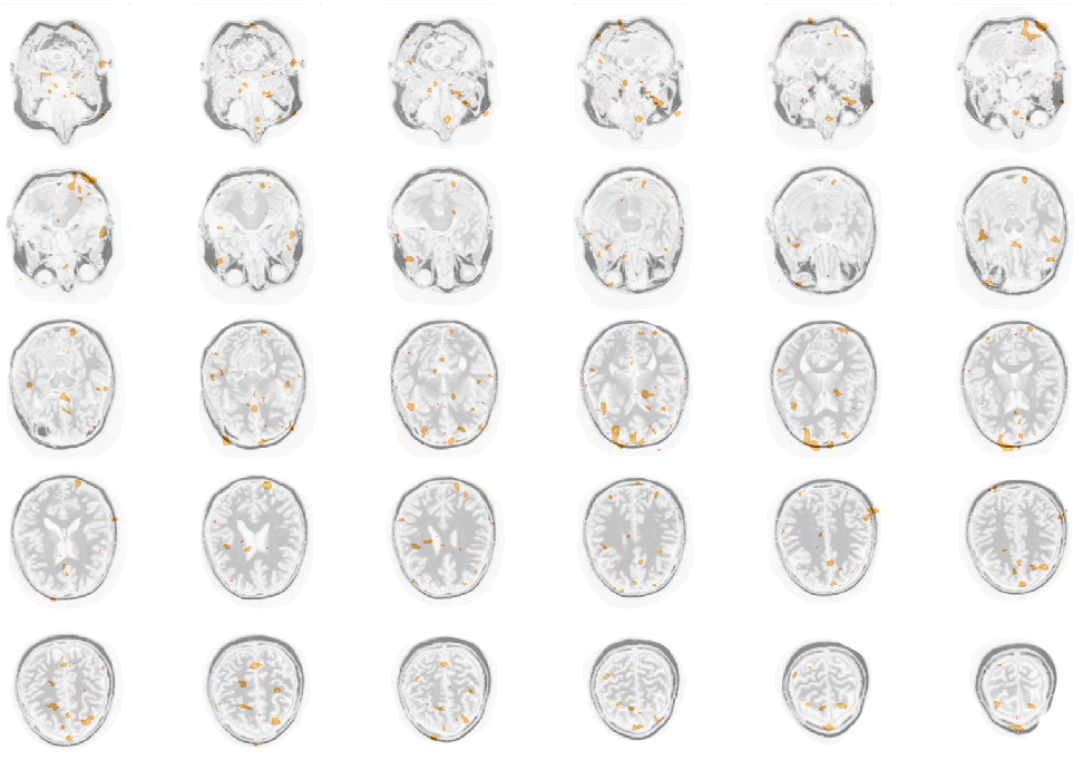


(b)

Figure 5.10: *PP-ICA components. These components will together form the component in Figure 5.9. Hence PP-ICA may express a single component obtained with PCA-ICA as two independent components.*



(a)



(b)

Figure 5.11: *For comparison; visualization of a time sample before analysis (a) and a component from a pure PCA decomposition (b).*

Chapter 6

Conclusions and Further Work

The PCA-ICA method described in the previous chapters is the most common method for ICA analysis of fMRI data. We argue in this thesis that the PCA-ICA method may cause misleading results in some cases. The method is compared to a Projection Pursuit ICA method based on existing theory from FastICA and to a new Lanczos based ICA method.

The use of PCA as a prepossessing step for FastICA and other ICA-algorithms has been widely studied. Hyvärinen et al. [15] show examples where over-learning makes the ICA models unsuccessful without a PCA preprocessing stage. Nadal et al. [24] investigate from a general view if data may be lost under the PCA-process and concludes that in most applications PCA-ICA will perform well. However, both Hyvärinen and Nadal point out that information may be lost in the PCA step. This loss of information may cause false components and is confirmed in this thesis by several examples. We have argued that components with small standard deviation but high kurtosis value are running a particular risk of being left out. Hence PCA-ICA must be used with care and alternative methods dealing with large datasets are needed.

Based on the observations in Chapter 5 and the related theory, we can not advocate Lanczos-ICA as a full-worthy alternative to the commonly used PCA-ICA in fMRI analysis. The results in this thesis are based on a single subject in a resting state study. Further studies using more subjects and different datasets may verify the quality of the Lanczos-ICA method.

