Wavelets and Sparse Methods for Image Reconstruction and Classification in Neuroimaging

Michal Piotr Romaniuk

A dissertation submitted in partial fulfilment of the requirements for the degree of **Doctor of Philosphy** of **Imperial College London**

Department of Computing, Imperial College London

November 2016

Declaration of originality

I declare that the work presented in this dissertation is my own, unless specifically acknowledged.

Michal Piotr Romaniuk

Copyright declaration

Abstract

This dissertation contributes to neuroimaging literature in the fields of compressed sensing magnetic resonance imaging (CS-MRI) and image-based detection of Alzheimer's disease (AD). It consists of three main contributions, based on wavelets and sparse methods.

The first contribution is a method for wavelet packet basis optimisation for sparse approximation and compressed sensing reconstruction of magnetic resonance (MR) images of the brain. The proposed method is based on the basis search algorithm developed by Coifman and Wickerhauser, with a cost function designed specifically for compressed sensing. It is tested on MR images available from the Alzheimer's Disease Neuroimaging Initiative (ADNI).

The second contribution consists of evaluating and comparing several sparse classification methods in an application to detection of AD based on positron emission tomography (PET) images of the brain. This comparison includes univariate feature selection, feature clustering and classifiers that automatically select a small subset of features due to their mathematical or algorithmic construction. The evaluation is based on PET images available from ADNI.

The third contribution is proposing an extension of wavelet-based scattering networks (originally proposed by Mallat and Bruna) to three-dimensional tomographic images. The proposed extension is evaluated as a feature representation in an application to detection of AD based on MR images available from ADNI.

There are several possible extensions of the work presented in this dissertation. The wavelet packet basis search method proposed in the first contribution can be improved to take into account the coherence between the sparse approximation basis and the sensing basis. The evaluation presented in the second contribution can be extended with additional algorithms to make it more comprehensive. The three-dimensional scattering networks that are the core part of the third contribution can be combined with other machine learning methods, such as manifold learning or deep convolutional neural networks.

As a whole, the methods proposed in this dissertation contribute to the work towards efficient screening for Alzheimer's disease, by making MRI scans of the brain faster and helping to automate image analysis for AD detection.

Acknowledgements

First and foremost, I would like to thank my supervisor Prof. Daniel Rueckert for his support and patience throughout my PhD studies. I would also like to thank my second supervisor Prof. Jo Hajnal for many useful and interesting discussions.

I would like to thank Dr Katherine Gray and Dr Robin Wolz for providing processed ADNI data that was used throughout this thesis.

I would like to acknowledge Dr Anil Rao and Dr Kanwal Bhatia for many useful discussions.

I owe very special thanks to Prof. Marek Sergot and Dr Amani El-Kholy. Their dedication and understanding was essential to making the completion of this dissertation possible.

I would like to acknowledge EPSRC for funding my research and ADNI for providing data that was used throughout this dissertation.

Last but definitely not least, I am grateful to my friends and especially my family without whose support getting through my PhD studies would not have been possible.

Contents

CHAPTER 1

Introduction

Imaging technology has had a remarkable impact on medical research and clinical practice in recent decades. Modern imaging tools, such as ultrasonography, magnetic resonance imaging (MRI), computed tomography (CT) and positron emission tomography (PET) provide the means to produce detailed maps of the internal structure and function of living organisms. This wealth of information allows researchers and practitioners to diagnose diseases more accurately and monitor their progression in unprecedented detail while avoiding many invasive tests.

The development of medical imaging has also prompted interest of computer vision researchers in designing algorithms that derive qualitative and quantitative metrics from medical images. Within the medical image analysis community, researchers have proposed numerous algorithms to solve problems such as image registration (estimating spatial correspondence between two or more images), segmentation (partitioning an image into anatomical or functional regions) and computer-aided diagnosis. These tools automate tedious manual tasks, making it easier for medical research to be done efficiently and on a large scale.

In its first part, this introductory chapter covers the basic principles of the medical imaging technologies that were used to acquire the images used in this dissertation: magnetic resonance imaging (MRI) and positron emisson tomography (PET). The second part of this chapter covers some basic background on Alzheimer's disease, which is important because much of this dissertation is concerned with applying machine learning algorithms to detection of Alzheimer's disease. Finally, this chapter also enumerates the contributions to the literature made by this dissertation and presents an outline of the remainder of it.

1.1. Medical imaging modalities

Imaging techniques can be categorised into modalities according to the physical process used to produce the images. Modern imaging modalities can produce three-dimensional tomographic images mapping specific physical properties of tissues across space and each modality has its own specific advantages and disadvantages.

Perhaps the most important characteristic describing a modality is whether it uses ionising radiation or not. This is important because ionising radiation poses a health risk. Table 1.1.1 lists several tomographic medical imaging modalities of ionising and non-ionising types.

The remainder of this section presents in some detail the physics of two imaging modalities that were used to acquire the images studied in this dissertation: magnetic resonance imaging and positron emission tomography.

Table 1.1.1: Selected tomographic medical imaging modalities.

Non-ionising	Ionising
Magnetic resonance imaging (MRI)	Computed tomography (CT)
Ultrasond imaging	Positron emission tomography (PET)
Photoacoustic imaging	Single positron emission computed tomography (SPECT)

1.1.1. Magnetic resonance imaging. Although quantum mechanical in nature, the physics behind magnetic resonance imaging can be explained in a simplified way using classical principles [**109**]. The following introduction is based on[**179**].

MRI relies on a phenomenon called nuclear magnetic resonance. Atomic nuclei with an odd number of protons or neutrons have a property called spin, which can be thought of as associating a tiny magnet with each nucleus. MRI typically focuses on the ^{1}H (hydrogen-1) nucleus which appears in water and organic molecules.

In the presence of an external magnetic field the spins precess about the axis of this field in analogy to the motion of a gyroscope. The direction of alignment is either

parallel or antiparallel to the direction of field lines, with a small net surplus of spins aligned in parallel, due to the lower energy of this state.

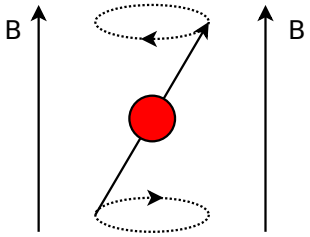


Figure 1.1.1: Precession of a spin can be viewed as similar to the motion of a spinning top.

The frequency of precession is proportional to the strength of the applied magnetic field, according to the Larmor equation:

$$\omega_0 = \gamma B_0$$

where ω_0 is the precession frequency (Larmor frequency), γ is the gyromagnetic ratio and B_0 is the magnetic field. The gyromagnetic ratio of ^{1}H is 42.58 MHz/T, so at magnetic field strengths generated by main magnets of MR scanners (typically 1.5T or 3T, or in some cases 7T) its Larmor frequency is in the radio frequency (RF) range.

If the nuclei are excited with an RF pulse tuned to the Larmor frequency, the precessive motion of their spins tips away from the direction of the magnetic field. The energy absorbed from the pulse is then re-emitted in the form of radio waves as the spins return to their low-energy state where they are aligned with the field. These radio waves are picked up by a receiver coil and recorded for further processing. The process follows an exponential decay curve and is characterised by a time constant referred to as T_1. The variation of T_1 between different types of tissue can be used as a source of image contrast.

The RF pulse not only causes the spins to tip away from the external magnetic field lines, but also synchronises the phases of their precessive motion [**179**]. However,

this synchrony gradually decays as the individual rates of precession are affected by the electromagnetic interactions between the spins. As a result, the signal at the receiver coil also decays. This process follows an exponential decay curve too, with the time constant T_2 associated with it. The differences in T_2 between tissues can be used as an alternative image contrast mechanism.

Spatial encoding in MRI is achieved with additional magnetic field gradients superimposed on the field generated by the main magnet. These additional gradients are generated with gradient coils installed inside the scanner (shown in figure 1.1.2). Three types of spatial encoding techniques are used to produce three-dimensional images: slice encoding, frequency encoding and phase encoding.

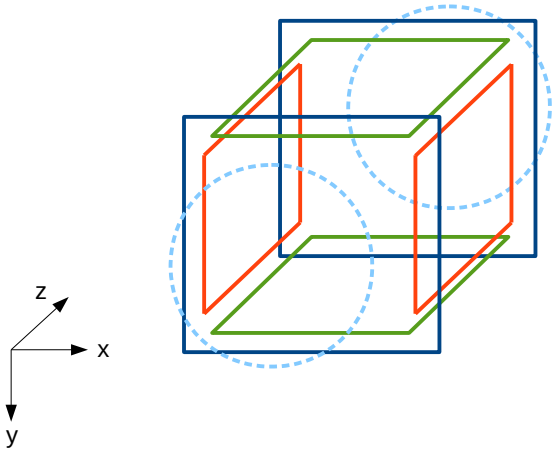


Figure 1.1.2: Gradient coils and the MRI coordinate system (radiological convention). The patient is surrounded by three sets of coils (simplified in this diagram) designed to generate magnetic fields in three directions: x, y and z (red, green and blue, respectively). The ends of the scanner bore are marked with light blue dashed circles.

Slice encoding is achieved by applying a field gradient during the RF excitation pulse. This gradient causes Larmor frequency to vary spatially with field strength, so it is possible to excite a thin slice of the sample by shaping the excitation pulse to contain a narrow range of frequencies.

Spatial encoding in the plane of the slice is achieved by the means of frequency encoding and phase encoding.

Frequency encoding consists of applying a field gradient (orthogonal to the slice selection gradient) during slice readout, resulting in a proportional variation of Larmor frequency across the slice. This means that different temporal frequencies in the signal emitted by the sample represent different locations in space. The signal amplitudes at different temporal frequencies can be recovered from the MR signal with the Fourier transform, giving a one-dimensional distribution of signal in the imaged sample. In other words, the temporal MR signal corresponds to the image in the spatial frequency space (k-space).

Phase encoding relies on applying an additional magnetic field gradient (orthogonal to both the slice selection gradient and frequency encoding gradient) between frequency encoding gradients. The effect of this gradient is to introduce an additional phase shift in the gyration of the spins across the slice, hence the name phase encoding. This is equivalent to starting a frequency encoding readout along the phase encoding direction k_y but interrupting it mid-way through k-space at some k'_y, so that now an ordinary frequency encoding readout along k_x will capture a part of k-space with $k_y = k'_y$. By repeating the process with different k'_y, a complete two-dimensional representation of k-space is acquired and this signal is then processed with a two-dimensional Fourier transform to give the spatial image of the slice.

