BACHELOR THESIS

Classification inference for event-related fMRI data

Author:
J.M. Wolterink *s0616451*

Supervisors:
Dr. J. Farquhar
Prof. Dr. P. Desain

Radboud University Nijmegen

March 16, 2010

Abstract

This study presents methods, results and conclusions used and drawn from an attempt to cluster fMRI brain activation patterns in an unsupervised manner in order to retrieve a categorization in a set of nouns denoting objects. Over the last few years, a lot of work has been done on the classification of neural responses patterns corresponding to the presentation of different objects or concepts, sometimes called brain reading. Here, we try to infer a categorization from such a set of neural response patterns based only on the similarities between these patterns. By assigning each neural response pattern for the presentation of an object to a cluster in a multi-dimensional space, a categorization in the set patterns can be induced. Furthermore, clustering methods are used to address the question which number of clusters is most plausibly present in the pattern set. As it turns out, an experiment design with long inter trial intervals and multiple stimulus presentations is critical for such an approach to be successful.

Contents

Chapter 1

Introduction

Over the past few decades, machine learning methods have been used in hundreds of ways to solve complex or time-consuming classification problems. In a wide range of research topics, machine learning has shown to be a quick and efficient way of retrieving information and assigning patterns to their appropriate class. When used together with human physiological data, machine learning methods could be used to either recognize individuals or evaluate the state in which an individual is. For the first problem, applications like handwriting recognition, speaker recognition and writer identification are already in use in everyday life. The second problem, identifying the state of an individual, implies methods like facial expression recognition and emotion recognition in speech. Logically, a next step would be to be able to infer what cognitive state a person is in, or even what a person is thinking about. To do this appropriately, one should know how concepts are organized in the brain. Obtaining a hierarchy of how the brain organizes objects could really increase the understanding and recognition of human thought. In this thesis, one possible approach to this problem is explored.

Quite recently, human state recognition has been taken one step further to the level of classifying neural activity patterns. This means that classification of human thought has already come within reach. Machine learning methods have been applied to the problem of classifying human brain activity associated with different nouns (2). Models have been trained that are able to retrieve nouns belonging to specific neural activity patterns. Others have applied machine learning methods for detection and classification of fMRI patterns in the visual cortex (3). In (3), participants were looking at a picture drawn from one of ten classes, after which the classifier could correctly identify from which of those classes the image was drawn, based on the neural activity corresponding with the trial. Measuring the neural response to objects further along the object visual pathway has also shown to pay off (4). It was found that widely distributed and overlapping patterns of neural responses in the ventral temporal cortex could be used to correctly classify image viewing trials.

This kind of research has been rapidly gaining ground, with techniques now even being able to classify neural patterns excluding visual areas (5). Furthermore, the model used by Shinkareva et al. was able to classify trials using a classifier trained only on neural patterns of other participants. This shows that neural patterns are present, can be identified using machine learning algorithms and are robust enough to be generalized among different human subjects. This has implications for the use of this kind of brain reading in real life situations. Classifying neural patterns elicited by the thought of a certain noun category could be used as a paradigm in brain computer interfaces. Or even better, applied as a fast way of communication by concept

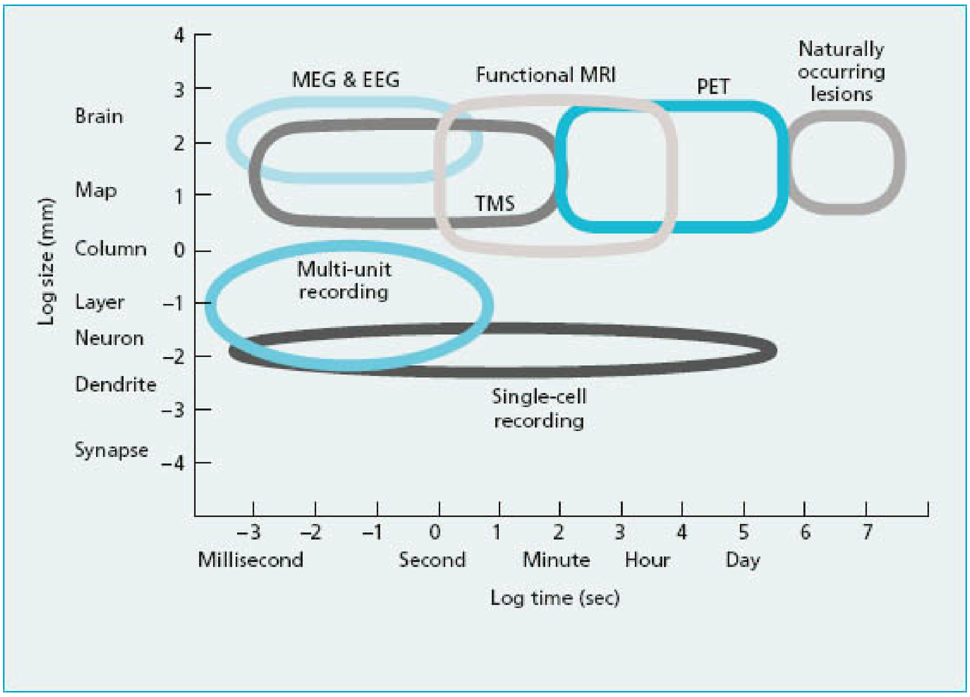


Figure 1.1: Spatial and temporal resolution of several brain function assessing methods. Adapted from (1).

with locked-in patients. When applied with the appropriate concern for ethical issues, these methods can be used in crime suspect interrogation (6). Methods could be used to gain insight in the thoughts of seriously autistic children or give us new insights in diseased like Alzheimer. On a less serious note, brain reading could be used in games.

Where these studies offer a way to distinguish between predetermined classes of objects (e.g. tools and dwellings (5)), in this thesis we try to infer a hierarchy in a set of objects based only on the characteristics of the neural activation patterns that presentation of these objects elicits. The difference is that while previous work has been done in a top-down fashion, here we will attempt to use a bottom-up fashion, where the only source of information is the data. This yields possibilities for finding new structures in the way the brain organizes the object set.

Functional magnetic resonance imaging (fMRI) has shown to be the appropriate tool for research of this kind. As a noninvasive imaging method, fMRI has some advantages over other imaging and electrophysiological methods for assessing brain function. In comparison with its most direct competitor positron emission tomography (PET), fMRI does not require radioactive injections and has a relatively high temporal resolution. Due to its high spatial resolution in comparison with electrophysiological methods for assessing brain function such as electroencephalography (EEG) and magnetoencephalography (MEG), fMRI offers a detailed view in the workings of the different parts in the human brain. Though their temporal resolution is much higher than that of fMRI, electrophysiological methods suffer from a trade off between invasiveness and localization accuracy. It is mathematically impossible to uniquely identify the locations of the neural sources that cause a given pattern of activity on the skull. This problem, also called the inverse problem, limits the value of using EEG and MEG in creating maps of brain function (7). More invasive electrophysiological methods such as ECoG, which uses electrodes planted on the cortex, have a higher spatial resolution, but their use involves medical risks for the subject and a very restricted pool of subjects for empirical studies. As can be seen in Figure 1.1, fMRI offers a fairly high spatial resolution, with the big drawback being its somewhat low temporal resolution. Though this lack of temporal resolution puts constraints on experiment design, fMRI has emerged as the most widely used tool in neuroimaging studies.

In this thesis, fMRI data is used. Neural activation patterns are extracted from fMRI data used in a study by Rueschemeyer et al. (8). In that study, participants were shown words corresponding to objects in one of two classes. Objects were either functionally manipulable (FM) or volumetrically manipulable (VM). The FM class contains objects such as a cup and a pen, that need manipulation to be of use. The VM class contains objects such as a bookend or a clock, which can be manipulated, but also function when they are not. Standard event related fMRI analysis has already been performed on this data, yielding significant differences in particular brain areas. We try to retrieve these classes using machine learning techniques on the individually acquired trial specific neural activation patterns.

In the experiment by Rueschemeyer et al., stimuli were presented with an 8 second inter trial interval (ITI). This relatively short interval, combined with the low temporal resolution of fMRI and the slow nature of the blood-oxygen-level dependent (BOLD) signal that fMRI measures, puts heavy constraints on the possible analysis methods. In previous brain reading studies, stimuli were either presented multiple times or with a large interval to get a more reliable estimate of the neural activation corresponding to a stimulus presentation. For example, (9) propose an ITI of around 20s to obtain a good estimate of activation for a single trial. The problem with ITI's of this size is that an experiment will take a very long time and participants might easily lose their concentration after a while.The question is whether neural activation patterns for individual trials can be extracted from fast event-related fMRI experiments with an ITI in the range of 8 seconds. Such an extraction method would allow for online classification of brain activation in a reasonably fast setup. One of the challenges in fMRI research is to find methods that are capable of dealing with such ITIs.

One exploratory method of machine learning is unsupervised clustering, algorithms that try to partition a set of patterns so that all patterns in a subset show a high similarity while the similarity between subsets of patterns is kept low. These methods perform their classification based only on the features in the patterns and the relationship between different patterns. An optimal clustering of patterns contains compact and well-separated clusters. Over the years, different clustering methods have been proposed, the most common being agglomerative algorithms which combine clusters in a bottom-up fashion with each pattern starting out as a cluster, and error-minimizing methods such as the *k*-means algorithm. The latter one tries to find cluster centroids such that the total distance of patterns from cluster centers is minimized.

In this thesis, we try to retrieve a categorization in nouns from neural activation patterns derived from a fast event-related fMRI experiment. For this, an fMRI data set already known to contain two classes is used. In the set of nouns, we expect to find groups of which the members have similar neural activation patterns. If any, we expect to find the distinction between FM and VM words. The groups or clusters of neural activation patterns are found using various unsupervised clustering methods and different clustering validity assessment methods are employed to assess the found clusterings. More precisely, we will derive trial-specific patterns from a data set by (8) and cluster these patterns in order to retrieve the FM and VM classes. Along with this, attention will be given to possible other existent classes in the data set.

Chapter 2

Theory

2.1 BOLD functional magnetic resonance imaging

Since the late 1960s, research has been conducted on measuring magnetic resonance (MR) in biological tissue, early applications including detection of cancerous cells in rats. Through the past few decades, MR has evolved to fMRI ((7) for an extensive overview). Magnetic resonance imaging (MRI) makes use of the magnetic properties of protons. A strong magnetic field is induced with an excitation pulse, forcing the protons to magnetically align with the field and in this process absorb energy. When the magnetic field is removed, the protons return to their original state, emitting a certain signal to a receiver coil. The protons return to their equilibrium state at a different rate, depending on the kind of tissue in which they are contained. This creates a contrast between different kinds of tissue.

	Gray Matter	White Matter	Cerebrospinal Fluid
T_1	900 ms	600 ms	4000 ms
T_2	100 ms	80 ms	2000 ms

Table 2.1: Values for time constants T_1 and T_2 at field strength of 1.5 T. From (7).

The detected signal M_{xy} at time t depends on two tissue-specific time constants: T1 is the time it takes a proton to return to the original energy level, T2 is the time it takes to retrieve its original magnetic orientation. Table 2.1 shows the differences in these values for different kinds of brain tissue. Equation 2.1 shows how M_{xy} depends not only on the original signal M_0, but also on the values of T_1 and T_2. Furthermore, M_{xy} depends on the time between an excitation pulse and data acquisition (TE) and the time between different excitation pulses (TR).

$$M_{xy}(t) = M_0(1 - e^{-TR/T_1})e^{-TE/T_2} \tag{2.1}$$

A contrast can be created by setting TE and TR such that M_{xy} is high for one kind of tissue, while it is low for another kind of tissue. This gives equation 2.2 for the contrast between tissues A and B. Different kinds of TE/TR combinations elicit different contrasts, the most common being T_1-weighted contrast which shows differences in T1, and T_2 or T_2^*-weighted contrast which uses differences in T2.

$$C_{AB} = M_{0A}(1 - e^{-TR/T_{1A}})e^{-TE/T_{2A}} - M_{0B}(1 - e^{-TR/T_{1B}})e^{-TE/T_{2B}} \tag{2.2}$$

In blood oxygenation level dependent (BOLD) fMRI, the most commonly used fMRI technique, the magnetic properties of hemoglobin, the oxygen transporter in blood, are used. Active neurons need more oxygenated blood and it is believed that there is a correlation between neural activation and the amount of oxygenated blood in a brain area. Oxygenated hemoglobin (Hb) has no magnetic moment, but deoxygenated hemoglobin (dHb) has a significant magnetic moment. In very strong magnetic fields, this difference is magnified. During the late 1980s, it was found that a contrast could be made based on the amount of Hb. From this, images could be constructed showing where in their brain there was more Hb and thus more neural activity. Over time, the signal on T_2^* images shows a particular shape for activated brain areas, called the hemodynamic response function. Though the actual shape may vary across individuals, brain areas or even sessions (9), a canonical function can be drawn as in Figure 2.1.

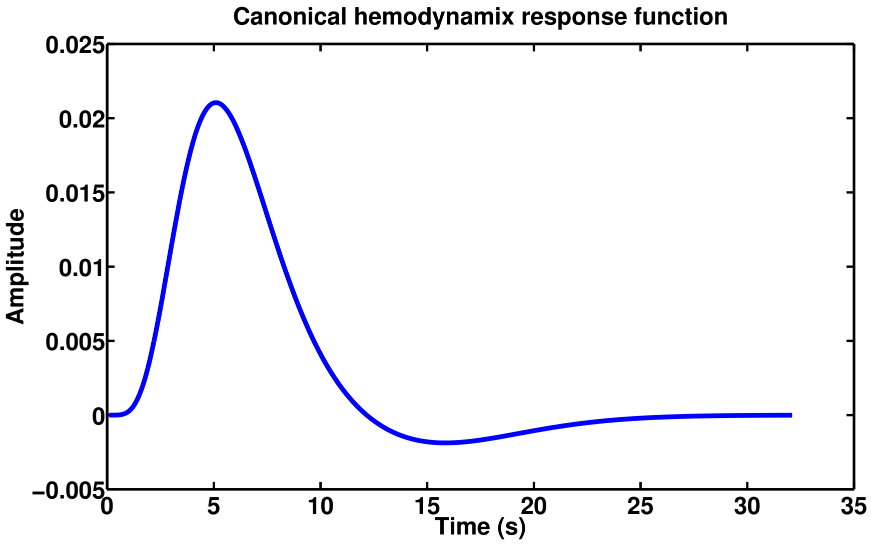


Figure 2.1: Canonical hemodynamic response function, derived from the Matlab SPM toolbox.

fMRI experiments are usually aimed at localizing brain areas that are involved in a certain cognitive task. In a review of 275 fMRI and PET studies, Cabeza and Nyberg present an overview of the broad range of subjects for which fMRI has been applied (10). Applications of fMRI range from attention, perception, language and all kinds of memory studies. Most of these studies use either a blocked design or an event related experiment design. In a blocked design, two or more different conditions are presented in an alternating pattern, e.g. ABABAB for conditions A and B. In an event-related design, stimuli are presented as individual trials, their order being independent of the condition they belong to. In slow event-related designs, the time between trials allows for the BOLD signal to return to its original state. In fast event-related designs, methods such as linear regression, deconvolution and finite impulse response (FIR) deconvolution have to be used to retrieve activations.