We have also seen that the eigenvalue spectre of the fourth cumulant tensor yield an undesirable convergence for some of the components. The spectre is by definition unknown prior to the analysis due to the assumed blind source separation model where no information about the sources is

given. However, with better knowledge to the spectrum, convergence properties can be changed by different pre-conditioners. This is a separate research field within numerical linear algebra where a range of pre-conditioners are known. In practice pre-conditioners can be recognized as e.g. low-pass or high-pass filters. The attempt to make ICA a semi-blind method by introducing prior knowledge is not new. It is also discussed in the framework of the PCA-ICA model [3]. On the other hand, too much guidance by prior information remove us from the advantages of a blind model, namely that we may encounter unexpected but interesting results. By fitting the data to a model constructed only to give certain output, we may only confirm results we already have some knowledge about.

Furthermore, the Lanczos-ICA method is opening several new possibilities by giving the fMRI analysis a different mathematical perspective. Using more advanced mathematical theory within linear algebra and eigenvalue extraction, the next generation Lanczos-ICA methods may offer an interesting alternative to PCA-ICA. There has been done work in this direction by e.g. Lim and Morton [23]. They introduce a more general solution strategy to the blind source problem with no assumption on independence called Principal Cumulant Component Analysis (PCCA). Both the PCCA method and the method and examples given in this thesis point out promising alternative views on the solution of the blind source problem in fMRI analysis.

Appendix A

Visualization Methods

Visualization of 3D brain data is not trivial. Effective visualization tools are implemented in commonly used toolboxes such as GIFT [4] and FSL [27, 31]. An example of a fMRI component visualization in GIFT is given in Figure A.1.

In general, visualizations of fMRI components are done by thresholding the spatial signal data and superimposing it to a structural map. In this way the components are possible to localize spatially in the brain. The thresholding is often done manually as different components have different signal to noise ratio. It is also common to show the time-course of the component. The time-course will simplify the identification of the different spacial components.

For the purpose of this thesis, a simple fMRI visualization method for Matlab[1] is developed. All 3D volumes in this thesis are represented as a mosaic of 2D slices using this method. An examle of a slice obtained from a volume is shown in Figure A.3. In each slice, fMRI data are superimposed on a structural image from the same scan. Since the fMRI and the structural data are from the same subject, no registration is done other then a interpolation of the fMRI data to fit the smaller voxel size of the structural data. The interpolation is cubic and makes the final image appear smoother. No time-course is displayed as the neurobiological identification of the components is outside the scope of this thesis. An example of a visualized component is given in Figure A.2. This component can be compared to the one in Figure A.1. Notice that in Figure A.2 the head is oriented with nose downwards and with the lower slice in the upper left corner.