Sweeping through k-space via frequency encoding and phase encoding requires changing the magnetic field within the scanner, which is done by switching currents in the gradient coils. Stronger gradients with rapid switching can in principle give faster scans but also cause problems with peripheral nerve stimulation [**54, 122**]. This limits the rate at which slices can be acquired, making it difficult to obtain multi-slice images when the patient is moving, *e.g.* in cardiac imaging or fetal imaging [**31, 144**].

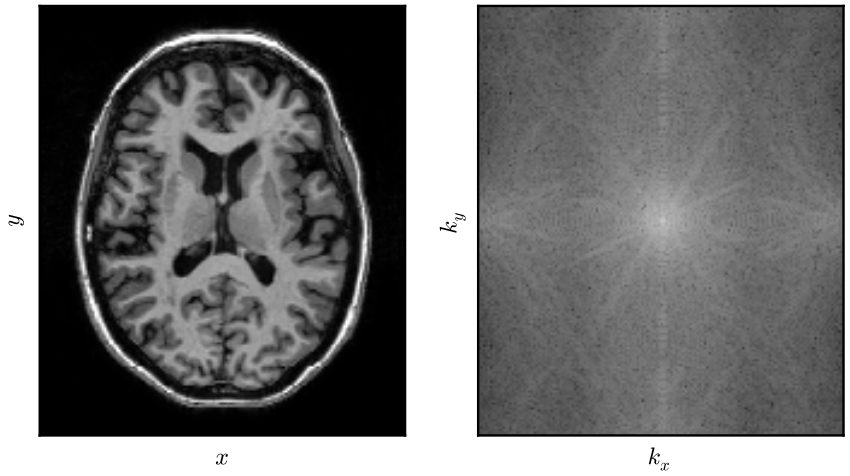


Figure 1.1.3: A brain MRI image (left) and its k-space representation (right).

The T_1 and T_2 properties change from tissue to tissue, allowing different tissues to be distinguished. In addition, MRI can be adapted for imaging of diffusion of water molecules (which contain protons) by diffusion weighted imaging (DWI) [**14**]. Diffusion tensor imaging (DTI) [**160**] is a related MRI technique that measures the diffusion of water in specific directions. Since water tends to diffuse along the fibers, DTI can be used for tracing neural fibers in the brain [**16**]. MRI can also be adapted for blood oxygenation dependent (BOLD) contrast, which enables detection of blood flow, which in turn allows for functional imaging [**178**].

MRI has several advantages as a medical imaging modality. It is non-ionising, which makes it safer than ionising modalities such as CT (which relies on X-rays) and PET (which relies on radioactive tracers). It produces high-resolution images and it offers many useful contrast mechanisms.

The main disadvanage of MRI is the high cost of the scanner and its supporting infrastructure. Additionally, there are dangers associated with imaging patients with metallic implants, which can be subject to large forces and cause damage in the presence of strong magnetic fields, or heat up due to absorbing RF energy.

MRI scans are used in clinical and scientific applications to study nearly all systems in the human body and they stand out particularly in neuroimaging. With clear contrast between grey matter and white matter, structural T1 scans can be used to study diseases affecting brain structure, including neurodegenerative disease such as Alzheimer's disease (AD). Diffusion imaging can be used to study the structural connections within the brain and functional MRI can map the brain's response to sensory stimuli and its functional connectivity.

1.1.2. Positron emission tomography. Positron emission tomography relies on radioactive decay of positron-emitting atomic nuclei to produce images that represent the spatial distribution of these nuclei. When one of these nuclei decays, the emitted positron travels a short distance through the surrounding tissue, losing kinetic energy due to Coulomb scattering [**109**]. Once it slows down, it annihilates with an electron, which produces two gamma ray photons emitted in opposite directions [**109**].

A PET scanner detects gamma ray photons with a ring of gamma ray detectors placed around the patient. The detector ring is equipped with electronics that register events where two detectors receive a gamma ray photon at the same time. When such event occurs, it likely means that the simultaneous detection is due to a positron decaying somewhere on the straight line between those two detectors (called the line of response). These events are counted and their number for each line is recorded. The data is then formatted as a matrix called a sinogram: each row corresponds to a different angle and each column to a different offset from the centre. An image is reconstructed from the sinogram with the filtered back-projection or maximum likelihood expectation maximisation (MLEM) [**109**].

In order to image a patient with PET, positron-emitting radionuclei are embedded in molecules involved in the biological processes under study and then introduced into the patient's body. For example, ^{18}F-fluorodeoxyglucose (FDG) is commonly

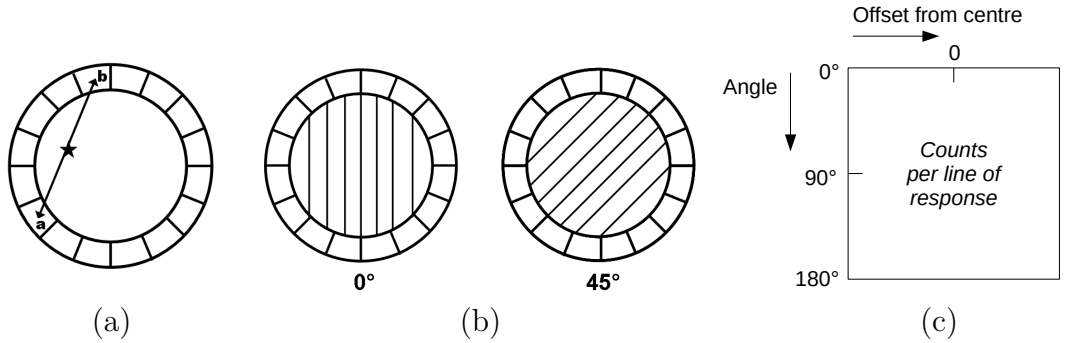


Figure 1.1.4: PET image acquisition. Left to right: (a) line of response due to positron-electron annihilation, (b) parallel lines of response corresponding to the same angle, and (c) sinogram constructed by grouping detection counts by angle and line of response. Images (a) and (b) from [**109**], used with permission; image (c) based on [**109**], with changes.

used in neuroimaging as an indicator of cerebral metabolic rate of glucose (CMRgl) [**109**].

The main advantage of PET as a functional imaging technique is that a radiotracer can be designed to target a specific biochemical process by including a positron-emitting isotope in a molecule involved in that process or a chemically similar molecule. In particular, FDG targets glucose metabolism, which makes it useful in the study of diseases which involve either increased or reduced glucose metabolism. The disadvantages of PET include potential harm due to ionising radiation, low signal to noise ratio (SNR) and high cost.

PET can be used scientifically to study a variety of biochemical processes in the body, depending on the specific radiotracer used. In neuroimaging FDG-PET is used for imaging of Alzheimer's disease patients because reduced metabolism in specific brain regions was found to be an indicator of AD [**155, 195**]. Another type of PET scan useful in AD studies uses Pittsburgh compund B (PiB) as radiotracer, which binds to amyloid plaques associated with AD [**109**].

1.2. Alzheimer's disease

Alzheimer's disease (AD) is the most common type of dementia, accounting for an estimated 60% to 80% of cases [**10**]. Its distinctive characteristic is the presence of amyloid beta (Aβ) plaques outside neurons and protein tau tangles inside neurons [**10**]. The loss of neural cells results in shrinkage of the brain [**10**].

AD may begin up to 20 years before symptoms appear and progress slowly without causing any effect noticeable to the patient or other people [**10**]. First symptoms to be noticed often include short-term memory problems and at this stage a person may be diagnosed with mild cognitive impairment (MCI) [**10**]. Patients diagnosed with MCI have an increased risk of developing AD and other types of dementia, although some of them remain stable or even improve [**10**]. Probable Alzheimer's disease is often diagnosed when further decline in memory and cognitive function starts to impair a person's daily life. In addition to impaired memory, AD symptoms include apathy and depression, problems with language, confusion with time and place, and behavioural changes [**10**]. Patients with advanced AD lose their ability to communicate, fail to recognise family members and become completely reliant on others for the simplest daily activities and eventually bedbound [**10**]. A definite diagnosis of AD requires examination of brain tissue samples [**109**].

1.2.1. Neuroimaging in Alzheimer's disease. Neuroimaging is used extensively in the scientific study of AD. The most established modalities for this application include structural MRI, FDG-PET and fibrillar Aβ PET. In addition, functional MRI, DTI, and several other techniques were also used to learn more about the disease [**216**].

Structural MRI is used to study the brain atrophy observed in AD. The hippocampus and entorhinal cortex of AD patients have reduced volume, gray matter and cortical thickness [**216**]. Many other cerebral regions are also affected [**216**]. Meanwhile, sulcal and ventricular volumes are larger in AD [**216**]. The presence of these changes

in asymptomatic subjects and higher rate of their progression indicate an increased risk of developing MCI or AD [**216**].

FDG-PET is used to map CMRgl across the brain. In AD patients CMRgl is reduced in the posterior cingulate, precuneus and parietotemporal regions [**155**]. More advanced AD also affects CMRgl in the frontal cortex and the rest of the brain [**216**]. These changes correlate with disease severity and they also have predictive value [**216**].

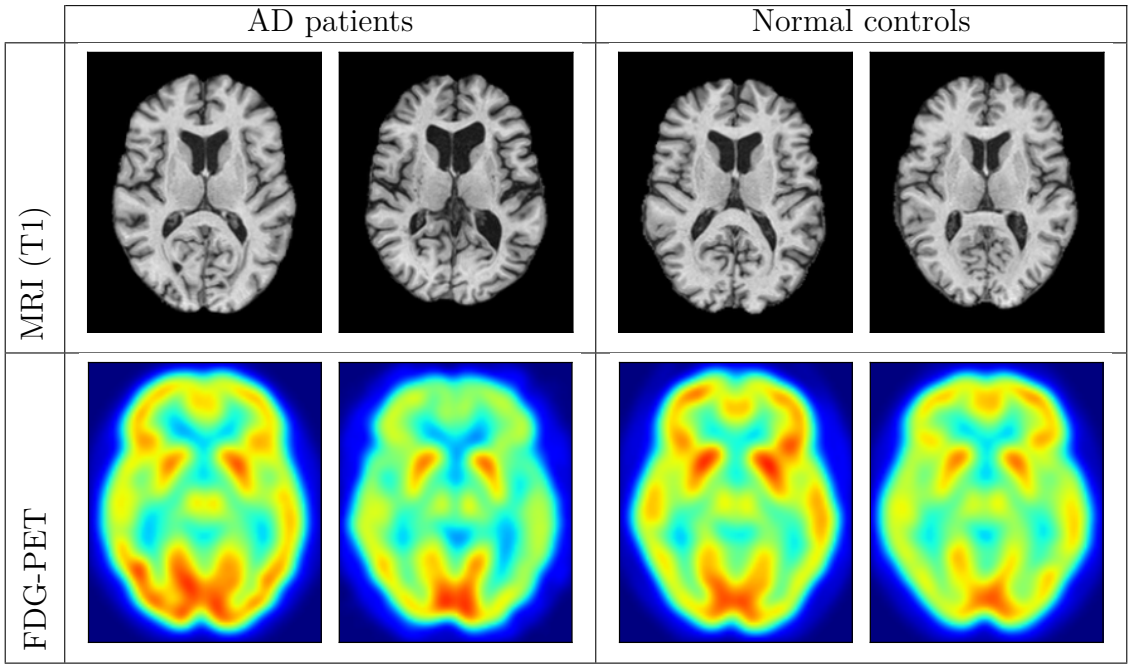


Figure 1.2.1: MR T1-weighted (top) and FDG-PET (bottom) images of AD patients (left panel) and cognitive normal controls (right panel).