2.2 Unsupervised Clustering

We attempt to apply unsupervised clustering methods to patterns of neural responses to stimuli. These methods produce a clustering of the nouns, from which we try do infer a categorization in the brain in a bottom-up fashion. According to (11), clustering is the unsupervised classification of patterns (observations, data items or feature vectors) into groups (clusters). It can be noted that the clustering problem has been adressed in a broad variety of disciplines because of its usefulness as an exploratory step in data analysis. In this thesis, we employ several clustering algorithms. Figure 2.2 offers a coarse taxonomy for different kinds of clustering algorithms. The most important division here is between algorithms that work bottom up, sequentially grouping smaller clusters into bigger clusters, starting with the individual patterns and partitional algorithms, that divide the whole set of patterns to optimize some objective function.

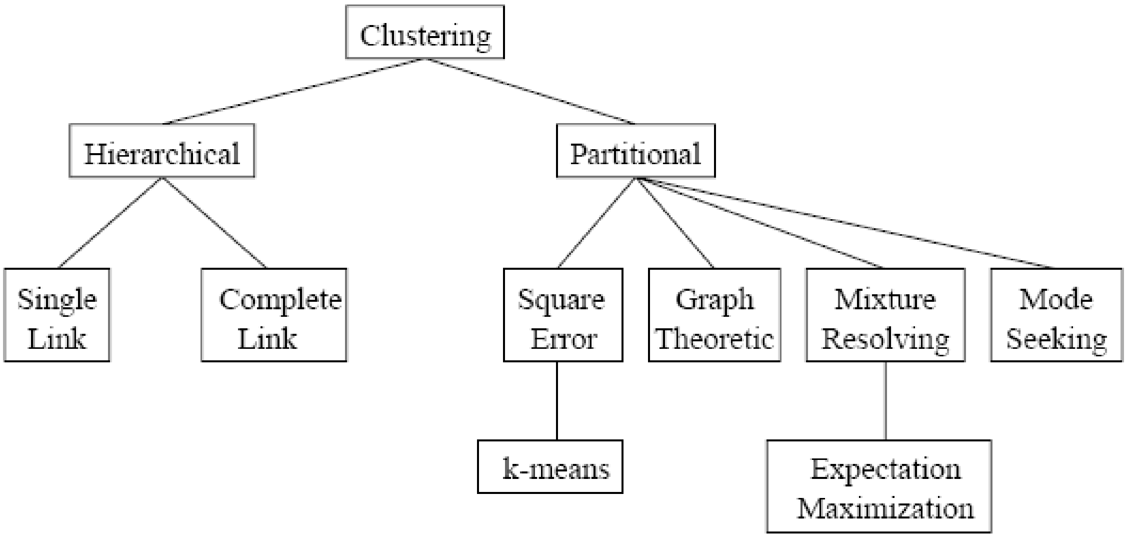


Figure 2.2: A taxonomy on unsupervised clustering. From (11).

Apart from this, other dimensions can be thought of on which to locate clustering algorithms. Jain et al. propose some issues, the most important of which are:

- agglomorative vs. divisive. Methods are either agglomorative (starting with one cluster per pattern, then merging clusters) or divisive (starting with all patterns in one cluster, then splitting cluster).

- hard vs. fuzzy. Clusterings can be hard (each pattern can only belong to one cluster) or fuzzy (patterns can have degrees of membership for several clusters).

- deterministic vs. stochastic. Deterministic methods have only one possible outcome, while stochastic methods may have several possible outcomes, depending on the order in which the patterns are clustered.

We will apply both agglomorative and divisive methods to the pattern set obtained by the preprocessing steps. We will only use hard clustering methods, since we assume a strict categorization among words for this problem. Methods to be used can either be deterministic (e.g. single link clustering) or stochastic (e.g. error-minimizing methods such as *k*-means clustering). All used clustering methods will be introduced in depth in Chapter 3.

In their 2001 review on clustering validation techniques, (12) identify four stages in the unsupervised clustering process:

1. Feature selection
2. Clustering algorithm selection
3. Validation of results
4. Interpretation

For feature selection, this thesis relies on literature from the brain reading field. Regarding the second step, selecting a clustering algorithm, we will compare the results of different algorithms and try and find which methods are most suitable in this case. Validation of results can happen internally, based on the metrics of the found clusters. Clusters that are compact and well-separated are thought to be more valid then overlapping clusters. External validation is also possible, where validity equals the found clustering's similarity with an other partition, in this case the predefined categorization in the stimulus set. The last step, interpretation, depends solely on the view of the researcher.

In addition to this, (11) and (13) propose a step between step 1 and 2. Measuring clustering tendency can save a lot of work. If the pattern set seems to random, there is no need to cluster, since all found clusters will not be any better than those of a random pattern set. In our case however, we assume that there are clusters present among the stimulus set because this has already been shown by (8). Therefore, our hypothesis is that the patterns we find have a tendency to cluster.

2.3 Machine Learning in fMRI classification

Over the past decade, large advances in so-called 'brain reading' have been made. In their paper on brain reading, (3) pose the brain reading problem as a pattern recognition problem where 'given a pattern of brain activity across space at a given point in time as measured by fMRI, a pattern recognition approach seeks to infer what percept a subject was experiencing'. In short, a pattern of activity over a set of voxels is extracted for a visually presented object after which this pattern is fed to a classifier together with its label. The classifier learns a mapping between patterns and stimulus categories. In a test session then, the classifier is fed new patterns for which it should correctly predict the category of the corresponding stimulus. This means that the brain reading problem consists of two stages, pattern extraction and classifier learning, and an additional test phase.

While approaches differ in the methods used in pattern extraction as well as the used classifiers, all brain reading methods use an approach similar to the one posed in the previous paragraph. Results have been promising, with classifiers being able to make distinctions between faces and objects (4), animal species (3) or tools and dwellings (5). Furthermore, While these three examples are impressive as they are, new approaches taken by (2) and (14) yield even more subtle distinctions between thousands of nouns or visual stimuli. These approaches make use of intermediate semantic levels (Figure 2.3) or receptive field models, while earlier approaches used simpler linear discriminant classifiers, support vector machines or gaussian naïve Bayes classifiers (3), (5).

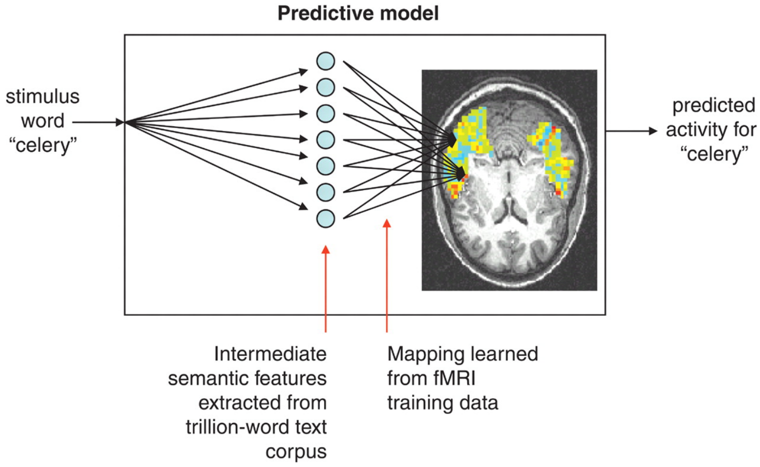

Predictive model
stimulus word "celery"
predicted activity for "celery"
Intermediate semantic features extracted from trillion-word text corpus
Mapping learned from fMRI training data

Figure 2.3: The brain reading model used by (2). The model uses an intermediate semantic features level to which each word is encoded. After that, the model predicts an fMRI image as a linear combination of fMRI signatures associated with the semantic features. From (2).

While brain reading approaches have been successful, the goal of this thesis is rather different from that of brain reading studies. Our problem has its first stage in common with that of brain reading, that is retrieving a neural activation pattern for a presented stimulus. But instead of then feeding this pattern to a classifier together with its label so that the classifier can learn a distinction between classes, an unsupervised clustering method does not leak any class label information. Instead, we try to infer a hierarchy based solely on the structure available in the pattern set. Therefore, the work in this thesis does not make use of methods posed for the second stage of brain reading, but instead only builds on pattern extraction methods used in brain reading studies. These should have advantages over standard fMRI techniques, since they try to infer stimulus-specific activation patterns, while event-related or blocked design fMRI techniques infer patterns per condition.

Brain reading techniques differ in the way that patterns are extracted. Pattern extraction methods heavily depend on the experiment design. Some studies deconvolve an average response over several repetitions of the same stimulus or category to allow for short inter trial intervals (e.g. (14),(3),(5)), while other studies use baseline conditions and long trials to allow for a trial-specific response to be found. In a tutorial on machine learning classifiers and fMRI, (15), the authors introduce several methods to obtain activation patterns from data. One method is to use the Percent Signal Change (PSC) relative to a baseline. An other method is to take the mean signal strength over all scans in a trial. A third method using linear regression is proposed in (15), where a pattern consists of β values for each voxel or signal. As we will see, for several reasons this is the best approach to our problem. In the same tutorial, it is stressed that the number of features should not exceed the number of patterns. Therefore, the number of features should be used, either through selection of particular Regions Of Interest (ROIs) or dimensionality reduction methods such as PCA or ICA. In the methods section, We will further elaborate on the lessons learnt from brain reading.

Chapter 3

Methods

The approach to fMRI analysis described in this thesis consists of two clearly distinguishable stages. The first stage amounts to pattern extraction and feature selection, while in the second stage the found patterns are used to retrieve a hierarchy among stimuli. In Figure 3.1, an overview of the approach as a whole is given. Along with this overview, this chapter describes which steps were taken.

As has been described in Chapter 2, for the first stage of our pipeline we have built on existing brain reading techniques, since our pattern extraction stage has a lot in common with the brain reading pattern extraction stage. In the present chapter, the experiment setup and resulting data set will be explained. Then, preprocessing steps and further processing steps will be described. In section 3.3, the general linear model and its workings will be explained. These steps all lead up to sets of patterns to be explored in the clustering steps. Section 3.4 will introduce each clustering method and the way it has been applied to our data. The last part of this chapter describes ways to measure validity of clusters.

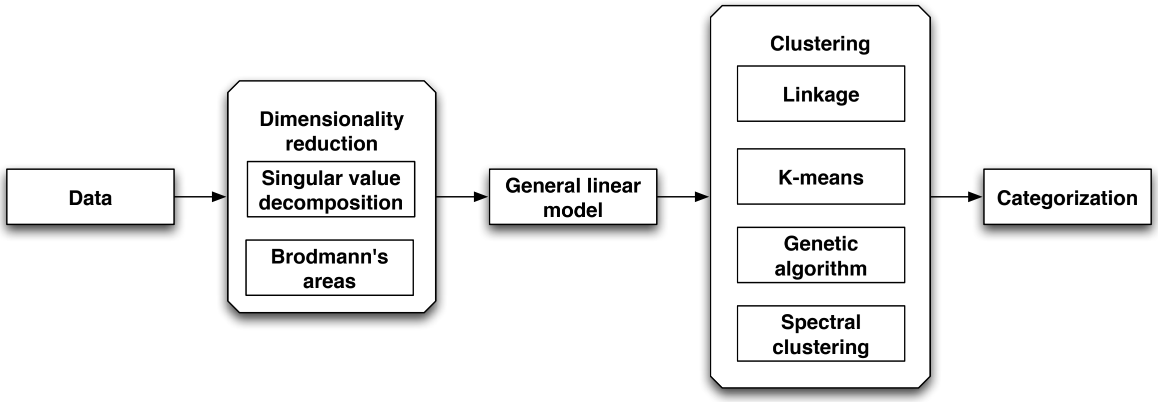


Figure 3.1: Overview of the pipeline approach used in this study.

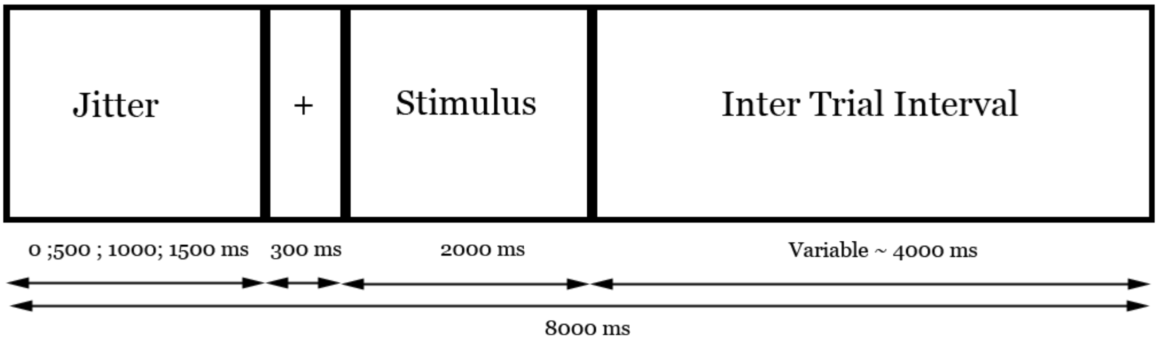


Figure 3.2: Standard temporal pattern for a trial. A trial consists of 4 stages: random jitter, fixation, stimulus presentation and variable ITI.

3.1 Data set

All analysis in this thesis is done on a data set by (8), that was gathered in an fMRI experiment which was part of a study on the distinction between objects with different manipulability. The objective was to see if there is a distinction between neural activation caused by viewing of either Functionally Manipulable (FM) objects, that need manipulation to be of use (e.g. a cup or scissors) and Volumetrically Manipulable (VM) objects, that do not necessarily need manipulation to function (e.g. a bookend or a clock). Here, I will briefly introduce the experiment and the characteristics of the data set.

3.1.1 Experiment setup

All FM and VM objects were represented by nouns. The list of VM and FM words was matched for relevant linguistic parameters, such as length, familiarity, imageability and frequency. There were 100 stimuli (40 FM, 40 VM, 20 nonwords) that were each shown once to a subject in an event-related design. Furthermore, there were 20 trials where there was no stimulus. This totals to 120 trials per subject. Subjects had the task to read each word thoroughly and respond to nonwords with a button press. Each of 120 trials was made up of four phases, which can be found in Figure 3.2. After an initial jitter of 0, 500, 1000 or 1500 ms and a fixation cross presentation of 300 ms, the target word was shown for 2000 ms or until a response was recorded. After that, the trial was filled up with an ITI so that every trial lasted exactly 8000 ms. After that, a new trial started with the same pattern.

3.1.2 Data acquisition

The following specifications are taken from (8). Functional images were acquired on a Siemens TRIO 2.0 T MRI System (Siemens, Erlangen, Germany) equipped with echo planar imaging (EPI) capabilities, using a birdcage head coil for radio frequency transmission and signal reception. The scanner acquired BOLD images (TE = 20ms, TR = 2000 ms) with a voxel size of 3.5 mm x 3.5 mm x 3.5 mm. Furthermore, it acquired anatomical images with a voxel size of 1 mm x 1 mm x 1 mm.

3.1.3 Data preprocessing

Data went through several preprocessing steps using the SPM5 (Statistical Parametric Mapping, `www.fil.ion.ucl.ac.uk/spm`) Matlab package. For every session, the first 3 volumes were removed to allow for T_1 equilibration effects. Body registration was applied along 3 translations and 3 rotations to correct for small head movements. This resulted in 6 motion registrations, which could be used in the GLM to account for subject movement. The time series for each voxel were realigned temporally to acquisition of the middle slice to correct for slice timing acquisition delays. Images were normalized to a standard brain and resampled at a voxel size of 6 x 6 x 6 mm as a first step in reducing the number of dimensions.