[1]The MathWorks, Matlab®, http://www.mathworks.com/products/matlab/

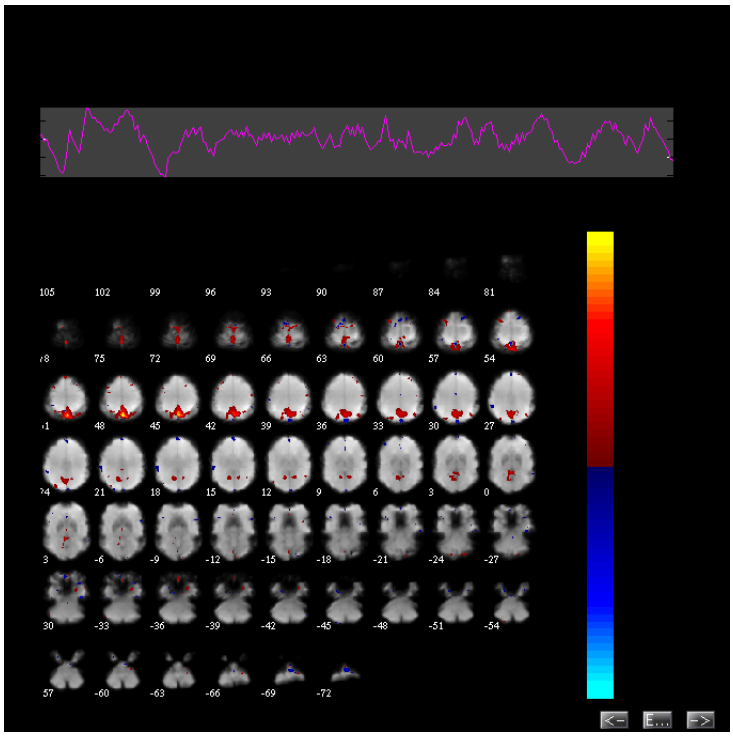


Figure A.1: *Example component decomposed and displayed with GIFT [4].*

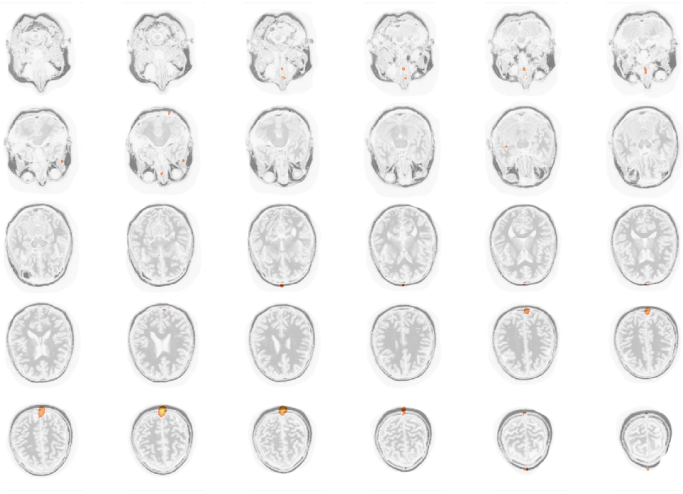


Figure A.2: *Example component decomposed with Lanczos-ICA and displayed with the standard method for this thesis.*

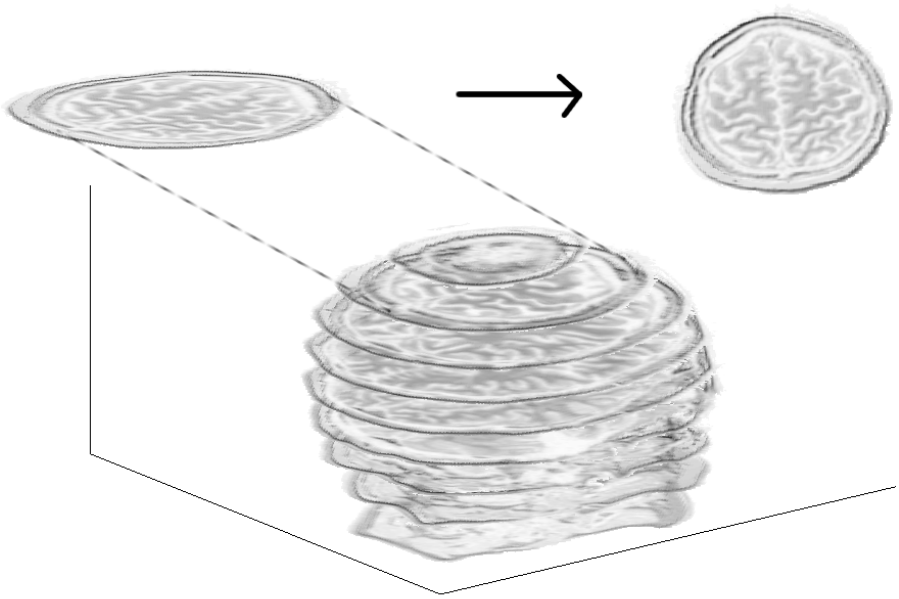


Figure A.3: *Slice obtained from a 3D volume of structural data. All 3D volumes in this theis are displayed as sets of slices.*

Appendix B

Properties of the Cumulant Operator F

B.1 Simplification of Cumulant Operator

A simplification of the cumulant operator (3.1) from Chapter 3 is given below. The simplification is valid for the normalized variable $z \in \mathbb{R}^n$ with $\sigma_{z_i} = 1 \quad \forall\, i$

$$\begin{aligned}
F_{ij}(Y,z) &= \sum_{kl} y_{kl} cum(z_i, z_j, z_k, z_l) \\
&= \sum_{kl} y_{kl} \Big(E\{z_i z_j z_k z_l\} - E\{z_i z_j\}E\{z_k z_l\} - E\{z_i z_k\}E\{z_j z_l\} - E\{z_i z_l\}E\{z_j z_k\} \Big) \\
&= \sum_{kl} y_{kl} E\{z_i z_j z_k z_l\} - \sum_{kl} y_{kl} E\{z_i z_j\}E\{z_k z_l\} \\
&\quad - \sum_{kl} y_{kl} E\{z_i z_k\}E\{z_j z_l\} - \sum_{kl} y_{kl} E\{z_i z_l\}E\{z_j z_k\}
\end{aligned}$$