Fibrillar Aβ PET is used to study the deposition of amyloid beta plaques in the brain [**216**]. This technique was used to confirm Aβ deposits in the brains of affected patients, particularly in the precuneus, posterior cingulate, parietotemporal and frontal regions [**216**]. In addition, these studies suggested that fibrillar Aβ PET levels are already near saturation in patients with MCI [**216**]. Fibrillar Aβ PET is expected to play an important role in evaluation of potential AD treatments that aim to clear Aβ deposits or prevent their accumulation [**216**]. These tests will also help to further investigate the role of Aβ deposits in AD [**216**].

The Alzheimer's disease neuroimaging intiative (ADNI) is a longitudinal observational study of a large group of Alzheimer's disease patients, MCI patients and a matched group of cognitively normal controls [**197**]. ADNI is a multisite collaboration with subjects receiving regular, standardised MR and (for some of them) PET scans, in addition to neuropsychologic, genetic and cerebrospinal fluid testing [**197, 135, 137**]. Anonymised data can be accessed by researchers as a unified database. The goals of ADNI are to develop standardised imaging protocols for AD and MCI studies, collect structural and metabolic imaging data, validate imaging biomarkers against standard clinical and cognitive measures, compare biomarkers with respect to their utility for AD and MCI diagnosis and tracking of effects due to treatment, and to create a generally accessible data repository [**197**]. ADNI data was used in a large number of scientific papers and the success of the initiative prompted its extension to ADNI-GO and ADNI2 stages [**262**].

1.3. Contributions

The contributions of this dissertation include applications of wavelets, sparse representations and machine learning to the problems of compressed sensing MRI reconstruction and image-based classification of Alzheimer's disease.

- The contribution to compressed sensing consists of adapting the wavelet packet best basis search algorithm [**55, 56**] for application to MR image reconstruction from undersampled data. An optimised basis is learned from a set of brain MR images from the ADNI database. This basis is shown to represent both training images and unseen brain images in a more sparse way than standard wavelets. The optimised basis is also compared to standard wavelets in reconstruction of brain MR images from simulated compressed sensing data. In the context of the rest of this dissertation, compressed sensing can be used to accelerate MRI scans for detection of AD, allowing

for more patients to be examined. This work was presented at the 2012 MICCAI Workshop on Sparsity Techniques in Medical Imaging [**217**].

- The first contribution to image-based classification consists of comparing several dimensionality reduction methods for AD detection based on FDG-PET data. Several feature selection and clustering algorithms are tested with a range of different classification algorithms to evaluate the potential benefits of feature selection in FDG-PET based AD detection. The algorithms are compared with respect to their classification performance as well as the distribution of selected features throughout the brain.
- The second contribution to image-based AD detection consists of extending the invariant scattering convolution network architecture proposed recently by Bruna and Mallat [**27**] to three-dimensional tomographic images and applying this image representation to the problem of AD detection based on MR images. The problems due to the very large dimensionality of 3D scattering data are addressed by applying the fast Johnson-Lindenstrauss transform introduced recently by Ailon and Liberty [**5**] as a form of data compression. The classifiers are learned and applied in the compressed domain, enabling efficient learning and classification in a situation where practical challenges would appear with learning from full-dimensional data.

As a whole, the work presented in this dissertation is aimed at making AD screening and prediction more accessible, by improving the efficiency of brain MRI scanning and developing tools for computer-assisted diagnosis of AD and detection of its early signs.

1.4. Thesis outline

The remainder of this dissertation is organised as described in the following.

Chapters 2 and 3 introduce the relevant background. Chapter 2 covers wavelet representations in signal processing as well as sparse representations and compressed

sensing. Chapter 3 presents an overview of the machine learning techniques that are used in subsequent chapters.

Compressed sensing MRI with optimised wavelet packet representations is discussed in chapter 4. The wavelet packet basis search algorithm [**55, 56**] is adapted for finding an optimally sparse wavelet packet basis for a set of images. This algorithm is then evaluated by fitting a wavelet packet basis to a set of brain MR images and measuring the sparsity of representations of other brain MR images in this basis. The adapted basis is also compared with wavelets in an application to MRI reconstruction from compressed sensing k-space data.

Sparse algorithms for image-based classification of Alzheimer's disease are discussed in chapter 5. Feature selection steps are added to several classification algorithms and these composite classifiers are evaluated on FDG-PET brain images of Alzheimer's disease patients and cognitive normal (CN) subjects. Those trained classifiers are also evaluated on the task of predicting whether an MCI patient will progress to AD or remain stable.

In chapter 6 scattering networks introduced by Bruna and Mallat [**27**] are extended to three-dimensional tomographic images. In addition, a fast Johnson-Lindenstrauss transform introduced by Ailon and Liberty [**5**] is proposed as a method to reduce the dimensionality of scattering network output to make application of machine learning more practical. These algorithms are then evaluated on MRI data from ADNI by classifying between AD patients and controls and in addition making predictions of whether MCI patients will progress to AD or remain stable.

Chapter 7 presents the outlook for the topics discussed in previous chapters, with a discussion of potential extensions of the work presented.

Chapter 8 presents a summary of the whole dissertation and general conclusions.

CHAPTER 2

Wavelets, sparsity and compressed sensing

In signal and image processing it is common to represent signals as sums of simple elements often referred to as atomic signals or simply atoms. Those representations make it possible to extract features of signals that may not be immediately apparent, or ones that appear salient to a human observer but difficult to distinguish automatically with an algorithm. Perhaps the most widely known example of this is the Fourier representation which models signals as sums of sinusoids or complex exponentials [**187**].

2.1. Wavelets

The sinusoids used as atomic signals in Fourier analysis have the benefit of distinguishing frequencies with high resolution but they are unable to localise the features of a signal in time (or space for spatial signals such as images). This has prompted the development of alternative representations which can localise signals in time (or space) at the expense of loss of some frequency resolution.

Wavelet transforms describe signals as sums of scaled and shifted versions of an atomic waveform known as the "mother wavelet". Large-scale atoms can provide a rough approximation of a signal. Meanwhile, compactly supported atoms add detail near discontinuities, such as edges in images. Large values of wavelet coefficients in the small scales appear near edges [**187**] while uniform regions have small coefficients at those scales, which enables efficient image compression.

Some basic concepts from wavelet theory are presented in the following. Since the subject of wavelets is very broad, this discussion is limited to the topics that are relevant to the algorithms discussed in further chapters.

2.1.1. Continuous wavelets in one dimension. A continuous wavelet analysis can be defined by choosing a mother wavelet, *i.e.* a function $\psi(t)$ such that $\int_{-\infty}^{\infty} |\psi(t)|^2 dt = 1$ and $\int_{-\infty}^{\infty} \psi(t)\, dt = 0$ [**187**] where the second condition means that the wavelet averages out to zero over its support. The "daughter wavelets" at scale $s \in \mathbb{R}^+$ and translation $u \in \mathbb{R}$ are then derived from the mother wavelet as follows [**187**].

$$\psi_{u,s}(t) = \frac{1}{\sqrt{s}} \psi\left(\frac{t-u}{s}\right)$$

A function $f(t) \in \mathbf{L}^2(\mathbb{R})$ is transformed into its wavelet representation $F_\psi(u,s)$ as follows [**187**].

$$F_\psi(u,s) = \langle f, \psi_{u,s} \rangle = \int_{-\infty}^{\infty} f(t)\, \psi^*_{u,s}(t)\, dt$$

The following theorem shows that the signal $f(t)$ can be reconstructed from $F_\psi(u,s)$ ($\hat{\psi}(\omega)$ denotes the Fourier transform of $\psi(t)$).

THEOREM 1. *(Calderon, Grossman and Morlet. This version quoted from* [**187**]*)* *"Let $\psi \in \mathbf{L}^2(\mathbb{R})$ be a real function such that*

$$C_\psi = \int_0^{\infty} \frac{|\hat{\psi}(\omega)|^2}{\omega} d\omega < \infty. \tag{2.1.1}$$

Any $f \in \mathbf{L}^2(\mathbb{R})$ satisfies

$$f(t) = \frac{1}{C_\psi} \int_0^{\infty} \int_{-\infty}^{\infty} F_\psi(u,s) \frac{1}{\sqrt{s}} \psi\left(\frac{t-u}{s}\right) du \frac{ds}{s^2} \tag{2.1.2}$$

and

$$\int_{-\infty}^{\infty} |f(t)|^2 dt = \frac{1}{C_\psi} \int_0^{\infty} \int_{-\infty}^{\infty} |F_\psi(u,s)|^2 du \frac{ds}{s^2}." \tag{2.1.3}$$

In other words, under a mild condition (equation 2.1.1) on the wavelet function, the continuous wavelet transform is invertible (with the inverse given by equation 2.1.2), and preserves signal energy up to a constant factor (as shown by equation 2.1.3).

The constant C_ψ ensures that reconstruction with equation 2.1.2 returns the original $f(t)$ rather than its scaled version.

Note that the condition in equation 2.1.1 (called the wavelet admissibility condition [**187**]) implies that it is necessary that $\hat{\psi}(0) = 0$. In other words, this explains why the average of the wavelet over its support must be zero [**187**].

The theory discussed so far explains how to transform continuous functions of time (or space) into wavelet representations that are continuous in time (or space) and scale. These results are important from a theoretical perspective but in computational applications the function to be transformed and the wavelet representation must both be discrete. The simplest way of addressing this problem is to sample the signal and its wavelet representation. The sampling rate of the signal is chosen so as to ensure that no information is lost due to aliasing. The sampling rate of the wavelet transform is often chosen so that the scale is on a logarithmic grid and the time intervals are matched to each scale individually. More specifically, [**105**]

$$s = 2^d, u = n2^d, \quad n \in \mathbb{Z}, d \in \mathbb{Z} \tag{2.1.4}$$

where d controls scale and n controls translation.