Because our fMRI signals were made up of ± 15 minutes of volumes each, it was inevitable that there was some low frequency scanner drift in the data. As is usual and advisable in the field of fMRI analysis (16), we applied a highpassfilter to the data that removed any noise with a frequency below a particular threshold. In event-related fMRI experiments, values between $\frac{1}{100}$ Hz and $\frac{1}{128}$ Hz for the cutoff point are typical. We chose the same value that (8) used, that is $\frac{1}{120}$ Hz. The cutoff point chosen makes sure that we do not lose the BOLD response, which has a period that is much shorter. The result is a signal more situated around its baseline, which makes regression results more reliable.

The resulting data set consisted of data for 14 subjects, each subject scanning session having around 480 volumes. After standardizing and resampling, every volume consisted of 27 x 32 x 25 voxels. Stimulus presentation onsets were registered so that they could be mapped to the acquired data. Though stimulus labels were also available, we did not use these until the cluster validation step in order not to leak any information to the clustering algorithms.

3.2 Dimensionality reduction

The second step in our approach is dimensionality reduction, in this case bringing down the number of possible features to be used in the clustering analysis. With only 80 critical patterns (FM and VM) per clustering analysis, it is important to reduce the number of features to a value well-below 80. High dimensionality is not so much a problem in standard fMRI experiments, where voxel activations are analyzed in a univariate manner and a large number of voxels would only increase computation time. But for brain reading high dimensionality is a problem, since an abundance of features make classification very easy for the training cases, but causes overfitting and thus very low generalizability for new cases. Though reducing dimensionality is necessary for a good clustering analysis, it has its drawbacks. We run the risk of losing perhaps very subtle information while going through the process of dimensionality reduction. We could easily throw the baby out with the bath water in averaging over signals or just removing signals, that is using methods too coarse for this type of signal. We should not forget that the signal to noise ratio in fMRI is very low, which means that the signals are very fragile. A good dimensionality reduction method should make sure not too lose the signal.

There are multiple ways of reducing dimensionality in fMRI data, both in standard fMRI analysis and in brain reading analysis. In cognitive fMRI studies, it is common to focus only on a part of the brain that the task or stimuli are already known to activate. This reduces the number of voxels to be analyzed and thus the feature space. This approach, using regions of interest (ROIs) has been used in e.g. brain reading studies, where experimenters were just looking at activations in the visual cortex. As an other example, experimenters interested in language might focus their analysis on brain regions known to be associated with language. A review on 275 fMRI and PET studies by (10) shows that ROIs have been used in studies on a broad variety of cognitive tasks (e.g. attention, perception, working memory ,episodic memory).

We could use a ROI approach like this on our data, but this would not suit the goal of this study. We try to infer an object hierarchy from the brain on a purely data driven basis. Of course we already know that there should be a class distinction in the motor cortex, but by narrowing our search down only to the motor cortex we run the risk of missing object distinctions that are not associated with manipulability. These might as well be registered in other parts of the brain, which means that we might miss new and unexpected hierarchies. Apart from this, narrowing our analysis down to just one part of the brain would be too much like peaking, since we increase the chance of finding the manipulability class distinction. But if we would apply our approach to a whole new set of nouns with no intended distinction, we could not use a ROI approach. For generalizability of our approach thus, it is better to do a whole-brain analysis.

An other way to bring down dimensionality is to combine chunks of voxels and derive a signal from each of these chunks. Basically, this is what also happens in spatial down-sampling, where nearby voxels get combined and an average signal is calculated for the new, larger voxel. Note that this step in preprocessing left us with 27 x 32 x 23 = 19872 voxels. If we were to perform an extra step of down-sampling to bring the number of voxels down to around 40, we would have to concatenate chunks of $\frac{19872}{40} = 497$ voxels. Voxels combined to one chunk might be in completely different brain regions, could exist of non-neural matter or could even be outside the brain. This would leave each chunk with a lot of noise. Averaging over such a chunk would probably throw away all interesting signal. Instead, it makes more sense to combine voxels that are in the same brain region, because on the basis of the similar function of neurons in the same brain regions, we can assume that they would respond in a similar way to stimuli. Furthermore, such an approach would only take into account in-brain and neural matter. Instead, we used two other techniques.

Area	Comments
12	Too small for detection.
14	Only for non-human primates.
15	Only for non-human primates.
16	Non-existent.
26	Too small for detection.
27	Only for non-human primates.

Table 3.1: List of Brodmann's areas excluded from our analysis.

3.2.1 Brodmann's areas

In their review on 275 fMRI and PET studies (10) use Brodmann regions (Figure 3.2) to describe the localization of cognitive components. It is of course possible to divide the brain into areas based on other criteria, but both the functional characteristics of the Brodmann's areas and their size make them the ideal candidate for this problem. Their nature is such that we can still oversee the dimensionality of the obtained patterns, while functional similarities within areas are preserved. Since in the human brain there is only a limited number of Brodmann's

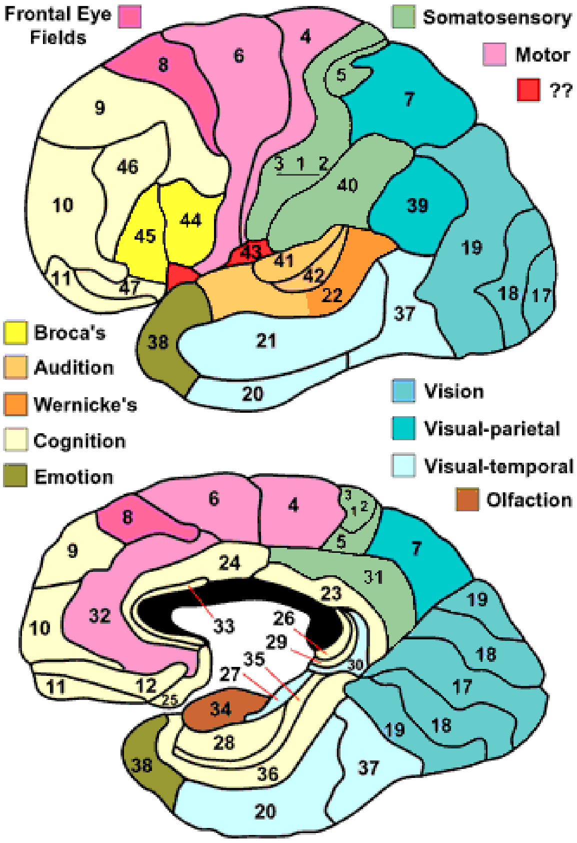


Figure 3.3: Korbininan Brodmann's anatomical description of the brain, based on the organization of neurons he observed in the cortex. From (17).

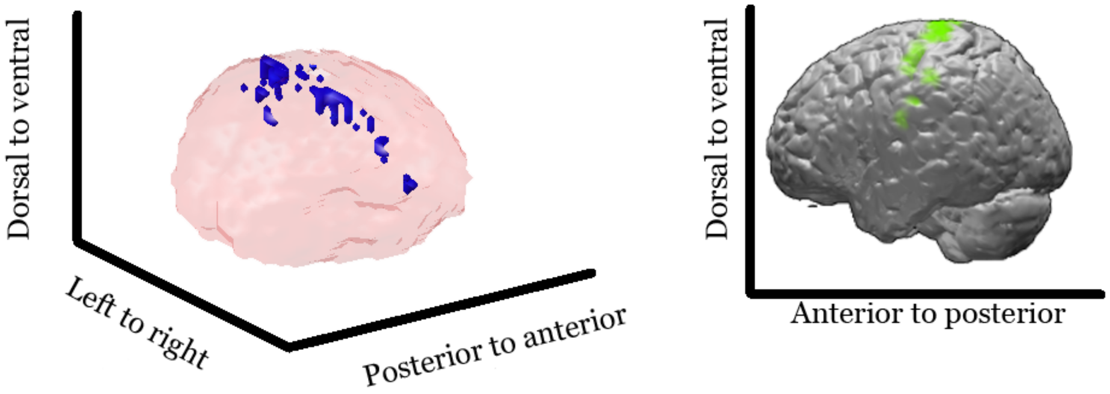


Figure 3.4: Labeling of voxels by mapping our brain to a standard brain. The left subfigure contains the mapping of Brodmann's area 4 (primary motor cortex) on the brain of subject 3. The right subfigure shows Brodmann's area 4 on a standard brain, adapted from (18).

areas, averaging over Brodmann's areas would bring the size of neural activation patterns down from 19782 to somewhere around 40, depending on the number of areas mapped. For every Brodmann's area, we average the signals of the voxels in that area to get one signal. For this, we map our brain to a standard Brodmann's area brain atlas ((19), (20)) using the Fieldtrip Matlab package. This leaves us with quite accurate estimates of the brain regions (Figure 3.2). Due to the size of our voxels and the characteristics of the mapping procedure, some areas did not get any signal. Table 3.2 lists the areas that were excluded from our analysis. In total, 41 Brodmann areas were included.

3.2.2 Principal component analysis

An other way to go would be to use dimensionality reduction techniques that are based solely on the information in the data. These multivariate techniques transform the set of signals to a new coordinate system based on components, where the first component explains the greatest variance in the data, the second coordinate the second greatest variance and so on. Two of these techniques are very commonly used in the field of fMRI analysis, either as a stand alone analysis method (21) or in combination with existing methods such as linear regression (22), (23): independent component analysis (ICA), which tries to find components that are statistically independent, and principal component analysis (PCA), where components only have to be uncorrelated. A third method, singular value decomposition (SVD) is less used, but yields results identical to those of PCA in less time and is thus a worthy opponent of PCA.

Component analyses are used as an exploratory analysis in many fields of research, since they have the capability of translating a large set of variables to a much smaller set of components. In our case, we would have to run a component analysis on the set of 19872 voxel signals and take an acceptable number of components from this to reduce the number of dimensions in the clustering analysis, but on the other hand get as good a representation of the original signals as possible. To get pattern lengths similar to those of the Brodmann's area approach, we would have to take around 40 components. On the other hand, there are some statistical restrictions to the number of components used. A rule of thumb is to use those components that together make up 99% of the variance in the data.

Of the three methods proposed here, ICA is the most commonly used in fMRI. ICA could be applied either spatially (sICA, finding a set of mutually independent activation images) or temporally (tICA, finding a set of mutually independent time courses). Furthermore, one could use spatio-temporal ICA (stICA), which finds a trade off between independence in time and independence in space. Using either sICA, tICA or stICA has yielded good results compared to SVD and PCA methods (24). The big drawback in using ICA is its computation time, which by far exceeds that of PCA and SVD. When differences between SVD/PCA and ICA performance are small, one might want to use the computationaly more attractive SVD/PCA approach. As (25) points out, using ICA on fMRI data is all about separating signals of interest (e.g. task-related signal, transiently task-related signal) from signals not of interest (e.g. physiology-related signals, motion-related signals). As ICA does, SVD and PCA try to find components that explain the variation in the data. These components however have to be uncorrelated, but not independent, which makes computation faster. Box 3.1 shows the calculation steps in PCA and SVD. As this box shows, both methods generate eigenvectors as components. The difference is in the calculation. PCA uses a covariance matrix, the generating of which could take a lot of time when using 17982 signals. SVD calculates the components directly on the data, thus taking less time. Therefore, we chose to use SVD as our method of data-driven dimensionality reduction.

Principal Component Analysis

1. $X = N$ x d data matrix, with one row vector x^n per signal.
2. Subtract $\bar{x}$ from each x^n.
3. Calculate covariance matrix of X, Σ.
4. Find eigenvectors and eigenvalues of Σ.
5. First M eigenvectors are M principal components.

Singular Value Decomposition

1. $X = N$ x d data matrix, with one row vector x^n per signal.
2. Subtract $\bar{x}$ from each x^n.
3. Solve $X = USV^T$
4. First M columns of V are M principal components.

Box 3.1 PCA and SVD calculation steps.

3.3 General Linear Model analysis

The dimensionality reduction stage outputs sets of either 41 or 50 signals for each subject. In the univariate GLM stage, we induce a response for each stimulus from each of these signals using linear regression. For each stimulus then, we get either 41 or 50 β-values, one for the strength of the stimulus' regressor in each signal. These β-values make up the patterns to be analyzed in the unsupervised clustering stage. The basis for using a GLM in fMRI data analysis is very strong, with model-based approaches like this one being used in a majority of fMRI studies (7). Most GLM fMRI studies use the SPM Matlab package (Statistical Parametric Mapping, `www.fil.ion.ucl.ac.uk/spm`) , which also contains options for preprocessing and statistical assessment.

Using a general linear model in fMRI analysis implies making the assumption that the fMRI data is a linear model made up out of model factors with parameter weights and an added amount of noise. Though most fMRI studies use a GLM approach, it is debatable whether the linearity assumption holds when the same stimuli are presented with a small interval (26), (7). In that case, the refractory period of the activated brain area might cause the response to be smaller, with the linear model overestimating the actual response. Though models that work under a nonlinear assumption are available (27), most studies are performed under the linearity assumption when intervals between subsequent trials are longer than the refractory period. A good compromise is that ITIs of 6s form the boundary between seeing the fMRI signal as a nonlinear system and assuming that the signal is linear (7). Since our experiment used ITIs of 8s, we can thus assume that we are dealing with a linear system.

$$\mathbf{X} = \mathbf{G} \cdot \beta + \varepsilon, \quad \mathbf{X} = [d \times T], \quad \mathbf{G} = [T \times R], \quad \beta = [R \times d] \tag{3.1}$$

In the standard GLM for fMRI (Equation 3.1), data is denoted as a matrix **X** with d voxels signals as rows where the number of columns T equals the number of samples, the model factors are gathered as R column regressors in a design matrix **G** and the parameter weights are represented as a vector β, with one parameter for the contribution of each regressor. Then there is an additional amount of unexplained signal, which is noise denoted with ε. Figure 3.5 shows the basic principles of a GLM in fMRI. Regressors in the design matrix can be thought of as a set of basis functions that make up the signals in **X**. What is important here is that there are two known variables, **X** and **G** and that there is one variable β that is not known. This problem can easily be solved using least squares linear regression as

$$\hat{\beta} = (\mathbf{G}^t\mathbf{G})^{-1}\mathbf{G}^t\mathbf{X} \tag{3.2}$$

Since the design matrix models the factors in the experiment, it is important to decide upon what regressors to use. The GLM analysis was originally performed on the whole signal set by (8), using 13 regressors. For each of four conditions (FM, VM, Nonword, Null) there was one regressor, there were six regressors for motion correction, one regressor for a constant scanner baseline and there were two regressors for the derivative of the HRF in the FM and VM conditions. These last two regressors are used to account for individual differences in BOLD response latency. In our analysis, we use one regressor per stimulus, since we want to infer one β-value for each stimulus instead of one β-value for each condition. Since we do data-driven analysis, we cannot make any assumptions about which stimuli belong together. Depending on the dimensionality reduction method taken, our design matrix will include motion regressors. The Brodmann's area preprocessed data will be treated with a set of motion regressors, since motion artifacts are assumed to be preserved in the preprocessing method and thus in all signals resulting from preprocessing. For the SVD preprocessed data, no motion regressors are used, since the motion artifacts are believed to be captured in specific components and not in all components. Furthermore, a baseline regressor has to be included to amount for the constant proportion of the signal.