We further use the normality of z giving $E\{zz^T\} = I \Longrightarrow E\{z_i z_j\} = \delta_{ij}$;

$$\begin{aligned}
F_{ij}(Y,z) &= \sum_{kl} y_{kl} E\{z_i z_j z_k z_l\} - \sum_{kl} y_{kl}\delta_{ij}\delta_{kl} - \sum_{kl} y_{kl}\delta_{ik}\delta_{jl} - \sum_{kl} y_{kl}\delta_{il}\delta_{jk} \\
&= E\{(\sum_{kl} z_k y_{kl} z_l) z_i z_j\} - y_{i,j} - y_{i,j} - \delta_{ij}\sum_{kl} y_{kl}\delta_{kl} \\
&= E\{(z^T Y z) z_i z_j\} - 2y_{i,j} - \delta_{ij}\mathrm{trace}(Y).
\end{aligned}$$

Hence the operator on full form is given by:

$$F(Y,z) \;=\; E\{(z^TYz)zz^T\} - 2Y - \text{trace}(Y)I.$$

B.2 Rank-one Eigenmatrix

The rank-one matrices $w_m \cdot w_m^T$ is assumed to be eigenmatrices for the cumulant operator F in Expression (3.1) when w_m is a column in the whitened mixing matrix (2.2). A proof is given in [15] and outlined here. Recall that the whitened ICA model is given by $z = W^T s \Longrightarrow z_i = w_i^T s$.

$$\begin{aligned} F_{ij}(w_m \cdot w_m^T, z) &= \sum_{kl} w_{mk}w_{ml}\text{cum}(z_i, z_j, z_k, z_l) \\ &= \sum_{kl} w_{mk}w_{ml}\text{cum}(\sum_q w_{qi}s_q, \sum_{q'} w_{q'j}s_{q'}, \sum_r w_{rk}s_r, \sum_{r'} w_{r'l}s_{r'}) \\ &= \sum_{klqq'rr'} w_{mk}w_{ml}w_{qi}w_{q'j}w_{rk}w_{r'l}\text{cum}(s_q, s_{q'}, s_r, s_{r'}) \end{aligned}$$

We also recall that the forth order cumulant is diagonal for independent data, giving $\text{cum}(s_q, s_{q'}, s_r, s_{r'}) \neq 0$ for $q = q' = r = r'$. And $\text{cum}(a,a,a,a) = \text{kurtosis}(a)$. Hence several summation indices cancel:

$$\begin{aligned} F_{ij}(w_m \cdot w_m^T, z) &= \sum_{klq} w_{mk}w_{ml}w_{qi}w_{qj}w_{qk}w_{ql}\text{cum}(s_q, s_q, s_q, s_q) \\ &= \sum_{klq} w_{mk}w_{ml}w_{qi}w_{qj}w_{qk}w_{ql}\text{kurtosis}(s_q) \end{aligned}$$

The vectors w_i are orthogonal leading to $\sum_n w_{mn}w_{qn} = \delta_{mq}$ which will simplify the summation even more:

$$\begin{aligned} F_{ij}(w_m \cdot w_m^T, z) &= \sum_q \delta_{mq}\delta_{mq}w_{qi}w_{qj}\text{kurtosis}(s_q) \\ &= w_{mi}w_{mj}\text{kurtosis}(s_q). \end{aligned}$$

Which gives the matrix expression

$$F(w_m \cdot w_m^T, z) = \text{kurtosis}(s)w_m \cdot w_m^T,$$

proving the eigenmatrix property, but also pointing out the relation between the eigenvalues and the kurtosis values of the sources.

Bibliography

[1] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.

[2] D. R. Brillinger. *Time Series Data Analysis and Theory*. International Series in Decision Processes. Holt, Rinehart and Wilston, 1975.

[3] V. D. Calhoun and T. Adali. Unmixing fmri with independent component analysis. *Engineering in Medicine and Biology Magazine, IEEE*, 25(2):79–90, March-April 2006.

[4] V. D. Calhoun et al. Group ICA of fMRI toolbox (GIFT). *Online at http://icatb. sourceforge. net*, 2004.

[5] J. F. Cardoso. Eigen-structure of the fourth-order cumulant tensor with application to the blind source separation problem. In *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, pages 2655–2658 vol.5, Apr 1990.

[6] J. F. Cardoso. Infomax and maximum likelihood for blind source separation. *IEEE Signal processing letters*, 4(4):112–114, 1997.