The Gabor wavelet is a commonly used wavelet defined as a continuous function. It consitsts of a complex exponential modulated by a Gaussian window [**187**]:

$$\psi_G(t) = \frac{1}{(\sigma^2\pi)^{1/4}} \exp(i\omega_0 t) \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

This function is not a wavelet in the strict sense because it does not average out to zero. However, this problem can be addressed by a simple adjustment, giving the

Morlet wavelet [**27**]:

$$\psi(t) = \alpha\left(\exp(i\omega_0 t) - \beta\right)\exp\left(-\frac{t^2}{2\sigma^2}\right)$$

where β is chosen so that $\int \psi(u)\,du = 0$ [**27**]. The Morlet wavelet is shown in figure 2.1.1.

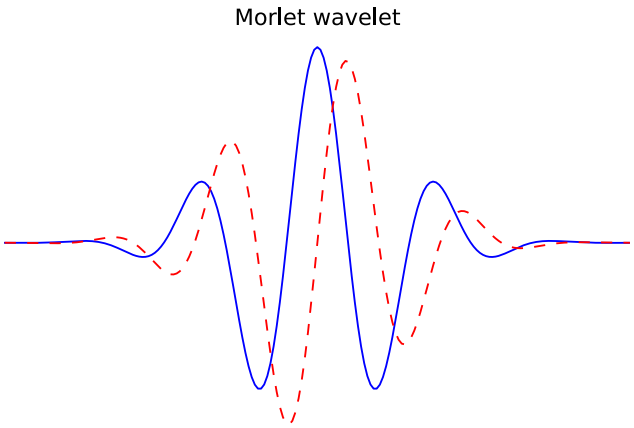


Figure 2.1.1: Morlet wavelet (real part in solid blue, imaginary part in dashed red).

2.1.2. Discrete wavelets in one dimension. An alternative way to define wavelet transforms is to start with multirate filter banks and scaling functions. The structure of a filter bank implementing the discrete wavelet transform (DWT) and the corresponding reconstruction filter bank are shown in figure 2.1.2. The filters $h_0[n]$, $h_1[n]$, $g_0[n]$ and $g_1[n]$ are chosen to ensure perfect reconstruction ($y[n] = x[n]$).

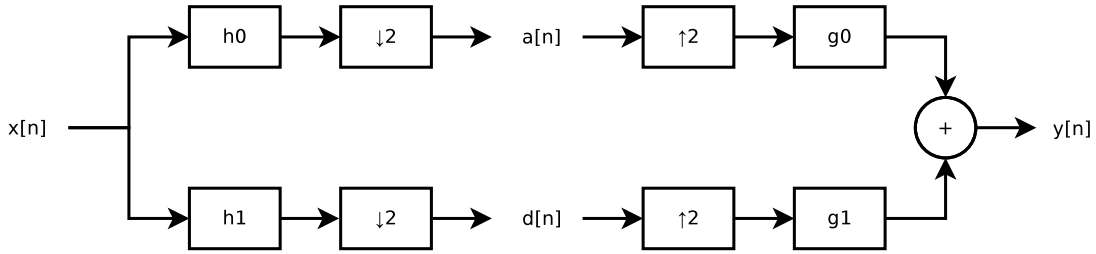


Figure 2.1.2: Discrete wavelet decomposition (left) and reconstruction (right).

The perfect reconstruction condition leaves some freedom to the designer of the filter banks, allowing other objectives to be met. For example, compact support of the

filters (*i.e.* finite length of their impulse responses) is a convenient property because it allows a very efficient implementation with Mallat's multiresolution analysis algorithm [**188**]. The filters can be designed to be orthogonal, so that the wavelet transform is also orthognal [**187**].

Daubechies wavelets [**66**] are a commonly used family of orthogonal, compactly supported wavelets. The Daubechies family consists of a wavelet for each integer number of vanishing moments of the filter $g[n]$, with a minimal support length for that number of vanishing moments, allowing these wavelets to represent polynomials very efficiently [**187**].

A multi-level discrete wavelet decomposition is obtained by iterating the filter bank in figure 2.1.2 on the low-pass branch. For example, the filter tree in figure **2.1.3** decomposes its input into two detail bands and one approximation band. The corresponding reconstruction filter tree is shown in **2.1.4**. Mallat's multiresolution analysis algorithm uses these filter trees to implement a fast wavelet transform [**188**].

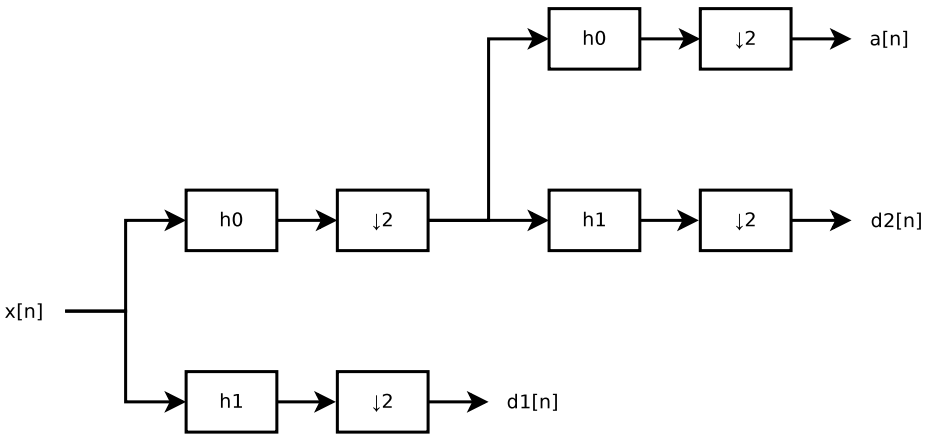


Figure 2.1.3: Multi-level discrete wavelet decomposition.

Filter banks used for a discrete wavelet transform can be used to derive a continuous wavelet and the associated scaling function (the basis function associated with the approximation band) using an iterative refinement algorithm [**66, 30**]. The db4 scaling function, wavelet and the filters $h_0[n]$ and $h_1[n]$ are shown in figure 2.1.5.

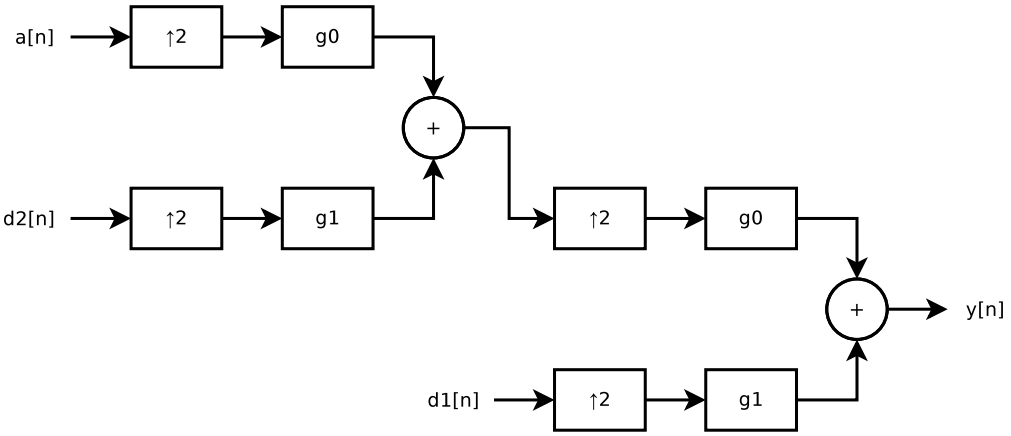


Figure 2.1.4: Multi-level discrete wavelet reconstruction.

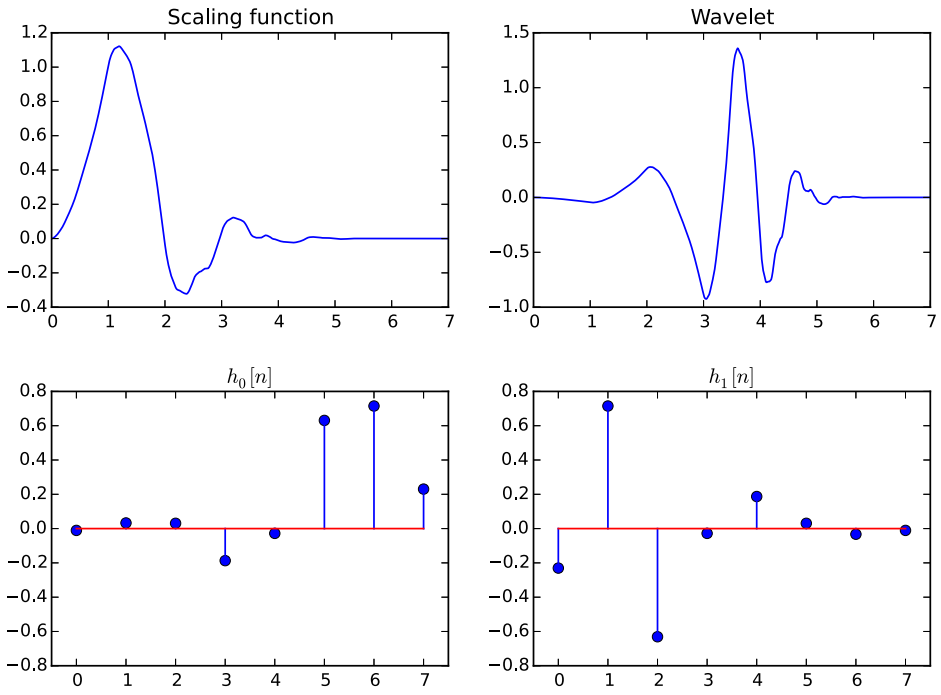


Figure 2.1.5: The *db*4 scaling function, wavelet and decomposition filters $h_0[n]$ and $h_1[n]$.

Discrete wavelet transforms can easily be expressed as matrix multiplication of a transform matrix and a signal vector [**30**]. The transform matrix is formed from the basis vectors of the wavelet representation as rows, and both single-level and multi-level transforms can be represented in this way. In computational applications the direct implementation of filter banks is much faster but the matrix representation is helpful in the analysis of theoretical aspects of the discrete wavelet transform.

Unless the signals are of infinite duration, both continuous and discrete formulations of wavelets have to handle boundary conditions. Some extension (padding) of the signal has to be assumed at the start and end of the signal. Periodic padding (*i.e.* concatenating a copy of the signal before its start and after its end) is elegant from a mathematical point of view because the associated transform matrix is circulant and with orthogonal filter banks it is also orthogonal. The disadvantage of this approach is that it often introduces a discontinuity at the point of concatenation. An alternative apprach is symmetric padding, where a reversed version of the signal is concatenated at both of its ends. This avoids the discontinuity but orthogonality is lost.