A stimulus regressor is made up of the convolution of the canonical hemodynamic response (Figure 2.1) with a constant signal that has a spike at the time t_i when stimulus i is presented. Since we have 120 stimulus presentations, we have 120 stimulus regressors in our design matrix. Also, there are six motion regressors and one baseline regressor in our GLM. Regression thus takes into account all trials.

The matrix β contains either 41 or 50 β-values for every trial. We can consider the rows of the β-matrix to be the activation patterns for the trials, where every β-value is an activation feature. Using the experiment setup files, we then match every pattern to a presented stimulus. Of the 120 trials, we only keep the 80 trials where a VM or FM word, a critical stimulus, was presented. Of these, we remove the false positive trials where the subject made an error in perceiving the stimulus as a nonword. At most, subjects made 3 of these mistakes. For a subject, we then get one pattern of β-values per critical stimulus.

3.3.1 On the use of HRF time derivatives

Most standard fMRI experiments use first order time derivatives of the condition regressors to account for slice timing differences or shifts in the hemodynamic response due to individual differences (28). Adding a temporal derivative shifts the HRF in time (Figure 3.3.1). Using second order derivatives of the HRF is also quite common and adds even more flexibility to the estimated HRF. It is common practice to include derivatives for every condition regressor. One then has the opportunity to either contrast conditions based only on the condition regressors or to also include the β-values for the derivatives (28). According to (16), using a temporal

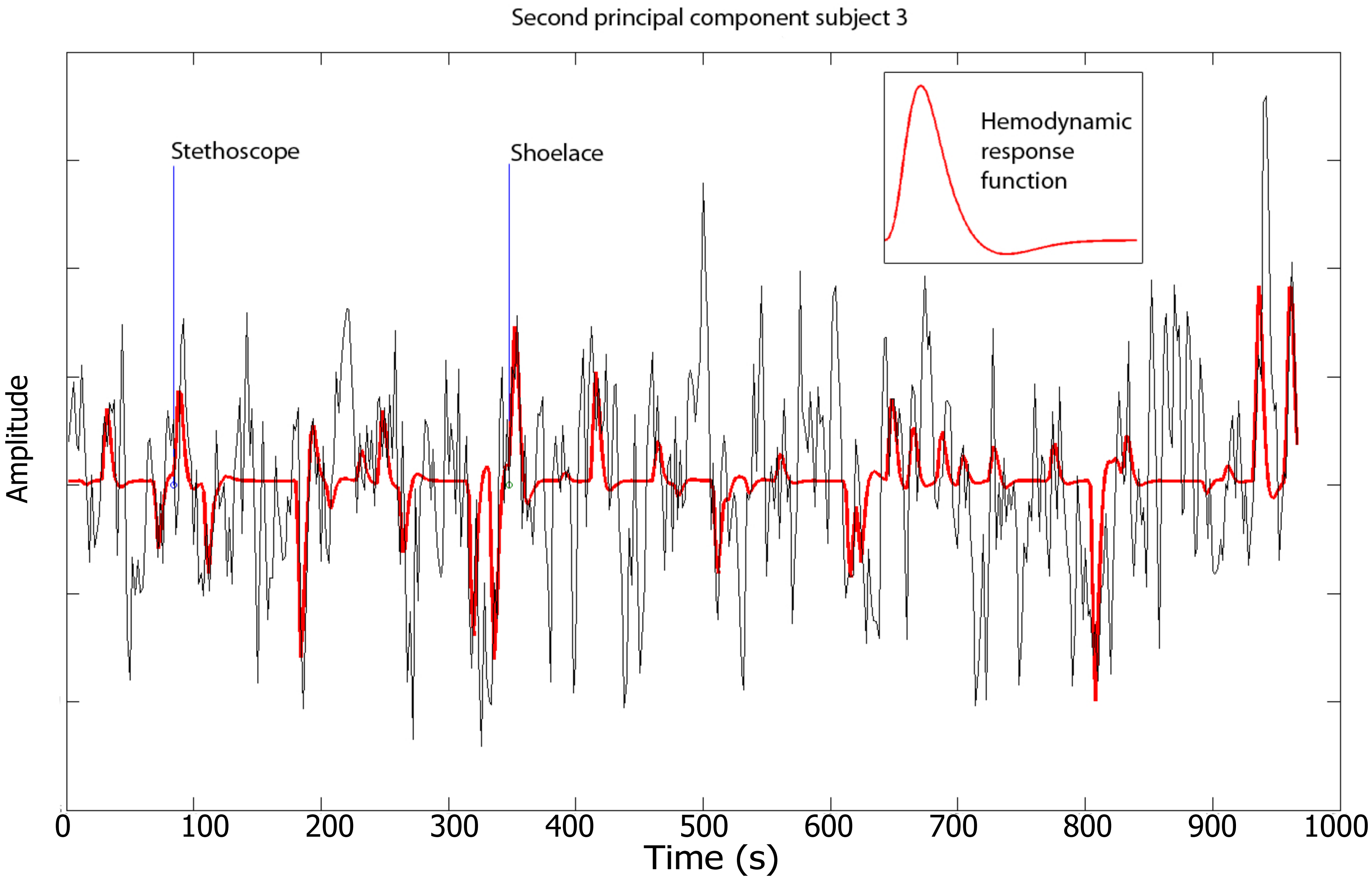


Figure 3.5: Basic principles of the general linear model in fMRI. For each stimulus, here ’shoelace’ and ’stethoscope’, the GLM estimates the contribution of the corresponding hemodynamic response to the overall signal. This illustration includes only stimuli with a FM label. Would it include also those of the other three classes, the fit would be near to perfect, with a very low variance.

derivative has its drawbacks. Though the regression model fits the residuals better, it can reduce the power of a t-test because it increases variance in the residuals. A comparable empirical GLM parameter study by (29) however states that using a temporal derivative can give a slight advantage towards estimation of the neural response. By all means, one should be careful when adding additional regressors to the design matrix, since they can have an effect on the estimated responses and overfitting mistakes are easily made.

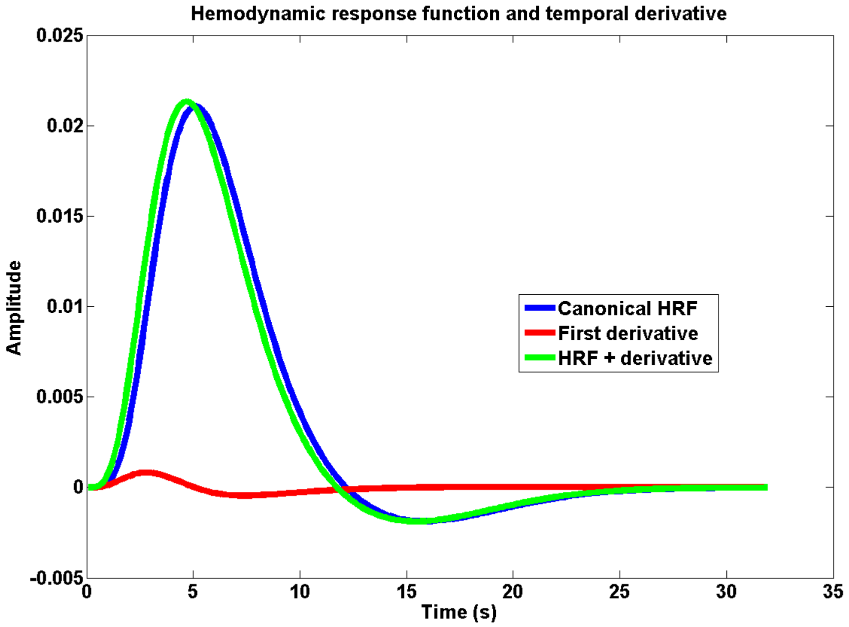


Figure 3.6: Shift of HRF by including the temporal derivative. The blue line is the original canonical HRF, the red line is the derivative of the HRF and the green line is the combination of HRF and temporal derivative

In our case, taking the temporal derivative into account would mean either adding one regressor to the model to account for hemodynamic and slice-timing delays in all conditions or adding one regressor per stimulus to account for the possible delay in the response elicited by that stimulus. Since we assume that we do not know which stimuli belong together, it is impossible to group stimuli and add one derivative per group. Adding one derivative per stimulus would mean that we would be fitting more than 240 (2 regressors per 120 stimuli, plus additional baseline and motion regressors) on a signal of around 480 time-points. This would mean that there are only 2 times more time-points then variables. Fitting such a model would result in very questionable activations estimates (30). It is recommended to use at least 10 to 20 times more data points than regressors, since adding more regressors will decrease statistical power of the model. Therefore, we do not take this approach to using the temporal derivative. In fact, we have decided to leave the temporal derivative out. Adding just one derivative term would mean fitting delay in 120 presumably very dissimilar hemodynamic responses with one derivative term. This didn't seem to make much sense either and would probably add very little to the estimations, while increasing the chance of overfitting.

3.4 Unsupervised Clustering

The result of the feature extraction step is a matrix $\beta = [R \times d]$ with one activation pattern per critical stimulus. These activation patterns are to be clustered using unsupervised clustering methods. As mentioned in Chapter 2, unsupervised clustering methods can be categorized among different dimensions. Here, we have used four different techniques that cover the domain of unsupervised clustering methods, ranging from simple to relatively complex methods. Essentially, clustering tries to minimize the dispersion within clusters and maximize the distance between clusters. Several distance metrics can be used to define dispersion and distance. Using different distance metrics can yield different results, because metrics handle e.g. outliers in different ways. We have employed three distance metrics (Figure 3.4) for the distance between two patterns **x** and **y** of dimensionality n. The simplest is Chebychev or L_0 distance (Equation 3.3), where the distance is given by the minimum distance on any one dimension i. For L_1 or City Block distance (Equation 3.4), the metric is given by the sum of the distance over all dimensions. The most common metric, Euclidean or L_2 distance (Equation 3.5), is given by the square-root over the sum of the squared distance on all dimensions.

$$d(\mathbf{x},\mathbf{y})_{L_0} = argmin_i |x_i - x_y| \tag{3.3}$$

$$d(\mathbf{x},\mathbf{y})_{L_1} = \sum_{i=1}^{n} |x_i - y_i| \tag{3.4}$$

$$d(\mathbf{x},\mathbf{y})_{L_2} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{3.5}$$

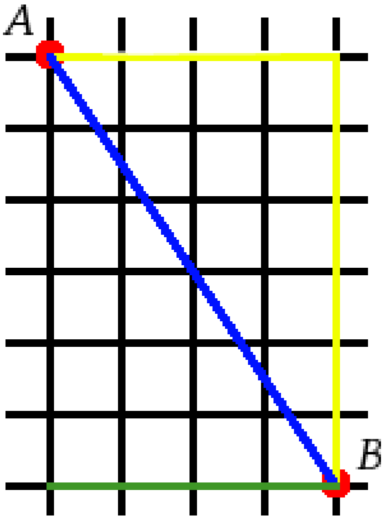


Figure 3.7: Distance metrics L_0 (green line), L_1 (yellow line) and L_2 (blue line) in two dimensional space.

3.4.1 Linkage clustering

The simplest clustering algorithms combine clusters of patterns in a bottom-up fashion according to some combination criterion. Starting with n clusters, one for every pattern, clusters are combined pairwise until all clusters are combined in one large cluster containing all individual patterns. This type of clustering yields a hierarchy among patterns, where categories of patterns that are similar are merged into a supercategorie. Hierarchical clustering algorithms like these yield dendrograms which show how the clusters are united (Figure 3.8). Intuitively, this would lead one to think that patterns corresponding to e.g. cats are clustered together as are patterns corresponding to dogs and that these two clusters are merged at a higher level. Different criterions for merging exist. The most common are single link clustering and complete link clustering, which are based on the distance between elements in clusters. We have also employed Ward linkage, which aims to minimize the gain in error by combining two clusters (31). One can cut off the hierarchy at a particular level to obtain k clusters.

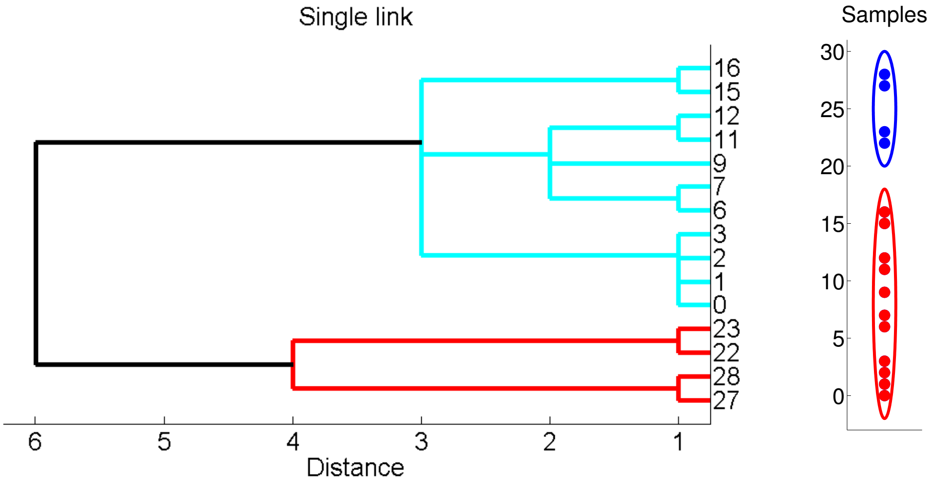
Linkage clustering algorithms try to optimize an objective function on every merging step. Single link clustering combines clusters on the basis of the distance between their closest elements. Pairs of clusters for which this distance is minimal are combined into new clusters. The complete link clustering algorithm also combines pairs of clusters on the basis of the distance between elements, but instead of taking the distance between the closest elements, it takes the distance between the elements furthest away from each other. Ward link clustering tries to minimize the increase in error that would result from combining two clusters. In Joe Ward's 1963 paper, the objective is to minimize the increase in the error sum of squares ESS on every merging of clusters. The ESS for a particular clustering is given by

$$ESS = \sum_{c=1}^{k} \sum_{i=1}^{n_c} d(\mathbf{x_{ic}}, \mathbf{\bar{x}_c})^2 \tag{3.6}$$

where k is the number of clusters, n_c is the number of elements in cluster c and $d(\mathbf{x_{ic}}, \mathbf{\bar{x}_c})$ is the distance between element i in cluster c and the centroid of cluster c. This ESS is zero in the original partitioning, where every element equals the centroid of its cluster and the sum of the distances is thus zero. As the number of clusters decreases, ESS increases. For a total of $n-1$ steps, in step t Ward's algorithm then chooses to unite two of the clusters from $t-1$ whose combination minimizes $ESS_t - ESS_{t-1}$.

3.4.2 *k*-means clustering

While linkage clustering is an agglomerative method, seeking to combine clusters bottom-up, k-means clustering is a divisive method, trying to find positions for cluster centroids which minimize the dispersion in every cluster. The algorithm starts out with k random centroids. Every element is appointed to the nearest of these centroids. Then, every centroid is moved so that it is in the center of its appointed elements. Again, elements are appointed to their nearest centroid after which the centroids are repositioned. On every step, the ESS is calculated. The objective is to minimize the ESS. Once there is no more decrease in the ESS or membership of elements is constant, the convergence criteria are locally set to be met and the centroids that minimize the ESS are found. Several variants of this algorithm exist, some trying to start off with good initial centroids based on a heuristic , others having a dynamic number of clusters k (11). The latter variant poses a solution to the problem where the exact number of groups is



(a)

(b)

Figure 3.8: Dendrogram (a) for the single link clustering of a one-dinmensional sample set $\{0,1,2,3,6,7,9,11,12,15,16,22,23,27,286\}$ (b) using the L_2 distance metric.

unknown. Since the algorithm is usually very fast, we will run it several times with random starting centroids on each combination of subject and preprocessing pipeline.