[7] J. F. Cardoso and P. Comon. Independent component analysis, a survey of some algebraic methods. In *IEEE International Symposium on Circuits and Systems*. IEEE, 1996.

[8] J .F. Cardoso and A. Souloumiac. Blind beamforming for non Gaussian signals. *IEE proceedings-f*, 1993.

[9] J. F. Cardoso and A. Souloumiac. Jacobi angles for simultaneous diagonalization. *SIAM J. Mat. Anal. Appl.*, 17(1):161–164, January 1996.

[10] C. Chen, W. Hardle, and A. Unwin. *Handbook of data visualization*. Springer Verlag, 2008.

[11] W. Cheney. *Analysis for Applied Mathematics*. Springer, 2001.

[12] I. Daubechies, E. Roussos, S. Takerkart, M. Benharrosh, C. Golden, K. D'Ardenne, W. Richter, J. D. Cohen, and J. Haxby. Independent component analysis for brain fMRI does not select for independence. *Proceedings of the National Academy of Sciences*, 106(26):10415–10422, 2009.

[13] M. E. Fox, M. D.and Raichle. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nat Rev Neurosci*, 8(9):700–711, Sep 2007.

[14] G. H. Golub and C. F. Van Loan. *Matrix computations*. Johns Hopkins Univ Pr, 1996.

[15] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley, 2001.

[16] A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural computation*, 9(7):1483–1492, 1997.

[17] A. Hyvärinen, J. Sarela, and R. Vigário. Spikes and bumps: Artefacts generated by independent component analysis with insufficient sample size. In *First International Workshop on Independent Component Analysis and Signal Separation*, 1999.

[18] C. Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *J. Res. Nat. Bur. Standards*, 45(4):255–282, 1950.

[19] David C. Lay. *Linear algebra and it's applications*. Pearson/Addison-Wesley, Boston, MA, USA, third edition, 2006.

[20] Y. O. Li, T. Adali, and V. D. Calhoun. Sample dependence correction for order selection in fMRI analysis. In *3rd IEEE International Symposium on Biomedical Imaging: Nano to Macro, 2006*, pages 1072–1075, 2006.

[21] A. Lundervold. On consciousness, resting state fmri, and neurodynamics. *Nonlinear Biomedical Physics*, 2010. 4(Suppl 1):S9.

[22] D. J. C. MacKay. Maximum likelihood and covariant algorithms for independent component analysis. In *Available electronically at http://www. inference. phy. cam. ac. uk/mackay/abstracts/ica. html*. Citeseer, 1996.

[23] J. Morton and L.-H. Lim. Principal cumulant component analysis,. Technical report, Stanford University, 2009.

[24] J. P. Nadal, E. Korutcheva, and F. Aires. Blind source separation in the presence of weak sources. *Neural Networks*, 13(6):589–596, 2000.

[25] B. N. Parlett and J. K. Reid. Tracking the progress of the Lanczos algorithm for large symmetric eigenproblems. *IMA Journal of Numerical Analysis*, 1(2):135, 1981.

[26] J. J. Pekar. A brief introduction to functional MRI. *IEEE engineering in medicine and biology magazine: the quarterly magazine of the Engineering in Medicine & Biology Society*, 25(2):24, 2006.

[27] S. M. Smith, M. Jenkinson, M. W. Woolrich, C. F. Beckmann, T. E. J. Behrens, H. Johansen-Berg, P. R. Bannister, M. De Luca, I. Drobnjak, D. E. Flitney, et al. Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage*, 23:S208–S219, 2004.

[28] J. V. Stone. Independent component analysis: an introduction. *Trends in Cognitive Sciences*, 6(2):59–64, 2002.

[29] L. N. Trefethen and D. Bau. *Numerical linear algebra*. Society for Industrial Mathematics, 1997.

[30] R. E. Walpole, R. H. Myers, S. L. Myres, and K. Ye. *Probability & Statistics for Engeneers & Scientists*. Prentice Hall, seventh edition, 2002.

[31] M. W. Woolrich, S. Jbabdi, B. Patenaude, M. Chappell, S. Makni, T. Behrens, C. Beckmann, M. Jenkinson, and S. M. Smith. Bayesian analysis of neuroimaging data in FSL. *Neuroimage*, 45(1):S173–S186, 2009.

[32] M. Ystad, T. Eichele, A. J. Lundervold, and A. Lundervold. Subcortical functional connectivity and verbal episodic memory in healthy elderly - a resting state fmri study. *NeuroImage*, In Press:–, 2010.