2.1.3. Image processing with wavelets. Discrete wavelet transforms for images are computed by applying the filters $h_0[n]$ and $h_1[n]$ to all rows of an image individually and then to the columns of the resulting representation. After downsampling, this gives four frequency bands: approximation, horizontal detail, vertical detail and diagonal detail. Alternatively, the frequency bands can be computed by pre-computing a two-dimensional kernel for each frequency band and convolving it with the image, followed by downsampling.

An example of a single-level two-dimensional wavelet decomposition is shown in figure 2.1.6. A multi-level decomposition for images is obtained by iterating the single level decomposition on the approximation band.

This definition of a wavelet transform for images is easily extended to three-dimensional tomographic images by applying a third pair of filters along the z-direction. A multi-level decomposition is then derived analogously by iterating the filter bank on the approximation band.

An alternative way of applying wavelets to images is to define a set of directional filters with different orientations. The filters have a band-pass profile along their longitudinal axis and low-pass profile along the transverse axis (or axes in case of

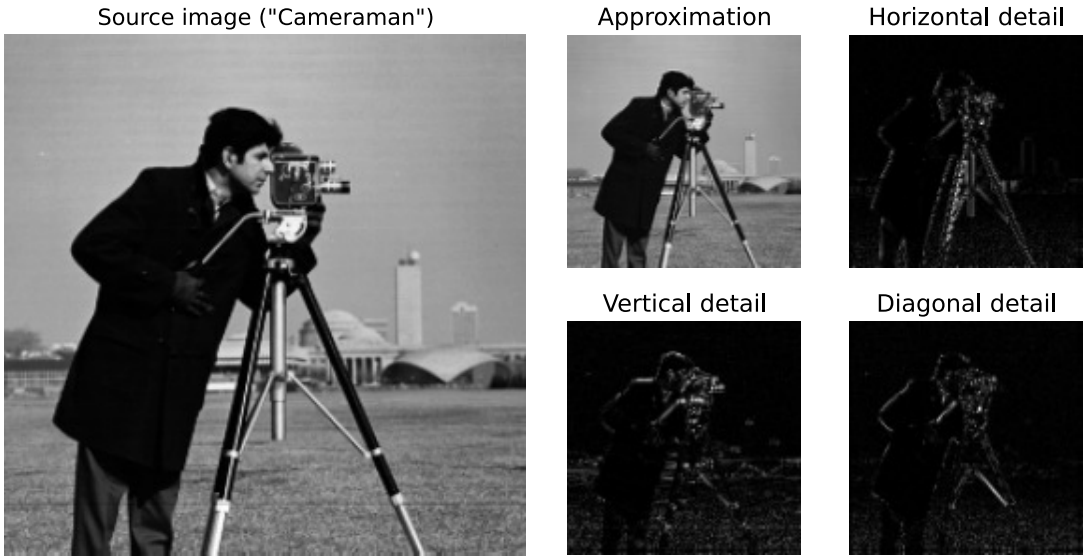


Figure 2.1.6: Single-level 2D wavelet decomposition. The source image (left) is the standard "Cameraman" test image. To the right, the approximation and directional detail bands. Note that the intensities of the images on the right were individually rescaled to cover the grayscale colour range, *i.e.* they are not directly comparable in terms of coefficient size.

3D), which makes this method useful for detecting edges (in 2D) or surfaces (in 3D). There are several transforms of this type, including wedgelets [**76**], bandelets [**161**], curvelets [**40**] and shearlets [**154, 115**].

2.2. Sparse representations

Since wavelet atoms are localised and separate data into different scales, they can approximate images in a sparse way. The low-pass bands are sufficient to reconstruct smooth regions in an image with good accuracy so the detail bands in the corresponding regions are close to zero and their ommision would have little effect on the quality of the image. This can be seen *e.g.* in figure 2.1.6, where only a small proportion of coefficients in the detail bands are significantly far from zero. Figure 2.2.1 shows the decay of coefficients in the wavelet representation of the "Cameraman" image with three levels of decomposition. When the image is reconstructed from approximately 10% of its largest wavelet coefficients, the result (figure 2.2.2) is very close to the original.

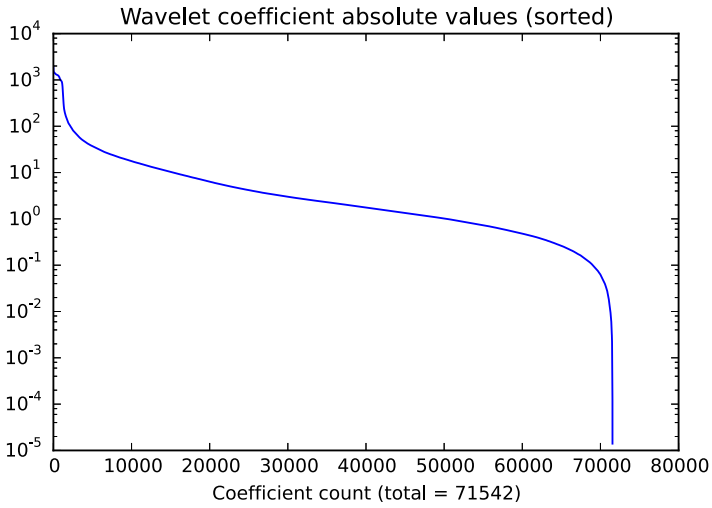


Figure 2.2.1: Decay of weavelet coefficients. The plot shows absolute values of wavelet coefficients for the "Cameraman" image, with three levels of wavelet decomposition. Note logarithmic vertical axis.

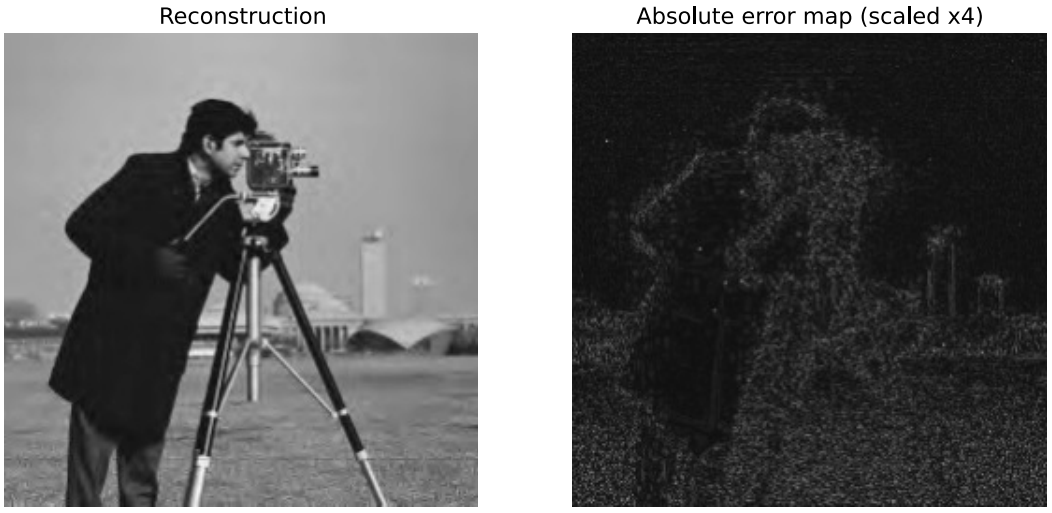


Figure 2.2.2: Left: sparse wavelet image reconstruction of the "Cameraman" image (shown in figure 2.1.6) from approximately 10% of its wavelet coefficients with largest absolute values. Right: error map of this reconstruction.

Sparsity in the strict sense can be defined as follows.

DEFINITION 2. (Tropp, [**251**]) s-sparse vector

A vector $\mathbf{x}$ is said to be s-sparse if $\|\mathbf{x}\|_0 \leq s$ where $\|.\|_0$ is a function that returns the number of non-zero components of its argument.

2.2.1. Optimising for sparsity. Orthogonal transforms provide unique representations for signals, so optimal approximations for different sparsity-accuracy trade-offs are easily obtained by setting the smallest coefficients to zero until the rquired level of sparsity is reached. However, non-orthogonal dictionaries require more advanced algorithms to find optimal sparse approximations.

Consider the representation equation $\mathbf{x} = \mathbf{D}\gamma$ where $\mathbf{x} \in \mathbb{R}^m$ is the vector representing the (discrete) signal, $\mathbf{D} \in \mathbb{R}^{m \times n}$ is the matrix representing the dictionary, with atoms as columns, and $\gamma \in \mathbb{R}^n$ is the vector of representation coefficients. With an overcomplete dictionary this system is underdetermined, with no unique solution. This means that there is some freedom in choosing a representation. This problem can be written as [**251**]

$$\min_{\gamma \in \mathbb{R}^n} \|\gamma\|_0 \quad s.t. \quad \mathbf{x} = \mathbf{D}\gamma \tag{2.2.1}$$

where $\|\gamma\|_0$ is the ℓ_0 pseudonorm, *i.e.* the nomber of non-zero components of the vector γ.

It turns out that the problem in equation 2.2.1 is very difficult to solve computationally. Some algorithms such as matching pursuit (MP) [**189**] and orthogonal matching pursuit (OMP) [**202**] can find solutions that are typically suboptimal (but these algorithms are fast). Alternatively, under some conditions ([**78, 77**]; also see the discussion on compressed sensing below) it is possible to solve an ℓ_1 version of the problem in equation 2.2.1, *i.e.*

$$\min_{\gamma \in \mathbb{R}^n} ||\gamma||_1 \quad s.t. \quad \mathbf{x} = \mathbf{D}\gamma \tag{2.2.2}$$

and with large probability get a solution that is also optimal for 2.2.1. The problem 2.2.2 is known in signal processing as basis pursuit [**45**]. It is a linear optimisation problem that has been studied extensively by mathematical optimisation researchers and a number of efficient algorithms are known that can be used to find a solution, as discussed by Tropp and Wright [**251**].