3.4.3 Genetic Algorithm

An other approach to solving clustering problems is to use a Genetic Algorithm. Such an algorithm could find the optimal partitioning of a data set based on recombination and mutation of a population of initially random partitionings. Typical steps in a GA are

1. *Population initialization*: Initialize a population of n random individuals.
2. *Selection*: Select individuals for recombination based on their fitness value, according to some fitness function.
3. *Reproduction*: Create new individuals by recombinination or mutation of the individuals selected in step 2.
4. *Termination*: After a fixed amount of generations or when a termination criterium is met, stop. If not, return to step 2.

GAs are infamous for the high number of free parameters in their implementation. Every fenotype, in this case a partitioning, is represented by a genotype or individual in the population. For a clustering implementation, there are two possible representations. One can make an integer string with length n is the number of patterns or elements. For every element, the integer at that position denotes the number of the elements cluster. This implementation was succesfully used by (32). The fitness function in their implementation was the inverse of the ESS. The higher this number is, the lower the ESS is and the better the partitioning minimizes the clusters dispersion. In step 2, the fittest individuals are selected, while in step 3 these individuals are combined via cross-over. The drawback of using this kind of representation is that its length increases with the number of samples. Since complexity increases with the length of the representation, it might be better to take a representation that is only dependent on the number of dimensions and the number of clusters, since these values are less flexible in our problem.

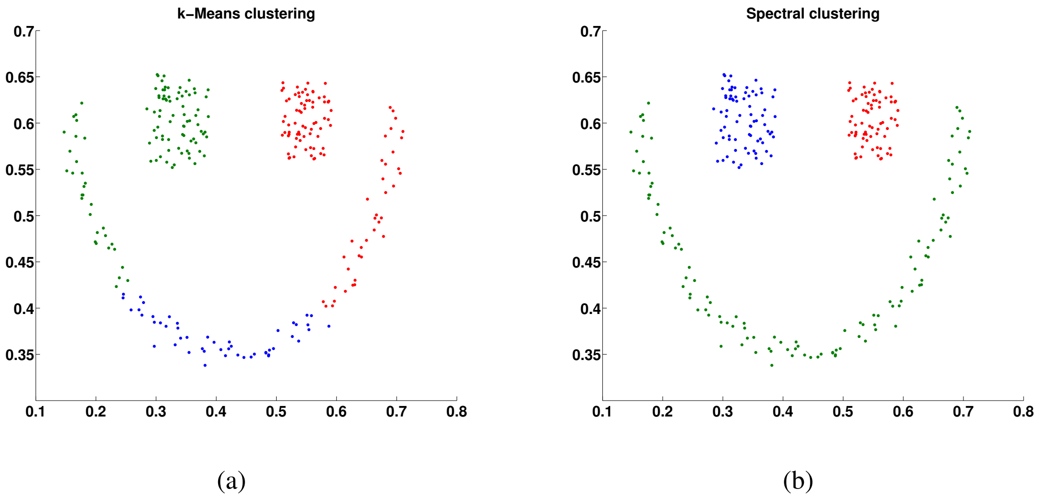
An other approach is to let the genotypes denote the cluster centroids. This yields a string of floats as a genotype on which selection and reproduction can be performed (33). The length of this string is $[D \times K]$, where D is the dimensionality of the element (in this thesis either 41 for Brodmann preprocessed data or 50 for PCA preprocessed data) and K is the number of centroids. In step 2, elements are first appointed to their nearest cluster centroid. Then, centroids are moved so that they are at the centre of their appointed elements. Then the fitness is computed on the float string as a whole. Again, fitness equals the inverse of the ESS. Note that this technique is quite similar to the k-means algorithm. The difference is that we run a whole population of k-means algorithms at the same time and try to find the optimal centroids position by searching as large a portion of the search space as the population size permits. In k-means, we are restricted by the initial positions of the centroids. The question is whether running a batch of k-means yields the same results as running a GA. In that case, using k-means would be the best choice because of its low computational cost.

In our implementation, we use the centroid coordinate representation by (33). Our fitness function is the same as that in (33) and (32), that is the inverse of the ESS. In their paper, (33) propose methods for recombination and mutation which we adopt. For crossover, simple single-point crossover is used, where two parent genotypes generate two new genotypes. For chromosomes of length l, a random integer is generated in the range $[;-l-1]$ and the portions of the chromosomes lying to the right are exhanged to produce two new offspring. On mutation, a small number $\pm\delta$ is added to the float v at a random position in the genotype. For selection, we use tournament selection, where two random parents are drawn from the population, crossover is performed and the fittest of these four is put into the new population so that the size of the new population is equal to that of the old one.

3.4.4 Spectral clustering

As a fourth clustering method we employed spectral clustering. In contrast to the other three clustering methods, spectral clustering does not cluster on the position of the patterns in high-dimensional space, but on the basis of a similarity matrix derived from the patterns. The spectral clustering algorithm is extensively described in a tutorial by (34) and in a paper by (35). In the latter paper, six steps in the spectral clustering process are described in a way that is reproduced here. Given our set of patterns $S = \{s_1, \ldots, s_n\}$ in $\mathbb{R}^D$, where D is the dimensionality of the patterns that we want to cluster into k subsets:

1. Form the affinity matrix $A \in \mathbb{R}^{D \cdot D}$ defined by $A_{ij} = exp(\frac{-||s_i - s_j||^2}{2\sigma^2})$ if $i \neq j$, and $A_{ii} = 0$.
2. Define D to be the diagonal matrix whose (i,i)-element is the sum of A's i-th row and construct the matrix $L = D^{-1/2} A D^{-1/2}$
3. Find $x_1, x_2 \ldots x_k$ the k largest eigenvectors of L and form the matrix $X = [x_1 x_2 \ldots x_k] \in \mathbb{R}^{D \cdot k}$ by stacking the eigenvectors in columns.
4. Form the matrix Y from X by renormalizing each of X's rows to have unit length.



(a) (b)

Figure 3.9: k-means clustering (a) and Spectral clustering (b) on a 2-dimensional data set.

5. Treating each row of Y as a point in $\mathbb{R}^k$, cluster them into k clusters via k-means or any other algorithm that attempts to minimize distortion.

6. Assign the original points s_i to cluster j if and only if row i of the matrix Y was assigned to cluster j.

This algorithm tries to find clusters based not on the points in high dimensionality, but only on the points in k-dimensionality. Then, the original data points are mapped onto the clusters in k-dimensionality. This method has shown to outperform simpler methods like k-means in many situations, for an example see Figure 3.4.4. One could think of spectral clustering as a method that creates a graph representation of the data and then removes particular vertices to split the graph in k pieces. These pieces then are the clusters. The decision on what vertices to cut is based on the eigenvalue decomposition and this decomposition is based on the affinity matrix A. The way in which this matrix A is formed, is very important. Here, we have used the function proposed in step 1, but many other functions are possible. In our implementation, there is one important scaling parameter σ, which defines the affinity between patterns. The higher σ is, the faster the affinity between two patterns falls off with the distance between the patterns. One option to find an appropriate value for σ is trying out different values and then picking the value that minimizes cluster dispersion. We have used a self-tuning technique for this.

3.5 Cluster Validation

We want to infer a hierarchy of nouns based on clusters of neural activation patterns, but in order to make claims about the reliability of the obtained partitioning of patterns, we need to have some measure of validity of this partitioning. Furthermore, having a measure of validity can help determine the optimal number of clusters k in the data. This in fact has by far been the most common application of cluster validity techniques (13). Methods to assess cluster validity can be roughly separated into two categories (36). Internal validity is based only on the characteristics of the found clusters and their patterns. External validity assessment methods measure the similarity between a found partitioning and an other, possibly known to be present, clustering. Numerous methods exist, but here we will present two internal validity indices and the single external validity index used.

3.5.1 Internal Validity

Different interval validity measures have been proposed. Usually, such measures allow for an investigation into the optimal number of clusters. If we take the clustering in Figure 3.4.4, we would say that there are three clusters there. One for each circle and one for the bowl. We could run a k-means or spectral clustering algorithm on this data with different values for k and measure the validity of the clustering for each k. If we have a good validity measurement, we would then find that three is the optimal number of clusters in this dataset.

The Dunn cluster validity index is based on the maximum cluster dispersion and the minimum intercluster distance ((37),(38)). When we have n_c clusters in a d-dimensional dataset, we will get a Dunn index of

$$D = \min_{i=1...n_c} \left\{ \min_{j=i+1...n_c} \left(\frac{d(c_i, c_j)}{\max\limits_{k=1...n_c} (diam(c_k)} \right) \right\} \tag{3.7}$$

$$d(c_i, c_j) = \min_{x \in c_i, y \in c_j} \{d(x, y)\} \tag{3.8}$$

$$diam(c_i) = \max_{x, y \in c_i} \{d(x, y\} \tag{3.9}$$

Here, equation 3.8 denotes that the distance between two clusters is given by the distance between their closest pair of members. In equation 3.9 the diameter or dispersion of a cluster is given as the distance between its two most distinct members. Equation 3.7 shows how in the partitioning a pair of clusters is found with minimal intercluster distance, while on the other hand the cluster with maximum intracluster dispersion is found. Separating the first by the latter will yield an index that is high for compact and well separated clusters but low for clusterings where clusters are very close to each other. Note that the distance between clusters is measured in a way that is similar to that used in single and complete linkage clustering.

While essentially very simple, this method has some drawbacks. The main disadvantage of the Dunn index is its sensitivity to noisy data. When a cluster is compact, but has a few outliers, the actual diameter of the cluster could easily be misestimated. This effect is strengthened when outliers of neighbouring clusters are close to each other. A better way would be to define cluster dispersion as the average distance to a centroid and use the distance between centroids as a measure of intracluster distance.

The Davies-Bouldin index for cluster validity matches these criteria (39) . It measures the average of similarity between each cluster and its most similar one by using a similarity measure R_{ij} between clusters

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}, d_{ij} = d(v_i, v_j), s_i = \frac{1}{||c_i||} \sum_x \in c_i d(x, v_i) \tag{3.10}$$

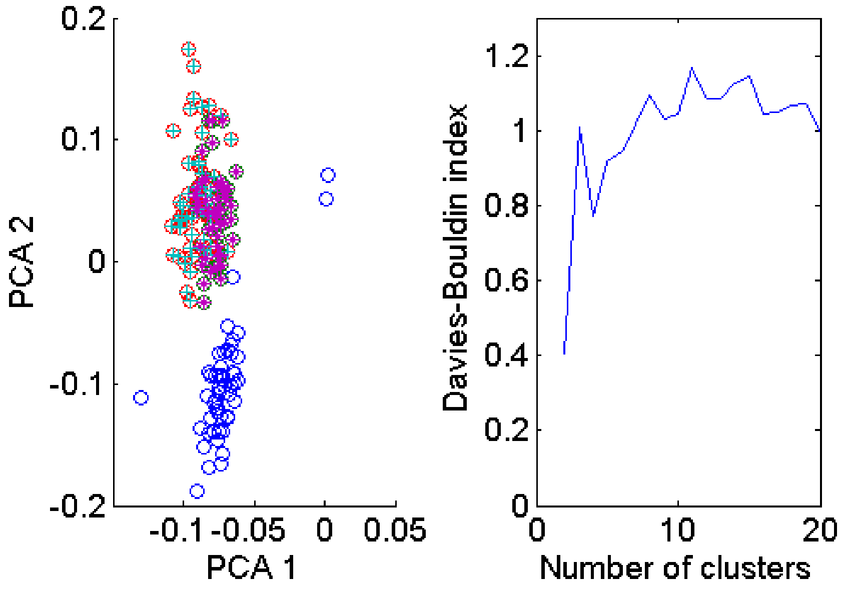


Figure 3.10: Davies-Bouldin index for the first two principal components in the Iris data set. The figure on the left shows the patterns in 2D-space. The figure on the right shows different values for the Davies-Bouldin index for different values of n_c.

where v_i and v_j are the centroids of cluster i and cluster j respectively, c_i is the i-th cluster and $||c_i||$ is the number of patterns in cluster i. The Davies-Bouldin index is then given by the average of the minimal value of all R_{ij}'s for every i or

$$DB = \frac{1}{n_c}\sum_{i=1}^{n_c} R_i \tag{3.11}$$

where

$$R_i = \max_{j=1\ldots n_c, i\neq j}(R_{ij}), i = 1\ldots n_c \tag{3.12}$$

A demonstration of this method in order to find the optimal number of clusters in a dataset can be found in Figure 3.10. The first two principal components where taken from a PCA on Fisher's Iris data set (40). It is clear here that the blue datapoints denoting the Iris Setosa are linearly separable from the red and green datapoints which denote Iris Virginica and Iris Versicolor. Based on this plot, a value of $n_c = 2$ seems to be the best setting for a k-means algorithm. The subfigure on the right, where the Davies-Bouldin index is plotted against different values for n_c, shows the same. Since a low index means that clusters are not similar, $n_c = 2$ seems to be the best pick.

3.5.2 External Validity

External validity measures are based on the similarity between the found partitioning Y and an other, external, partitioning Y'. Both Y and Y' have length n equal to the number of patterns in the dataset, where each pattern X_i has an assigned cluster k. We assume that the number of clusters in both partitionings is equal to n_c. In clustering methods, the labelling of clusters is often quite arbitrary. A cluster could be labelled 1 in one k-means run, while in another run it gets the label 2, depending on the order in which the random centroids where placed, while the patterns in each cluster are the same on both runs.

A similarity measure between partitionings should ignore the values of the labels and only take into account which patterns do and which patterns do not share the same label. The index proposed by William Rand (41) does just that. A $[n \times n]$ matrix R is formed for each clustering, where R_{ij} is 1 if X_i and X_j are in the same cluster and 0 if they are not. Then, matrix $\gamma = 1 - abs(R - R')$ is formed. This matrix has values γ_{ij} is 1 if two patterns are in the same cluster in both partitionings or if they are in different clusters in both clusterings. Otherwise, this value is 0. The number of zeros of this matrix is then taken and divided by the total number of possible pairings to yield the Rand index with

$$c(Y, Y') = \frac{\sum_{i<j}^{N} \gamma_{ij}}{\begin{pmatrix} N \\ 2 \end{pmatrix}} \tag{3.13}$$

This index always has a value between 0 and 1. A value of 0 means that two clusterings are maximally dissimilar, a value of 1 means that they are identical. We will use this method to assess how well the found clusterings match the distinction we already know to be present in the data set, that is the difference between FM and VM words. The clustering to which all found clusterings will be matched has a label 1 for all FM words and a label 2 for all VM words. Though we keep in mind that there is the possibility of finding more than only this class distinction, we know this distinction to be at least present in the stimulus set. It thus offers a good baseline to compare our clustering results to.

Chapter 4

Results

The methods introduced in Chapter 3, up to the point of unsupervised clustering, have been applied to our data set. All preprocessing steps were taken to the point where there were 28 pattern sets, one for each dimensionality reduction method performed on each subject session. Each pattern set consisted of 78 to 80 patterns, depending on the number of errors the subjects made in the lexical decision task. Here, the results of running the different algorithms will be presented. For every clustering configuration, we can look at both the interval validity, using Dunn and Davies-Bouldin index, and the external validity, taking the Rand index with relation to the FM-VM labeling of the patterns. Chance level for the Rand index lies at 50%. Furthermore, we will make a qualitative estimation of the best clustering results obtained (measured by internal validity) to see by eye if the obtained hierarchy contains any semantic information.