The discussion so far focused on the problem of sparse representation, *i.e.* finding a combination of atoms that represents a signal exactly. However, in reality this might not be the right problem to solve because the model with all signals composed of a small number of atoms each is an idealised one. A more realistic approach is to allow a trade-off between sparsity and aproximation error, which gives the following optimisation problem.

$$\min_{\gamma \in \mathbb{R}^n} \frac{1}{2} \|\mathbf{D}\gamma - \mathbf{x}\|_2^2 + \lambda \|\gamma\|_1 \tag{2.2.3}$$

This problem is known as basis pursuit denoising (BPDN) [**45**]. It can also be written in two alternative forms, as follows.

$$\min_{\gamma \in \mathbb{R}^n} \|\gamma\|_1 \quad s.t. \quad \frac{1}{2} \|\mathbf{D}\gamma - \mathbf{x}\|_2^2 \leq \epsilon^2 \tag{2.2.4}$$

$$\min_{\gamma \in \mathbb{R}^n} \frac{1}{2} \|\mathbf{D}\gamma - \mathbf{x}\|_2^2 \quad s.t. \quad \|\gamma\|_1 \leq \tau \tag{2.2.5}$$

Both of those two problems can be transformed into the form in equation 2.2.3 using the method of Lagrangian multipliers. The BPDN problem can be solved computationally with several types of algorithms, which are discussed in detail by Tropp and Wright [**251**].

In statistics the problem in equation 2.2.5 is also referred to as least absolute shrinkage selection operator (LASSO) [**246**].

2.2.2. Adpative sparse dictionaries. Sparse representations can also be constructed adaptively, by matching the dictionary of atomic signals to a particular image or set of images. This problem is referred to as dictionary learning [**198, 152, 3, 141, 88, 269, 29, 165, 250, 213**]. It can be formulated mathematically as

$$\min_{\mathbf{D}\in\mathbb{R}^{m\times n},\gamma\in\mathbb{R}^n} \left(\frac{1}{2}\left\|\mathbf{x}-\mathbf{D}\gamma\right\|_2^2+\lambda\left\|\gamma\right\|_1\right) \tag{2.2.6}$$

which is essentially the same problem as the one in formula 2.2.3, except that the minimisation is over $\mathbf{D}$ in addition to γ, making it much more difficult to solve.

Dictionary learning literature was reviewed by Tosic and Frossard [**250**]. The K-SVD algorithm [**3**] is worth mentioning in particular. It alternates between sparsely representing an image in a set of atoms and optimising the atoms, with efficient implementations for both of these steps.

Dictionary learning algorithms are usually applied to small image patches [**198, 3, 86, 141, 185, 186, 87, 220, 165**], rather than whole images which is the case with wavelets. Therefore, sparsity in dictionary learning is accomplished in the sense of individual patches being sparse in the learned dictionary of patch-size atoms. Olshausen and Field [**198**] found that the atoms in a dictionary optimised (with a cost function slightly different from the one in equation 2.2.6 and a different optimisation algorithm) for patches of natural images resembled directional wavelets.

The sparsity of wavelet transforms can be improved with adaptive wavelet packet representations [**55, 56**]. Wavelet packets are an extension of discrete wavelets where filter banks are iterated on the high-pass branches in addition to low-pass branches. An adaptive wavelet packet representation is then built by choosing the point in each branch where the iteration should stop. In contrast to the dictionary learning algorithms discussed above, a wavelet packet basis is normally optimised for a whole image or set of images instead of image patches. Wavelet packets are discussed in detail in chapter 4.

2.3. Compressed sensing

Compressed sensing (or compressive sensing) [**36, 38, 42, 82**], abbreviated as CS, is a mathematical signal processing technique that can be used to reconstruct sparse signals from a reduced number of linear measurements. A compressed sensing system can be modeled with the equation

$$\mathbf{y} = \mathbf{A}\gamma \tag{2.3.1}$$

where $\mathbf{y} \in \mathbb{R}^m$ is the vector of measurements, $\mathbf{A} \in \mathbb{R}^{m \times n}$ is the sensing matrix and $\gamma \in \mathbb{R}^n$ is the sparse vector that is being measured. We can think of γ as being in the "sparse representation space" and $\mathbf{y}$ as being in the "sensing space".

This model can easily accommodate the case where we are trying to reconstruct a data vector $\mathbf{x}$ in a "human readable space" which is not itself sparse but can be represented sparsely in another basis, *i.e.* when $\mathbf{x} = \mathbf{D}\gamma$ as discussed in section 2.2. In this case we have

$$\mathbf{y} = \mathbf{\Phi}\mathbf{D}\gamma$$

where $\mathbf{\Phi}$ is the "physical" sensing matrix that transforms an object from its natural representation (*e.g.* pixels or voxels) to compressive measurements. This reduces to 2.3.1 when $\mathbf{A} = \mathbf{\Phi}\mathbf{D}$, *i.e.* when we consider the system as a whole to be taking direct linear measurements of the sparse representation.

The structure of the matrix $\mathbf{\Phi}$ depends on the particular sensing system under consideration. For example, in magnetic resonance imaging $\mathbf{\Phi}$ is a matrix constructed from a subset of rows of the Fourier transform matrix.

2.3.1. Restricted isometries. Intuitively, to enable reconstruction of sparse signals, it is necessary for the matrix $\mathbf{A}$ to transform different sparse vectors γ into different measurement vectors $\mathbf{y}$. Otherwise, if there were sparse vectors γ_1 and γ_2 such that $\mathbf{y} = \mathbf{A}\gamma_1 = \mathbf{A}\gamma_2$, it would be impossible to distinguish γ_1 and γ_2 based only on their measurement vectors with any method. More formally, we define the *restricted isometry property*, which is characterised by the *restricted isometry constant* of a matrix.

Definition 3. (Originally Candes and Tao [**38**], this version from [**34**]) Restricted isometry constant

For each integer s, the restricted isometry constant δ_s of a matrix $\mathbf{A}$ is the smallest number such that

$$(1 - \delta_s) \|\gamma\|_2^2 \leq \|\mathbf{A}\gamma\|_2^2 \leq (1 + \delta_s) \|\gamma\|_2^2$$

holds for all s-sparse vectors γ.

Essentially, this definition formalises the requirement that the transformation $\mathbf{A}$ only distorts the square of the ℓ_2 norm of any s-sparse vector to a degree limited by δ_s.

If we consider two s-sparse vectors γ_1 and γ_2, the ℓ_0 pseudo-norm of their difference is at most $2s$. Therefore, the restricted isometry constant δ_{2s} places a limit on the degree to which the pairwise distances between s-sparse vectors can be distorted by the transformation $\mathbf{A}$ [**39**].

Intuitively, it makes sense to have a matrix $\mathbf{A}$ with δ_{2s} that is as close to 0 as possible, since this results in Euclidean distances between sparse vectors only becoming distorted to a small degree by the transformation defined by $\mathbf{A}$. In order for $\mathbf{A}$ to encode s-sparse vectors unambiguously, it is necessary that $\delta_{2s} < 1$, which ensures

that there are no $2s$-sparse vectors in the null space of $\mathbf{A}$ and thus that reconstruction by ℓ_0 minimisation has a unique solution [**39**]. Matrices with good restricted isometry constants can be generated using random matrix constructions (with high probability) [**15, 34**] or deterministically [**74**].

2.3.2. Noiseless compressed sensing recovery. Returning to the model in equation 2.3.1, the signal γ usually is not sparse in the strict sense but a sparse approximation can be constructed with a small number s of its components with the largest magnitude. Let us denote this approximation as γ_s. As already discussed in section 2.2, such approximations can be very accurate even when s is small relative to the dimension of γ if the right representation is chosen (see figure 2.2.1 for the wavelet example). The aim of compressed sensing recovery is to reconstruct γ_s from $\mathbf{y}$.

Compressed sensing reconstruction essentially consists of using a sparse coding algorithm to find the sparsest vector $\gamma^\star$ that is consistent with the measurements seen in $\mathbf{y}$. In particular, when ℓ_1 minimisation is used as the reconstruction algorithm, the following theorem states the requirements for exact reconstruction of γ_s.

THEOREM 4. *Noiseless recovery* [**34**]

If $\mathbf{y} = \mathbf{A}\gamma$ *and the matrix* $\mathbf{A}$ *has the restricted isometry constant* $\delta_{2s} < \sqrt{2} - 1$ *and* $\gamma^\star$ *is the solution to*

$$\min_{\tilde{\gamma} \in \mathbb{R}^n} \|\tilde{\gamma}\|_1 \quad s.t. \quad \mathbf{A}\tilde{\gamma} = \mathbf{y}$$

then

$$\|\gamma^\star - \gamma\|_1 \leq C_0 \|\gamma - \gamma_s\|_1$$

and

$$||\gamma^\star - \gamma||_2 \leq C_0 \frac{||\gamma - \gamma_s||_1}{\sqrt{s}}$$

for some constant C_0. In particular, if γ is s-sparse, the recovery is exact.

This theorem means essentially that if the sensing matrix transformation does not distort the pairwise distances between s-sparse vectors too much, then ℓ_1 minimisation recovers an s-sparse vector exactly. In addition, if the vector γ is only approximately sparse, then the reconstruction error is bounded by the error of the s-sparse approximation multiplied by a constant. This is important because it establishes that compressed sensing with ℓ_1 reconstruction is still effective when signal representations are only approximately sparse [**39**].

2.3.3. Noisy compressed sensing recovery. In reality the model 2.3.1 with approximate sparsity is still somewhat idealised in the sense that it does not include noise in the system. The model of a noisy compressed sensing system is

$$\mathbf{y} = \mathbf{A}\gamma + \mathbf{z}$$

where $\mathbf{z}$ is the noise term.

The following theorem then puts a bound on reconstruction error for reconstruction with ℓ_1 minimisation.

THEOREM 5. *Noisy recovery* [**34**]

If $\mathbf{y} = \mathbf{A}\gamma + \mathbf{z}$ and the matrix $\mathbf{A}$ has the restricted isometry constant $\delta_{2s} < \sqrt{2} - 1$ and $\|\mathbf{z}\|_2 \leq \varepsilon$ and $\gamma^\star$ is the solution to

$$\min_{\tilde{\gamma} \in \mathbb{R}^n} \|\tilde{\gamma}\|_1 \quad s.t. \quad \|\mathbf{y} - \mathbf{A}\tilde{\gamma}\|_2 \leq \varepsilon$$

then

$$\|\gamma^\star - \gamma\|_2 \leq C_0 \frac{\|\gamma - \gamma_s\|_1}{\sqrt{s}} + C_1 \varepsilon \tag{2.3.2}$$

for some constants C_0 and C_1. In particular, if γ is s-sparse, the recovery is exact.

The constants C_0 and C_1 are quite reasonable: for example, when $\delta_{2s} = 0.2$, the error in 2.3.2 is bounded by $4.2\frac{\|\gamma-\gamma_s\|_1}{\sqrt{s}} + 8.5\varepsilon$ [**34**].

Theorem 5 establishes that reconstruction error of ℓ_1 recovery scales linearly with both sensor noise and approximation error of the sparse signal model. Therefore, compressed sensing with ℓ_1 recovery is robust, which is important for practical applications of this theory.

2.4. Applications to neuroimaging

Wavelets and sparse methods have found many applications in neuroimaging. Several authors applied wavelet analysis to the problem of statistical testing of activation maps in fMRI analysis [**23, 222, 71, 196, 92, 28, 258, 146, 201**]. These methods are an alternative to conventional statistical parametric mapping (SPM) [**103**]. Wavelets are useful in fMRI analysis because of their denoising property: piece-wise smooth signals can be approximated closely by a relatively small number of large coefficients while noise is distributed evenly, so wavelet shrinkage tends to improve signal quality [**81, 80**]. Since wavelet analysis eliminates the requirement of image smoothing with Gaussian filters to reduce noise, these wavelet methods can map brain activity with higher resolution [**196**].

Voxel-based morphometry (VBM) [**9**], which allows SPM to be applied to structural brain images, was extended with wavelets by Canales-Rodriguez *et al.* [**33**].

Classification of Alzheimer's disease using structural MRI and the dual-tree complex wavelet transform was proposed by Hackmack *et al.* [**117**]. Lao *et al.* [**156**] proposed a method for morphological classification of brain images that uses the discrete wavelet transform to enable efficient reduction of data dimensionality.

Dictionary learning was applied in the neuroimaging context to fMRI analysis [**163, 255, 85**], hippocampus segmentation [**248**], lesion segmentation [**263**] and brain atlas construction [**229**].

Neuroimaging is also an important application of compressed sensing. MRI was used as an example application in early work on CS by Candes *et al.* [**36**] and a complete CS-MRI system was soon built by Lustig *et al.* [**183, 181**], using the discrete wavelet transform and image instensity variation for sparsifying structural MR images. Extensive literature on compressed sensing MRI is available, with a recent review by Hollingsworth available in [**132**]. Compressed sensing was also combined with dictionary learning for MR imaging [**211, 236**]. Chapter 4 starts with a more detailed discussion of compressed sensing MRI techniques that represent images with adaptive sparse dictionaries.

2.5. Conclusions

This chapter introduced some basic concepts in wavelets, sparsity and compressed sensing. The discrete wavelet transform is a prerequisite for wavelet packets which chapter 4 is focused on. Morlet wavelets are used in chapter 6 as a building block of scattering networks. Sparsity is important in chapter 5 where classification algorithms with ℓ_1 regularisation are evaluated alongside other methods. Next chapter introduces basic concepts in machine learning, including several classification algorithms.

CHAPTER 3

Machine learning

Machine learning (ML) is the study of algorithms that adapt to data, with particular emphasis on algorithms that detect nontrivial patterns or make predictions about future or missing data. The field of machine learning is closely related to computational statistics and those two fields intersect and inspire each other.

Machine learning problems can be divided into supervised learning and unsupervised learning. Supervised learning is concerned with problems where the reference values of the target variables ("ground truth") are available for the instances in the training set. Supervised learning models learn by adjusting their internal variables so that their outputs predict target variables based on input variables (predictors). Supervised learning can be categorised into regression (predicting a continuous-valued target variable) and classification (predicting a discrete-valued label).

Unsupervised learning is concerned with problems where reference values of the outputs are not available but patterns are still sarched for. Clustering, where one seeks to group data into clusters of similar samples, is an example of an unsupervised learning problem.

3.1. Classification

Classification is a type of supervised learning where the goal is to predict a discrete-valued label. For example, one may be interested in predicting based on an MR image if a patient would be diagnosed as healthy, suffering from MCI or from AD (three possible labels). This section is intended to provide the background on the

three classification algorithms that are used later in this dissertation: logistic regression, support vector machines and random forests.

3.1.1. Logistic regression. Logistic regression relies on the logit transformation to build a model that predicts class probabilities for each feature vector $\mathbf{x}$. A fitted logistic regression model can be used for classification by selecting the label that has the highest predicted probability for a given vector of predictors.

For the case of two classes, the predicted class probabilities are as follows [**123**].

$$\Pr(Y = 1|\mathbf{X} = \mathbf{x}; \beta) = \frac{\exp\left(\beta^T \mathbf{x}\right)}{1 + \exp\left(\beta^T \mathbf{x}\right)}$$
$$\Pr(Y = 0|\mathbf{X} = \mathbf{x}; \beta) = \frac{1}{1 + \exp\left(\beta^T \mathbf{x}\right)}$$

where $(\mathbf{X}, Y) \in \mathbb{R}^{p+1} \times \{0, 1\}$ is a random variable that represents data points and their labels. We use $(\mathbf{x}_i, y_i)$, $i \in [1, \ldots, n]$ to denote a specific sample. It is assumed that the feature vector $\mathbf{x}$ has a "1" prefixed to allow for an intercept in the model (*i.e.* $\mathbf{x} = \begin{bmatrix} 1 & x_1 & \cdots & x_p \end{bmatrix}$) and β is a vector with the same dimension as $\mathbf{x}$. The semicolon is used to separate variables from model parameters.

Given a set of training data $(\mathbf{x}_i, y_i)$, $i \in [1, \ldots, n]$, a logistic regression model is fitted by maximising the log-likelihood of the data in the following way (derivation based on [**123**]).

The log-likelihood of the data given the model is

$$\ell(\beta) = \sum_{i=1}^{n} \log \Pr(Y = y_i|\mathbf{X} = \mathbf{x}_i; \beta)$$

which can be written as

$$\ell(\beta) = \sum_{i=1}^{n} \left[y_i \log \Pr(Y = 1|\mathbf{X} = \mathbf{x}_i; \beta) + (1 - y_i) \log \Pr(Y = 0|\mathbf{X} = \mathbf{x}_i; \beta)\right]$$

which ensures that when $y_i = 1$ we take $\log \Pr(Y = 1|\mathbf{X} = \mathbf{x}_i; \beta)$, and when $y_i = 0$ we take $\log \Pr(Y = 0|\mathbf{X} = \mathbf{x}_i; \beta)$. After substituting the logistic regression probability model and simplifying, this becomes

$$\ell(\beta) = \sum_{i=1}^{n} \left[y_i \beta^T \mathbf{x}_i - \log\left(1 + \exp\left(\beta^T \mathbf{x}_i\right)\right)\right].$$

The derivative with respect to β is

$$\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{n} \mathbf{x}_i \left[y_i - \frac{\exp\left(\beta^T \mathbf{x}_i\right)}{1 + \exp\left(\beta^T \mathbf{x}_i\right)} \right].$$

This expression can be used to find the maximum with gradient descent or alternatively the Hessian can also be derived to enable solution with the Newton-Raphson method [**123**].

3.1.2. Overfitting and regularisation. If the number of features available for building a machine learning model such as logistic regression is large enough then it is possible to fit the data in a near-perfect way. However, the data available for training in most cases only represents a limited number of samples at a limited number of points in the feature space. In addition, data often contains noise, such as imperfect measurements or incorrect labels. The source of the the data may also be probabilistic in nature, with labels depending on features in a way which is to some degree random. These issues can lead to a problem called overfitting, where the model fits the training data very well but fails to generalise to new data.

The problem of overfitting can be alleviated by including a regularisation term in the cost function. Most commonly, regularisation is based on ℓ_1 or ℓ_2 norm of the weight vector β (excluding the intercept β_0).

The ℓ_1 version is fitted by solving the following optimisation problem [**123**].

$$\max_{\beta} \left(\sum_{i=1}^{n} \left[y_i \beta^T \mathbf{x}_i - \log\left(1 + \exp\left(\beta^T \mathbf{x}_i\right)\right)\right] - \lambda \sum_{j=2}^{n} |\beta_j| \right)$$

The ℓ_2 version is similar, with the following optimisation problem.

$$\max_{\beta} \left(\sum_{i=1}^{n} \left[y_i \beta^T \mathbf{x}_i - \log\left(1 + \exp\left(\beta^T \mathbf{x}_i\right)\right)\right] - \lambda \sum_{j=2}^{n} \beta_j^2 \right)$$

These are maximisation problems, so the norm has to be included with negative sign so that smaller values of the norm are preferred.

Regularisation biases the solution towards smaller weights, reducing model variance. If the weight of the regularisation term is adjusted well then the reduction in variance outweighs the bias introduced and generalisation performance improves relative to the unregularised model.

The regularisation term treats the coefficients associated with all predictors (except for the intercept) equally, so it is common to normalise the predictors before applying regularised logistic regression [**123**]. Algorithms for solving regularised logistic regression problems computationally were proposed by several authors [**283, 284, 145, 102, 273, 247**].

If there are more than two classes, the logistic regression model can be extended to accommodate that requirement, although this results in more mathematical complexity. Alternatively, it is possible to use a two-class version in a one-vs-rest framework. In this case a classifier is built for each class that can distinguish that class from all other classes combined in one set. The class label prediction is then assigned as the class with the highest probability as computed by its own vs-rest classifier. Another alternative framework for multi-class classification is one-vs-one, where a classifier is trained to distinguish between each pair of classes and the class label prediction is the class with the most pairwise tests resolved in its favour.

3.1.3. Support vector machines. The idea of support vector machines (SVM) [**254**] is to find a hyperplane that separates data points so that the two classes are on opposite sides of the hyperplane and there is a maximum possible margin between

the hyperplane and any of the points. The derivation of the SVM optimisation problem below is adapted from [**123**].

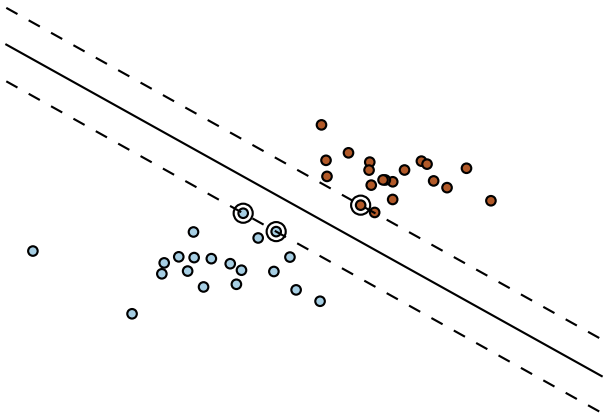


Figure 3.1.1: Classification problem with samples from two groups (marked with different colours). The separating hyperplane is marked with a solid line. The dashed lines bound the maximum margin. The samples at the edges of the margin (marked with larger circles) are the support vectors. (Image generated with Scikit-learn example code.)