4.1 Linkage Clustering

Complete, single and Ward linkage clustering have been performed on the data. Here, external validity is given by the Rand index of the two main clusters, obtained by cutting off the hierarchy at one step from the top. Figure 4.1 shows the external validity for all pattern sets, one plot per clustering configuration. For both single and complete link clustering, distances L_0, L_1 and L_2 have been used. For Ward linkage, only distance L_2 was used. As we can see, the external validity values are centered a bit below chance level. There seems to be little similarity between the FM-VM labeling and the found clusters. This might be because linkage clustering algorithms often produce one large cluster and one small cluster at the top level. The Rand index between such a clustering and an evenly spread clustering equals 0.5, because for half the pattern pairs there is no difference between the two clustering. In fact, both single linkage and complete linkage seem to give clusterings with an extreme ratio between clusters sizes, an example of which can be found in figure 4.3.

An other interesting thing to look at is the similarity between subject-specific clusterings. Figure 4.2 contains a mesh plot of the between-subject Rand indices, where we see a peak similarity value between subjects 2 and 4 on L_2 ward linkage clustering. Though this looks promising, when looking at the pattern distribution we see that the average cluster size ratio in single link clustering is $\approx 79 : 1$ with a SE of only 0.25 pattern. In Ward clustering, this is $\approx 73 : 7$ with a SE in the larger cluster of 7 patterns. It is highly unlikely that such pattern distributions contain information about the inherent classes in the pattern set. High intersubject Rand indices are either artifacts of extreme cluster size ratios, as with the single link and complete link methods, or of a large difference between cluster size ratios, as we have seen in calculating the similarity between found clusterings and FM-VM labeling. In fact, the cluster size ratios of subjects 2 and 4 in Figure 4.2 are 61:19 and 43:37. The difference between these two values, as the large variation in intersubject Rand indices, lead to the assumption that the peak value is probably caused by such algorithmic artifacts.

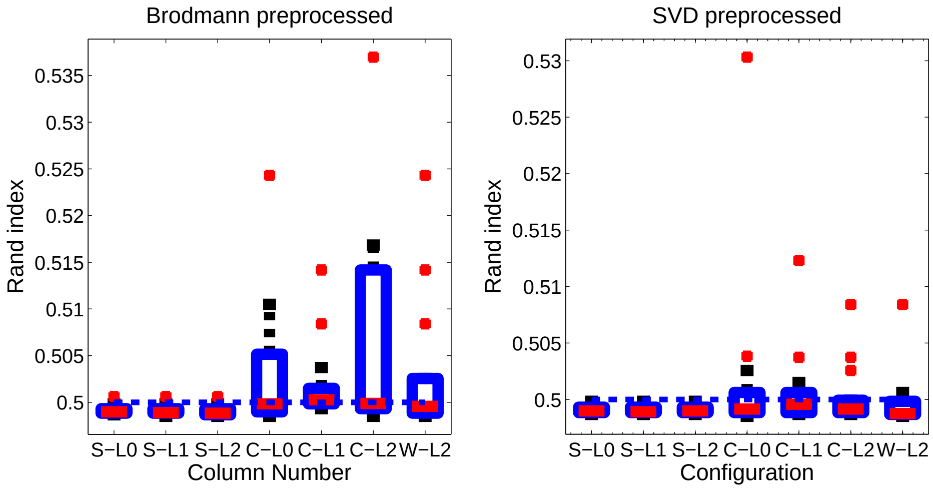


Figure 4.1: External validity of different linkage clustering configurations. S means ’single’, C means ’complete’, W means ’Ward’. L_0, L_1 and L_2 stand for the different linkage metrics. The figure on the left shows the distribution for the Brodmann preprocessed patterns, the figure on the right shows these for the SVD preprocessed patterns.

The internal validity of linkage clustering results can give an indication for the correct number of clusters. Internal validity indices are either minimal (Davies-Bouldin) or maximal (Dunn) when intercluster distance is maximal and intracluster dispersion is minimal. We can determine the optimal number of clusters by cutting off the dendrograms at different levels and calculating the interval validity for the obtained clusterings. As we can see in Figure 4.2, cutting off a single linkage clustering dendrogram at one level from the top would leave us with a cluster size ratio of 78:1:1. This will lead to a monotonous trend in internal validity index for these kinds of clusterings, with the optimal number of clusters being the number of patterns n. Though mathematically correct, this is very uninformative, since we are looking for some general similarity among patterns. Therefore, we choose to perform this analysis on the combination of linkage algorithm and distance measure that yields the fairest cluster size ratio between the two largest clusters. This is true for complete linkage clustering using L_0 distance metric, with an average ratio of $\approx 56:24$ for sets preprocessed using Brodmann’s area dimensionality reduction and $\approx 62:18$ for sets preprocessed using SVD. Also, we calculate the internal validity for Ward clustering using L_2 distances, the second best configuration in terms of cluster size ratio. The fact that there is a difference between these two preprocessing methods suggest that there might be a difference in the optimal number of clusters for the different preprocessing pipelines as well. Results of both analyses can be found in Figure 4.4, where we show the results for an analysis using Ward linkage and Euclidean distance. What the graphs tell us is that for both configurations there is an optimum in Dunn index and Davies-Bouldin index at $k = 2$ when SVD preprocessing is used. Calculating the Davies-Bouldin index

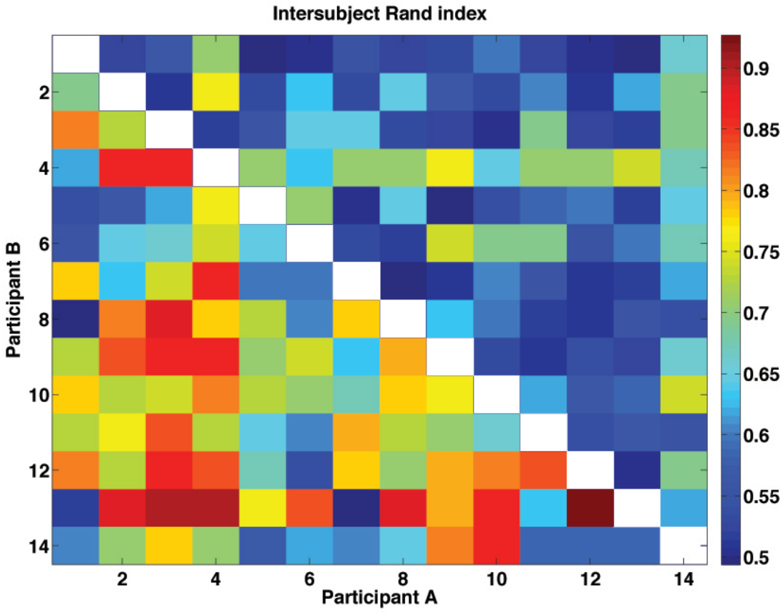


Figure 4.2: Intersubject Rand index, SVD preprocessed patterns clustered with L_2 distance Ward clustering.

on the Brodmann preprocessed data yields optima at $k = 2$ as well, but the Dunn index for these clusterings seems to increase monotonously. This could be due to the lowdimensional nature of the Dunn index, as explained in Chapter 3.

4.2 *k*-means Clustering

k-means clustering and the consecutive validity analysis was performed on the data. Only L_1 and L_2 distance were used for *k*-means clustering. As for the linkage clustering algorithms, it is interesting to look at the cluster size ratios when $k = 2$. These are $\approx 42 : 38 \pm 11$ and $\approx 41 : 39 \pm 12.5$ respectively for Brodmann preprocessing and SVD preprocessing. The external validity relative to the FM-VM labeling for all 28 datasets is given in Figure 4.2. Validity measures for each subject were calculated as the average over 100 *k*-means runs. As we can see here, the external validity values are somewhat below chance level. Of interest is also the similarity between clusterings for different subjects. These values are given for both L_1 and L_2 distance in Figure 4.2. Here, we see that there is very little similarity between clusterings of different clusterings when $k = 2$. Using other values for k yields similar results, with the similarity values centered around 0.5 The fact that the cluster size ratio is more symmetric here than with the linkage clustering algorithms, and the observation that here the intersubject similarity is very low, supports the idea that the high intersubject similarity values for the linkage clusterings was a mere artifact of the skewed cluster size ratios. The internal validity for different values of k showed a preference for $k = 2$, similar to the preference of the linkage clustering.

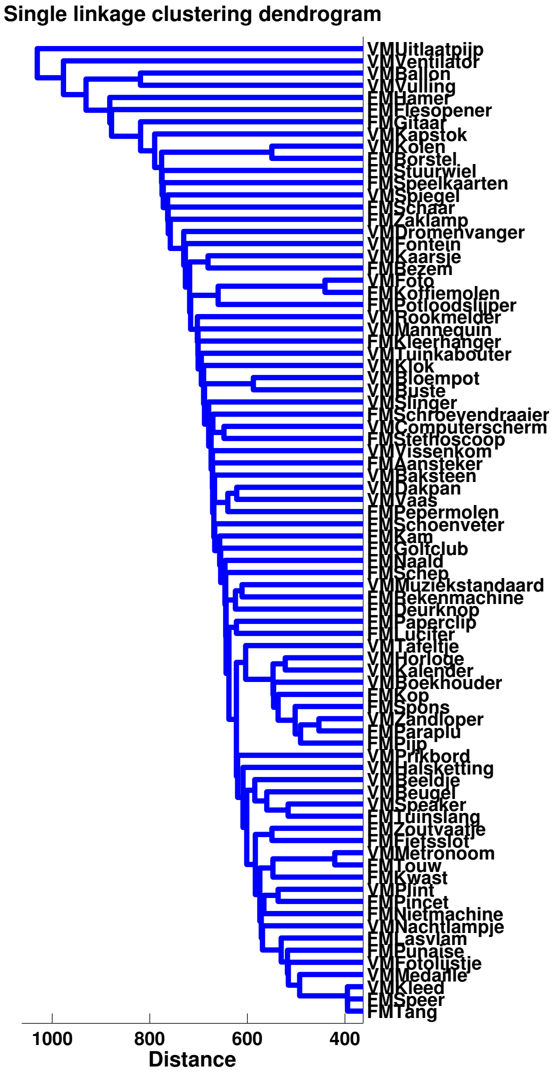
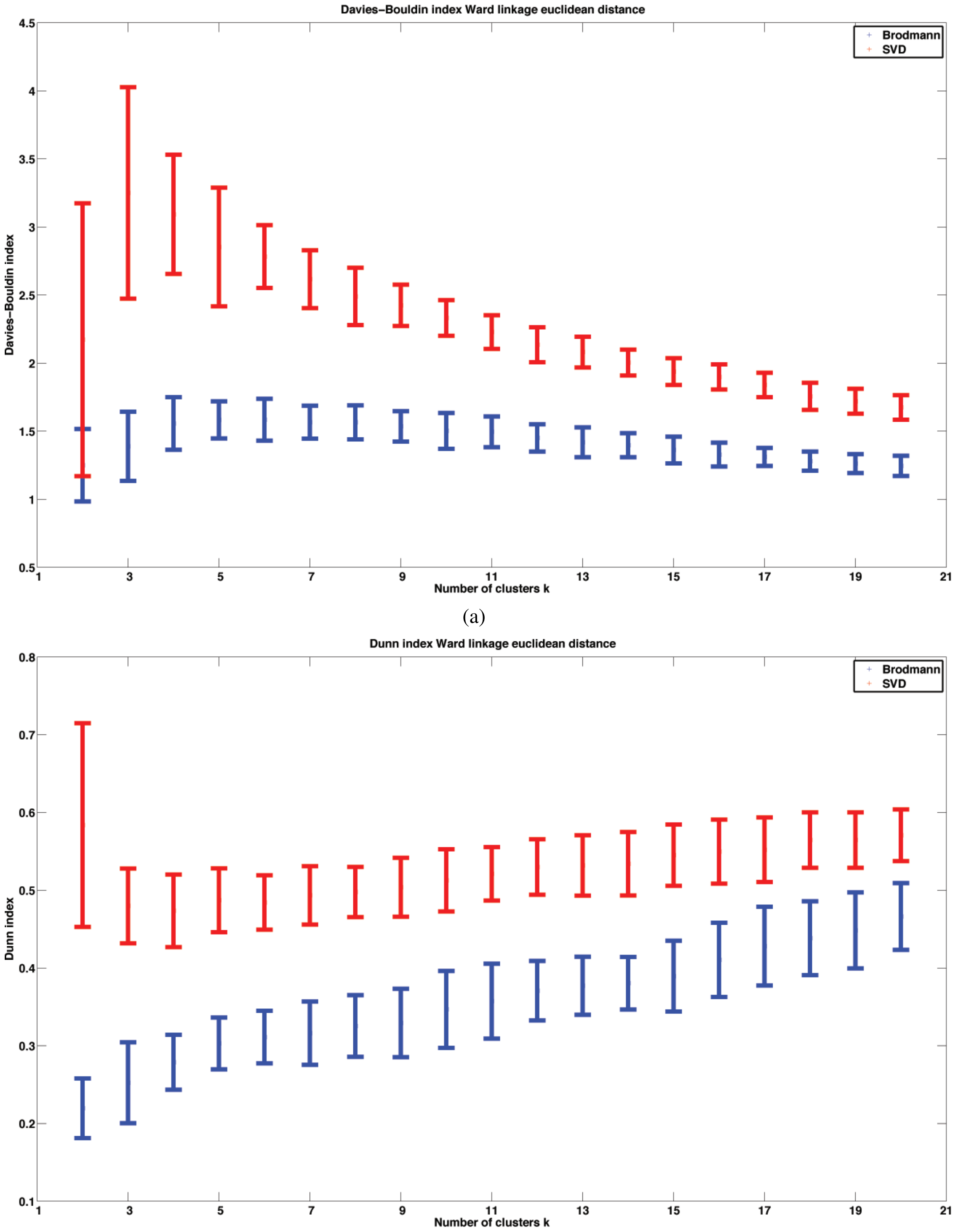


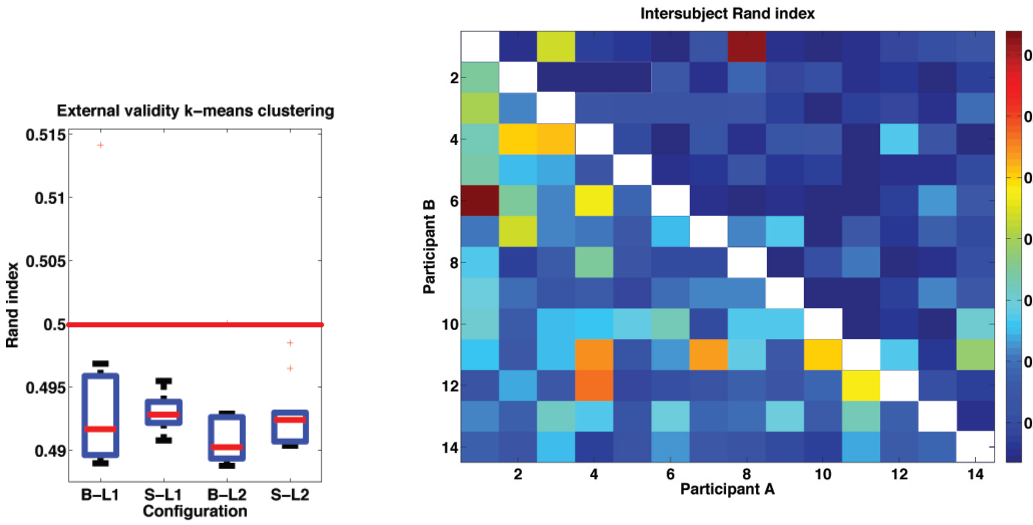
Figure 4.3: Typical single link clustering dendrogram of SVD preprocessed pattern set.



(a)

(b)

Figure 4.4: Interval validity of Ward linkage using Euclidean distance. Validity is measured as Davies-Bouldin index (a) or Dunn index (b).



(a) (b)

Figure 4.5: External validity of *k*-means clustering configurations.

4.3 Genetic Algorithm

We found that the genetic clustering algorithm yields results that are very similar to that of the *k*-means algorithm. On a population of 100 individuals or possible clustering, the genetic algorithm was run for 40 generations on each data set. We found that the external validity of the found clusterings was, as was the case with the linkage and *k*-means clusterings, centered around chance level. Taking a closer look at another performance measure for the genetic algorithm, the sum of squared distances within the found clusters, we see that these are approximately equal to those of the *k*-means algorithm for each data set ($p > 0.99$) in a one-factor ANOVA). When competing against an individual *k*-means run, the genetic algorithm seems to get some optima that the *k*-means algorithm misses, but when running against a batch it always yields the same dispersions. Performing the between cluster external validity analysis with the genetic algorithm yields similarity values that are equal to that of the *k*-means algorithm. All in all, the genetic algorithm seems to yield the same results in this approach. As was mentioned in Chapter 3, the genetic algorithm approach can be of use when there are multiple local optima in the fitness landscape and we do not want to be restricted to the original location of the centroids in the *k*-means algorithm. In such a scenario, using a genetic algorithm could be of use, since its individuals cover a large area of the fitness landscape. However, here it seems like the fitness landscape of this clustering problem is quite simple and that running a batch of *k*-means algorithms yields the same results with a lower computational cost.

4.4 Spectral Clustering

Spectral clustering was performed on the feature sets, using standard affinity matrices based on Euclidean distance. We found that the external validity was centered around 0.5 with a very small variance. The cluster size ratio in these clusterings was similar to that of the *k*-means

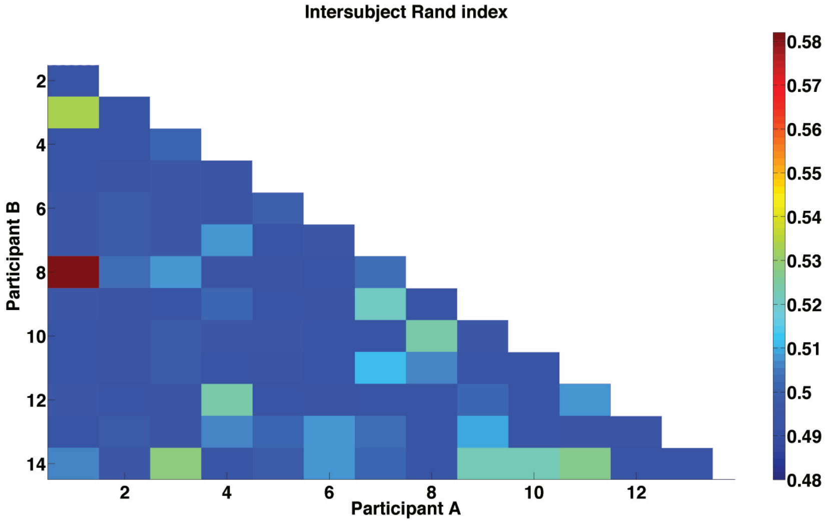


Figure 4.6: Intersubject validity for spectral clustering.

algorithm, so we can say that the external validity is an estimate of the similarity between the found clustering and the FM-VM labeling. Rand similarity between data sets is shown in Figure 4.6. The corresponding landscape is very flat, with outliers only reaching a similarity value of. Comparing Davies-Bouldin indices for different values of k gives an estimate of the number of clusters in the data set. Here we see an optimum at $k = 2$ (Figure 4.7). This supports the results we got from the $k-$means internal validity measures.

4.5 Group analysis

There are multiple ways to combine the data sets of all the subjects into one pattern set. One can either combine the data before preprocessing and then extract patterns from the preprocessed data, extract new patterns from the different pattern sets of all the subjects or concatenate all pattern sets to create one very large pattern set. Here, we choose to use the second method. For every stimulus (excluding the ones on which one or more subjects failed), we have 14 patterns. In total, there are 73 stimuli on which all subjects responded correctly. Using an independent component analysis, we can extract the main components from these patterns and perform clustering on the principal components of the patterns. The question is whether one component suffices or a combination of components is better. Furthermore, if one component alone optimizes an internal validity measure, which component is it? A third question is whether Brodmann preprocessing or SVD preprocessing better optimizes the internal validity. To allow for this question to be answered, we normalized the pattern sets, thereby normalizing the distances between patterns in both types of pattern sets.The clustering method we picked for this group analysis is spectral clustering. If there is anything to be found in the data, spectral clustering should be capable of doing so. As a measure of internal validity, we take the Davies index, which in individual analysis has shown to be a reliable estimate of the number of clusters.

Figure 4.5 shows the Davies index for different component settings. We have either used

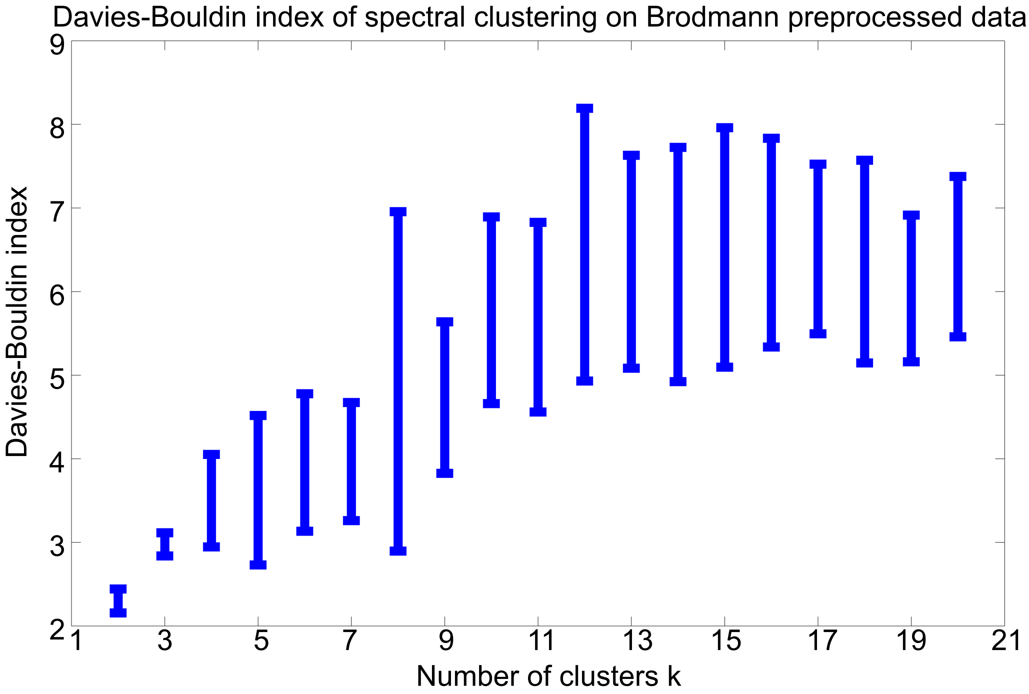


Figure 4.7: Davies-Bouldin index of spectral clustering using different values of *k*. Clustering performed on Brodmann's area preprocessed data.

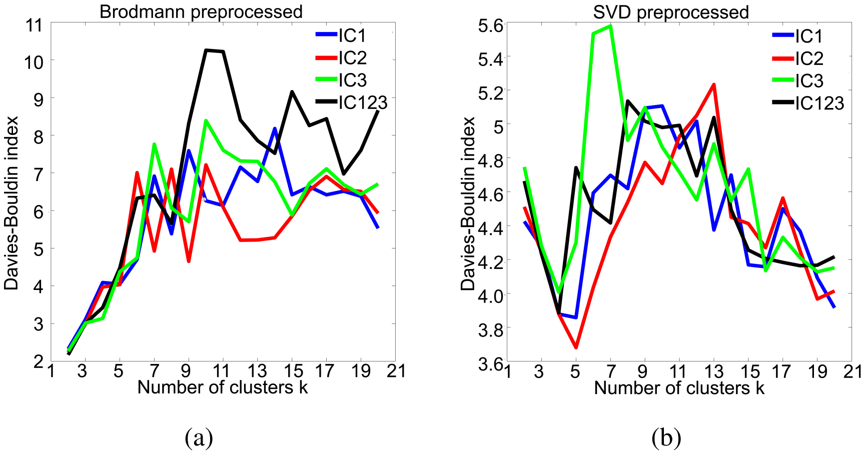


Figure 4.8: Group analysis of data for Brodmann preprocessing (a) and SVD preprocessing (b) using the first, second or third component of ICA on patterns or combination of these three.

Group 1	Group 2	Group 3	Group 4	Group 5
Rookmelder	Spiegel	Mannequin	Slinger	Horloge
Tuinkabouter	Tafeltje	Nachtlampje	Muziekstandaard	Zoutvaatje
Klok	Prikbord	Kolen	Medaille	Punaise
Halsketting	Kapstok	Kalender	Kleed	Pepermolen
Dakpan	Fotolijstje	Kaarsje	Dromenvanger	Paraplu
Bloempot	Fontein	Foto	Computerscherm	Kam
Beeldje	Tuinslang	Ballon	Beugel	Uitlaatpijp
Baksteen	Stuurwiel	Zaklamp	Touw	Deurknop
Schoenveter	Speer	Tang	Spons	Borstel
Schep	Schroevendraaier	Speelkaarten	Paperclip	Vaas
Rekenmachine	Schaar	Pincet	Kwast	Zandloper
Potloodslijper	Naald	Golfclub	Hamer	Ventilator
Pijp	Gitaar	Bezem	Fietsslot	
Nietmachine	Flesopener		Vissenkom	
Lucifer			Aansteker	
Kop			Speaker	
Vulling				
Koffiemolen				

Table 4.1: Optimal spectral clustering for group analysis.

only the first, second or third component or taken the average value of these three components. We see that using different components has little effect on the obtained validity values. In Figure 4.8(a) using the third independent component seems to give an increased similarity between clusters and in Figure 4.8(b) this is true for the combination of components, but all component options seem to follow the same trends in both figures. A large difference between the two figures is that Brodmann preprocessing seems to put the optimal number of clusters at 2, while SVD preprocessing puts k at 4 or 5, depending on the component. It is debatable which one of these two estimations is correct, but the fact that the Davies-Bouldin index values are much lower for the SVD preprocessed data set while using normalized data suggests that SVD clusterings are more reliable. This speaks in favor of the SVD estimation of k.

4.6 Qualitative analysis

A last interesting way of analyzing the clustering results is to look at some of the found clustering and explore them by eye to see if there are any sensible groups of nouns in them. Looking at all the different clusterings obtained for every data set would take an inappropriate amount of time, therefore we decided to look only at one of the clusterings obtained from the spectral clustering approach. This is the clustering obtained by by using group analysis with SVD preprocessing (Table 4.1). With no clear formalism at hand for this analysis, we are restricted to giving our own interpretation, which might be very subjective.

In the clustering we see five groups. At first sight, there are no clear semantic similarities between nouns in clusters. All clusters seem to have an approximately equal ratio of FM and

VM words. What's interesting to note is that in the first cluster there are some stone or baked objects like 'baksteen' (brick), 'dakpan' (roof tile), 'beeldje' (figurine), 'bloempot' (flowerpot), 'kop' (cup) and 'tuinkabouter' (garden gnome). In the second group of nouns, there seem to be a lot of sharp objects. Both a 'speer' (spear) and a 'schroevendraaier' (screw driver) as a 'naald' (needle) and a 'schaar' (scissors) can be used to pierce objects. In the third group, there are three nouns denoting objects that give light: a 'nachtlampje' (nightlight), a 'zaklamp' (flashlight) and a 'kaarsje' (candle). The fourth and fifth group do not seem to contain these kind of semantic similarities between nouns. Though it is notable that 'horloge' (wristwatch) and 'zandloper' (hourglass) are both in group five, one might wonder why they are not grouped with 'klok' (clock), which clearly also is involved with time measurements. What is interesting, is the fact that 'zoutvaatje' (saltshaker) and 'pepermolen' (pepper mill) are both in group five. When looking at the broader picture, one would expect words like 'fontein' (fountain) to be grouped with the stone words in group 1.

Though it is tempting to draw conclusions from this, it should be noted that is actually quite hard to assess the reliability of this clustering. The internal validity was high, but this could also be the case for a random data set. Were the sample nouns separated randomly into five clusters, one could also identify some similarity between nouns within groups.

Chapter 5

Conclusion and discussion

In the present study, an attempt was made to infer a categorization of neural activation patterns elicited by the presentation of nouns. The problem of obtaining a categorization of stimuli in a single-trial, bottom up way was approached with a two-stage model of whole-brain neural activation pattern extraction and unsupervised clustering methods. In a first step, neural activation patterns were extracted from fMRI data obtained from an event-related experiment design (8) where a subject was instructed to identify nonwords by pushing a nonword. Using a conventional statistical parametric mapping approach, two distinct classes of words in the stimulus set were found to elicit different neural activation patterns. Words denoting functionally manipulable (FM) objects elicited stronger activations in certain parts of the motor cortex than words denoting volumetrically manipulable (VM) objects. The hypothesis was that our unsupervised clustering approach could at the very least identify these two distinct classes to be present in the fMRI data.

Relying on both functionally (Brodmann's area downsampling) and mathematically (singular vector decomposition) grounded dimensionality reduction methods, dimensionality was reduced in order to prevent overfitting in the clustering step. In accordance with literature in the brain reading field and commonly used methods in general fMRI analysis, deconvolution using a general linear model was applied on the resulting signals. Out of 120 trial activation features per signal, the trials on which the subject correctly identified a word as being a word were selected. Trials where a nonword was presented or where the subject incorrectly identified a word as a nonword and baseline trials were excluded from further analysis. For each of 77 to 80 correctly identified words, the activation levels over the signals were used as features to form a set of patterns. As a result of this first step, 28 pattern sets were obtained, one per dimensionality reduction method per subject.

In the second stage, the resulting pattern sets were analyzed using a variety of unsupervised clustering algorithms. Though the specific approach of these algorithms differs, they all try to match two criteria. First, the within cluster dispersion should be minimized. Second, the similarity or distance between clusters should be maximal. The result of this is a clustering with *k* groups whose member patterns have a certain similarity while they differ from patterns in other groups. Three distance metrics were used in calculating similarity between sample patterns. Using a Chebychev distance metric, only the most discriminating features were rewarded. On a City Block and Euclidean metric, all features were taken into account.

In a series of analyses using different configurations of both distance metrics and clustering methods, we found no support for the hypothesis that the FM/VM class distinction was identifiable using an unsupervised clustering approach. The similarity between different clustering was measured using the Rand index, which takes into account the number of items that pair or not pair in both clusterings (41). We found that the similarity between the found clusterings and the FM/VM labeling of the patterns was around chance level for all clustering configurations. This means that the clusterings have as much as in common with the FM/VM labeling as would a clustering with 2 approximately equal groups consisting of randomly drawn patterns with a probability that is the same for FM and VM patterns. This implies that the FM/VM class distinction in the stimulus set is not propagated to the eventual categorization'of neural activation patterns. As in the second stage of our approach similarities between patterns should be identified, and because we have used intrinsically different clustering algorithms, this leads to the assumption that the distinction is lost in the pattern extraction stage. An obvious possibility is that the pattern extraction stage is too rigorous in its dimensionality reduction and activation estimation, thereby killing off any information that could be present in the fMRI data.

To test whether the pattern extraction stage removes the signal from the data, or that a class distinction stronger than the FM/VM distinction overrules the FM/VM distinction, we compared the clusterings for different subjects with each other. If there was high similarity between clusterings, this could imply that the stimulus set contained a categorization that was stronger than that between FM and VM words. It was found that for most of the applied clustering algorithms, these between subject clustering similarities were around chance level. For the linkage algorithm though, these similarities could reach values of up to 90%. A closer look at the clustering compared showed this to be merely an artifact of the similarity index. The Rand index performs best when all k groups in both clustering contain an approximately symmetric amount of samples. The problem with some of the linkage clustering configurations was that they yielded two clusters where one cluster contained only one sample and the other cluster contained the rest of the samples. This cluster size distribution, which in fact is specific for random data sets, causes the Rand index to give skewed estimates of the similarity between clusterings.

With the clusterings showing no clear FM/VM distinction, nor any other consistent stimulus grouping over subjects, the question remains whether there is a preferred number of clusters k in the pattern set. If there is such a number, this might suggest that there is indeed some structure in the data. For a random set of n patterns, the optimal number of clusters k would equal n, since such a clustering matches the clustering criteria best. Instead, we found $k = 2$ to be the number of clusters that optimizes these criteria for most of the clustering configurations. Only on a group approach, where independent components of the set of patterns for each stimulus were used as samples, did we find an optimal number of clusters $k = 5$. On a normalized pattern set, we found that this was true only for the SVD preprocessed pattern set and that the criteria were met better than for the Brodmann's area preprocessed pattern set, where the optimal numbers of clusters was 2. A qualitative analysis of the optimal group clustering with $k = 5$ showed some semantic similarities between nouns that are in the same group. It is however highly debatable whether these similarities would not have occurred when five random groups of stimuli were drawn. The existence of a clear formalism to analyze the semantic value of such groupings could help assess the results.

One could say that our endeavors have been futile and that the rather dissatisfying results are caused by our choices to do a whole-brain area analysis and use only the trials corresponding to existing words. No clear class distinctions have been found, not even for the FM/VM classes that are known to be present in the stimulus set. One could also suggest to narrow the analysis down to just those brain areas in the motor cortex that are known to contain the FM/VM distinction. As a matter of fact, to see whether this would enhance the results, we did an investigation in which we fed voxel locations to the two-stage pipeline and obtained features only from those areas in the motor cortex which were known to show a preference for FM-words. Though this clearly deviates from our main research question, where we try to infer a categorization from the whole brain and do not any assumptions about the nature of the stimuli whatsoever, this was useful to further narrow down where the exact location of the problem in the processing pipeline. The results of this investigation were very similar to that of the whole-brain analysis. Decreasing the factor in the dimensionality reduction step thus does not seem to make a very large difference. On the assumption that the word-nonword distinction was very strong, considering the activation of motor cortex areas as subject pushed a button on nonword presentations, we performed a last analysis to try and retrieve this distinction between trials. Estimation of the similarity between the clustering results and the word/nonword labeling using the Rand index showed no results deviating from chance level.

Considering that the clustering step in our two-stage approach is fairly trivial, we can only draw the conclusion that it is very hard if not impossible to obtain activation patterns from an event-related fMRI experiment with short inter trial intervals on a single-trial basis. As stated in the introduction of this thesis, the problem at hand was in fact very hard. Where some studies advice to use an inter trial interval of at least 20 seconds to allow for the hemodynamic response to return to baseline (9), the data we used was recorded in an experiment with an 8 second inter trial interval. Furthermore, we were trying to obtain activations based on only one presentation per stimulus. This is in fact very fragile, since fMRI has a very low signal to noise ratio and there are numerous ways in which a hemodynamic response could be corrupted by other processes in the brain or artifacts of the measurement methods. In fMRI studies, it is therefore common to present a stimulus multiple times to a subject in order to increase the reliability of the activation estimation. In fact, one might argue that the strength of the SPM GLM analysis lies in the fact that it extracts a stimulus specific response by regressing for all stimulus presentations at once. This makes the estimate much more reliable since it does not add trial-specific noise to a trial. Furthermore, using a shorter interval between scans (TD), could increase the sampling rate of the BOLD signal and the estimation of activation.

Another way to increase the reliability of the activation estimates is to add longer baseline conditions to the experiment design. In such a baseline condition, the hemodynamic response can return to baseline, giving an estimate of the brains rest state. This would make the calculation of a percent signal change (PSC) possible, which could be an even better estimate of the neural activation than just the β-values in a GLM-analysis. The percent signal change is the percentage of signal deviation from the rest condition. One could take this value for multiple timepoints t after stimulus presentation and append these in a feature vector, instead of using only one β-value per stimulus. Another possible analysis method would be finite impulse response analysis (FIR), which is very similar to the SPM approach, but uses one regressor for each of an arbitrary number of time bins after stimulus presentation. Like the PSC analysis, this yields a feature vector per stimulus per signal instead of just one feature. Both SPC and FIR analysis can only be used when two other criteria are met: a stimulus has to be presented multiple times and the inter trial interval should be such that there is a sufficient amount of time for the hemodynamic response to relax.

We see here that the number of presentations per stimulus, the time between presentations and the length of baseline conditions can enhance the quality of the neural activation estimates.

To further increase the reliability of the estimates, another lesson with respect to the nature of stimuli can be drawn from brain reading studies. In basically all brain reading studies (e.g. (2) (5) (3)), subject are presented with pictures of objects instead of words denoting objects. In accordance with this, one can narrow down the region of analysis to parts of the cortex that are related to visual processes, as is done in typical brain reading studies. Note however that this is quite different from the goal of this thesis, where we do not take into account any information about the nature of the stimuli. A more controlled set of stimuli could also increase categorization performance. For example, previous studies have used two clearly distinct groups of figures representing dwellings or tools. This leaves little space for ambiguity, while the stimulus set used in this study offers a lot of interpretations.

The lessons drawn from this study can be summarized in a set of advices for a possible follow-up study:

1. Present stimuli multiple times in separate trials, get an activation estimate based on the combination of these trials.

2. Use long inter trial intervals, preferably around 20 seconds.

3. Insert long baseline conditions at the start and at the end of the experiment.

4. Use PSC estimation and FIR analysis to get more informative activation estimates.

5. Use pictures as stimuli instead of words.

6. Prevent ambiguity about categorization, make classes among stimuli very clear.

7. Optionally narrow down analysis to particular brain areas.

We believe that, when these criteria are met, reliable activation estimates corresponding to different stimuli can be retrieved and a categorization of objects in the brain can be derived. Maybe one day we can really 'read' a mind. For now, we will have to do with these lessons.

Bibliography

[1] J. Ward, *The Student's Guide to Cognitive Neuroscience*, 1st ed. London, UK: Psychology Press, 2006.

[2] T. M. Mitchell, S. V. Shinkareva, A. Carlson, K.-M. Chang, V. L. Malave, R. A. Mason, and M. A. Just, "Predicting human brain activity associated with the meanings of nouns," *Science*, vol. 320, no. 5880, pp. 1191–1195, May 2008.

[3] D. D. Cox and R. L. Savoy, "Functional magnetic resonance imaging (fMRI) brain reading: detecting and classifying distributed patterns of fMRI activity in human visual cortex," *NeuroImage*, vol. 19, no. 2, pp. 261–270, June 2003.

[4] J. V. Haxby, I. M. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, and P. Pietrini, "Distributed and overlapping representations of faces and objects in ventral temporal cortex," *Science*, vol. 293, no. 5539, pp. 2425–2430, September 2001.

[5] S. V. Shinkareva, R. A. Mason, V. L. Malave, W. Wang, T. M. Mitchell, and M. A. Just, "Using fMRI brain activation to identify cognitive states associated with perception of tools and dwellings," *PLoS ONE*, vol. 3, no. 1, p. 1394, 01 2008.

[6] C. Davatzikos, K. Ruparel, Y. Fan, D. Shen, M. Acharyya, J. Loughead, R. Gur, and D. Langleben, "Classifying spatial patterns of brain activity with machine learning methods: Application to lie detection," *NeuroImage*, vol. 28, no. 3, pp. 663 – 668, 2005.

[7] S. Huettel, A. W. Song, and G. McCarthy, *Functional Magnetic Resonance Imaging*. Sunderland, Massachusetss, USA: Sinauer Associates, Inc, 2004.

[8] S.-A. A. Rueschemeyer, D. van Rooij, O. Lindemann, R. Willems, and H. Bekkering, "The function of words: Distinct neural correlates for words denoting differently manipulable objects." *Journal of cognitive neuroscience*, July 2009.

[9] C. Windischberger, C. Lamm, H. Bauer, and E. Moser, "Consistency of inter-trial activation using single-trial fMRI: assessment of regional differences," *Cognitive Brain Research*, vol. 13, no. 1, pp. 129 – 138, 2002.

[10] R. Cabeza and L. Nyberg, "Imaging cognition II: An empirical review of 275 PET and fMRI studies," *J. Cognitive Neuroscience*, vol. 12, no. 1, pp. 1–47, 2000.

[11] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Computer Survey*, vol. 31, no. 3, pp. 264–323, September 1999.

[12] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, “On clustering validation techniques,” *Journal of Intelligent Information Systems*, vol. 17, pp. 107–145, 2001.

[13] R. C. Dubes, “How many clusters are best? - an experiment,” *Pattern Recognition*, vol. 20, no. 6, pp. 645–663, 1987.

[14] K. N. Kay, T. Naselaris, R. J. Prenger, and J. L. Gallant, “Identifying natural images from human brain activity,” *Nature*, vol. 452, no. 7185, pp. 352–355, March 2008.

[15] F. Pereira, T. Mitchell, and M. Botvinick, “Machine learning classifiers and fMRI: a tutorial overview.” *NeuroImage*, vol. 45, no. 1, March 2009.

[16] V. Della-Maggiore, W. Chau, P. R. Peres-Neto, and A. R. Mcintosh, “An empirical comparison of SPM preprocessing parameters to the analysis of fMRI data,” *NeuroImage*, vol. 17, no. 1, pp. 19–28, September 2002.

[17] M. Dubin, “Brodmann areas in the human brain with an emphasis on vision and language,” 2009, `http://spot.colorado.edu/ dubin/talks/brodmann/`.

[18] Wikipedia, “Brodmann area 4 — Wikipedia, the free encyclopedia,” 2010.

[19] J. A. Maldjian, P. J. Laurienti, R. A. Kraft, and J. H. Burdette, “An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets,” *NeuroImage*, vol. 19, no. 3, pp. 1233 – 1239, 2003.

[20] J. A. Maldjian, P. J. Laurienti, and J. H. Burdette, “Precentral gyrus discrepancy in electronic versions of the Talairach atlas,” *NeuroImage*, vol. 21, no. 1, pp. 450 – 455, 2004.

[21] C. H. Moritz, B. P. Rogers, and M. E. Meyer, “Power spectrum ranked independent component analysis of a periodic fMRI complex motor paradigm,” *Hum. Brain Mapp*, vol. 18, pp. 111–122, 2003.

[22] M. J. McKeown, “Detection of consistently task-related activations in fMRI data with hybrid independent component analysis,” *NeuroImage*, vol. 11, pp. 24–35, 2000.

[23] D. Hu, L. Yan, Y. Liu, Z. Zhou, K. J. Friston, C. Tan, and D. Wu, “Unified SPM-ICA for fMRI analysis,” *NeuroImage*, vol. 25, no. 3, pp. 746–755, April 2005.

[24] J. Stone, J. Porril, C. Buchel, and K. Friston, “Spatial, temporal, and spatiotemporal independent component analysis of fMRI data,” in *Spatial temporal modelling and its applications*, 1999.

[25] V. D. Calhoun, T. Adali, L. K. Hansen, J. Larsen, and J. J. Pekar, “ICA of functional MRI data: An overview,” in *in Proceedings of the International Workshop on Independent Component Analysis and Blind Signal Separation*, 2003, pp. 281–288.

[26] A. L. Vazquez and D. C. Noll, “Nonlinear aspects of the BOLD response in functional MRI.” *Neuroimage*, vol. 7, no. 2, pp. 108–118, February 1998.

[27] C. Buchel, A. Holmes, G. Rees, and K. Friston, “Characterizing stimulus-response functions using nonlinear regressors in parametric fMRI experiments,” *NeuroImage*, vol. 8, pp. 140–148, 1998.

[28] V. D. Calhoun, M. C. Stevens, G. D. Pearlson, and K. A. Kiehl, “fMRI analysis with the general linear model: removal of latency-induced amplitude bias by incorporation of hemodynamic derivative terms,” *NeuroImage*, vol. 22, no. 1, pp. 252 – 257, 2004.

[29] J. Hopfinger, C. Buchel, A. Holmes, and K. Friston, “A study of analysis parameters that influence the sensitivity of event-related fMRI analyses,” *NeuroImage*, vol. 11, pp. 326–333, 2000.

[30] M. A. Babyak, “What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models,” *Psychosomatic Medicine*, vol. 66, no. 3, pp. 411–421, May 2004.

[31] J. H. Ward, “Hierarchical grouping to optimize an objective function,” *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963.

[32] V. V. Raghavan and K. Birchard, “A clustering strategy based on a formalism of the reproductive process in natural systems,” *SIGIR Forum*, vol. 14, no. 2, pp. 10–22, 1979.

[33] U. Maulik, S. Bandyopadhyay, and S. B, “Genetic algorithm-based clustering technique,” *Pattern Recognition*, vol. 33, pp. 1455–1465, 2000.

[34] U. von Luxburg, “A tutorial on spectral clustering,” Max Planck Institute for Biological Cybernetics, Tech. Rep. 149, August 2006.

[35] A. Y. Ng, M. I. Jordan, and Y. Weiss, “On spectral clustering: Analysis and an algorithm,” in *Advances in Neural Information Processing Systems 14*. MIT Press, 2001, pp. 849–856.

[36] R. C. Dubes and A. K. Jain, “Validity studies in clustering methodologies.” *Pattern Recognition*, vol. 11, no. 4, pp. 235–254, 1979.

[37] J. C. Dunn, “Well separated clusters and optimal fuzzy-partitions,” *Journal of Cybernetics*, vol. 4, pp. 95–104, 1974.

[38] C. Legány, S. Juhász, and A. Babos, “Cluster validity measurement techniques,” in *AIKED’06: Proceedings of the 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases*. Stevens Point, Wisconsin, USA: World Scientific and Engineering Academy and Society (WSEAS), 2006, pp. 388–393.

[39] D. L. Davies and D. W. Bouldin, “A cluster separation measure,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. PAMI-1, no. 2, pp. 224–227, January 1979.

[40] R. A. Fisher, “The use of multiple measurements in taxonomic problems,” *Annals Eugenics*, vol. 7, pp. 179–188, 1936.

[41] W. M. Rand, “Objective criteria for the evaluation of clustering methods,” *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971.