A hyperplane is defined by the equation

$$\mathbf{x}^T\beta + \beta_0 = 0$$

where $\mathbf{x} \in \mathbb{R}^p$ is a point on the hyperplane, $\beta \in \mathbb{R}^p$ is the normal vector of the hyperplane and $\beta_0 \in \mathbb{R}$ is a scalar. The signed distance of an arbitrary point $\mathbf{x}$ from this plane is [**123**]

$$d = \frac{1}{\|\beta\|_2}\left(\mathbf{x}^T\beta + \beta_0\right)$$

The sign of d is of particular interest because it indicates on which side of the hyperplane the point $\mathbf{x}$ lies. The problem of finding a separating hyperplane that maximises the margin can then be written as

$$\max_{\beta,\beta_0} M \quad s.t. \quad z_i \frac{1}{\|\beta\|_2}\left(\mathbf{x}_i^T\beta + \beta_0\right) \geq M,\ i = 1, \ldots, n \tag{3.1.1}$$

where M is the margin and $z_i \in \{-1, 1\}$ are the respective labels for the training instances $\mathbf{x}_i$, -1 for one class and 1 for the other. As argued in [**123**], "since for any

β and β_0 satisfying these inequalities, any positively scaled multiple satisfies them too, we can arbitrarily set $\|\beta\|_2 = 1/M$". Thus, 3.1.1 is equivalent to

$$\min_{\beta,\beta_0} \|\beta\|_2 \quad s.t. \quad z_i \left(\mathbf{x}_i^T \beta + \beta_0\right) \geq 1,\ i = 1, \ldots, n$$

which in turn is equivalent to

$$\min_{\beta,\beta_0} \frac{1}{2} \|\beta\|_2^2 \quad s.t. \quad z_i \left(\mathbf{x}_i^T \beta + \beta_0\right) \geq 1,\ i = 1, \ldots, n$$

Often it turns out that there is no hyperplane that can separate the classes perfectly. In this case it is possible to add slack variables that allow points to be on the wrong side of the hyperplane at some cost, giving the following optimisation problem.

$$\min_{\beta,\beta_0} \frac{1}{2} \|\beta\|_2^2 \quad s.t. \quad z_i \left(\mathbf{x}_i^T \beta + \beta_0\right) \geq 1 - \xi_i,\ \xi_i \geq 0, \sum_{i=1}^{n} \xi_i \leq \text{constant}, i = 1, \ldots, n$$

With Lagrange multipliers, this can be written as [**123**]

$$\min_{\beta,\beta_0} \frac{1}{2} \|\beta\|_2^2 + C \sum_{i=1}^{n} \xi_i \quad s.t. \quad \xi_i \geq 0,\ z_i \left(\mathbf{x}_i^T \beta + \beta_0\right) \geq 1 - \xi_i,\ i = 1, \ldots, n$$

which removes the constant and introduces the Lagrange multiplier C. This optimisation problem is a convex cost function with linear constraints, so it can be solved with efficient convex optimisation algorithms [**21**]. By adding further Lagrange multipliers α_i and μ_i the following Lagrangian form can be obtained [**123**].

$$L_p = \frac{1}{2} \|\beta\|_2^2 + C \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i \left[z_i \left(\mathbf{x}_i^T \beta + \beta_0\right) - (1 - \xi_i)\right] - \sum_{i=1}^{n} \mu_i \xi_i$$

The SVM cost function has the following dual [**123**] (see Appendix for the definition of the dual function).

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j z_i z_j \mathbf{x}_i^T \mathbf{x}_j \tag{3.1.2}$$

The vector $\hat{\beta}$ can then be reconstructed as [**123**]

$$\hat{\beta} = \sum_{i=1}^{n} \hat{\alpha}_i z_i \mathbf{x}_i \tag{3.1.3}$$

and the classification rule is [**123**]

$$g(\mathbf{x}) = \text{sign}\left(\mathbf{x}^T\hat{\beta} + \hat{\beta}_0\right)$$

which by substituting equation 3.1.3 becomes

$$g(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{n}\hat{\alpha}_i z_i \mathbf{x}^T\mathbf{x}_i + \hat{\beta}_0\right). \tag{3.1.4}$$

It follows from equations 3.1.2 and 3.1.4 that the computation of the vector $\hat{\beta}$ is not necessary if the $\hat{\alpha}_i$ are computed instead, and also that knowing the vectors $\mathbf{x}_i$ is not necessary as long as there is a way of computing the products $\mathbf{x}_i^T\mathbf{x}_j$. This becomes even more important if an additional transformation $h(\mathbf{x})$ is applied to the input vectors before taking inner products. Equations 3.1.2 and 3.1.4 then become [**123**]

$$L_D = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j z_i z_j \langle h(\mathbf{x}_i), h(\mathbf{x}_j)\rangle$$

$$g(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{n}\hat{\alpha}_i z_i \langle h(\mathbf{x}_i), h(\mathbf{x})\rangle + \hat{\beta}_0\right)$$

This means that it is not necessary to be able to compute $h(\mathbf{x})$ as long as there is a way of computing the kernel function $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \langle h(\mathbf{x}_i), h(\mathbf{x}_j)\rangle$, which should be symmetric positive semi-definite [**123**].

The support vectors can be identified in the dual representation of the SVM classifier as those vectors $\mathbf{x}_i$ that have non-zero $\hat{\alpha}_i$ associated with them. Equation 3.1.4 shows that the support vectors, along with their respective z_i and $\hat{\alpha}_i$, and $\hat{\beta}_0$ are sufficient to reconstruct the decision function [**123**].

For a practical example of a non-linar kernel, the radial basis function (RBF) kernel is $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|_2^2\right)$ [**123**]. Kernel SVM with a non-linear kernel corresponds to a non-linear decision boundary in the feature space (figure 3.1.2 shows decision boundaries for SVM with three different kernels) [**123**].

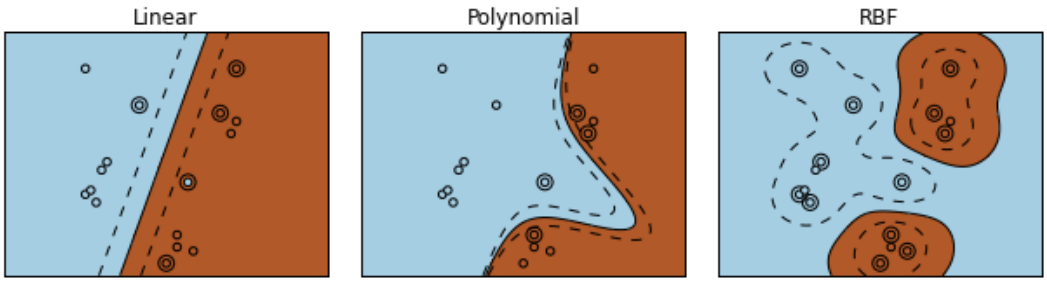


Figure 3.1.2: SVM with different kernels (left to right): linear, polynomial and RBF. Solid lines represent decision boundaries in the feature space. Support vectors marked with larger circles. Dashed lines represent level sets of the decision functions (not classification margins). (Images generated with Scikit-learn example code.)

As with logistic regression, the predictors used with SVM are usually standardised as a pre-processing step to ensure that all predictors are in a similar range of values.

3.1.4. Classification trees and random forests. Compared to the linear classsifers discussed above, classifcation trees [**25**] are a very different concept. An instance is classified by traversing a tree, starting from its root node. At each internal node a condition on the instance is tested and, depending on the result of the test, the tree is traversed towards one of the child nodes. This process is repeated until one of the leaf nodes is reached. Each leaf node has a label assigned to it and that label is returned by the algorithm as the prediction. An example classification tree is shown in figure 3.1.3.

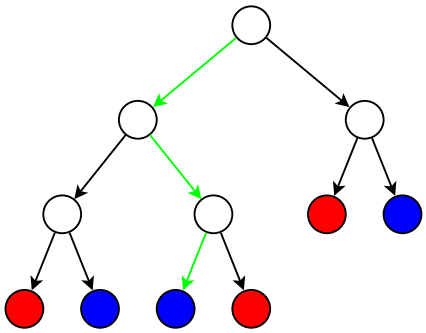


Figure 3.1.3: Classification tree example. Starting at the root, a comparison is made at each internal (white) node to decide which direction to follow. Once a leaf node is reached, the label of the leaf node becomes the classifier output. In this example green arrows are followed and the output is the blue label.

The choice of decision rules at internal nodes depends on the type of predictors available (discrete-valued or continuous-valued). In the continuous case usually one of the components of $\mathbf{x}$ is compared to a threshold value and depending on the result of the comparison we move towards either of the two child nodes.

Given a set of data with labels $\{\mathbf{x}_i, y_i\}, \mathbf{x}_i \in \mathbb{R}^p$, a classification tree is learned by recursively partitioning the data into two subsets. The partition of a node m into two children is defined by a feature index $r_m^\star$ and a threshold $t_m^\star$. One child node is assigned those samples for which the $r_m^\star$-th feature is less than or qeual to $t_m^\star$ and the other child node is assigned the remaining samples. The parameters $r_m^\star$ and $t_m^\star$ are chosen to maximise the reduction in some measure of node impurity (discussed below), weighted by the number of samples assigned to each node.

Let us define S_m to be the set of samples at node m, with cardinality N_m, and $\hat{y}_m$ to be the most common label among the samples at node node m. Also let us define q_{mk} to be the fraction of samples at node m that have the label k, *i.e.* $q_{mk} = \frac{1}{N_m} \sum_{\mathbf{x}_i \in S_m} I(y_i = k)$ with $I(\cdot)$ the indicator function. Then the following are definitions of some common measures used for deciding node partitions [**123**].

$$\begin{aligned}
\text{Misclassification error}: &\quad \frac{1}{N_m} \sum_{i \in S_m} I(y_i \neq \hat{y}_m) = 1 - q_{m\hat{y}_m} \\
\text{Gini index}: &\quad \sum_{k \neq k'} q_{mk} q_{mk'} = \sum_{k=1}^{K} q_{mk}(1 - q_{mk}) \\
\text{Cross} - \text{entropy}: &\quad -\sum_{k=1}^{K} q_{mk} \log(q_{mk})
\end{aligned}$$

The parameters $r_m^\star$ and $t_m^\star$ for each node m are computed by exhaustive search as follows. For each admissible feature index r_m the algorithm selects a threshold by iterating over the values of the r_m-th feature for all samples available at node m and choosing the value that results in the lowest weighted node impurity of the children of m. Then $r_m^\star$ is selected as the index of the feature which results in the lowest

weighted node impurity of the children of m, with its optimal threshold, and this optimal threshold becomes $t_m^\star$ [**123**].

The main weaknesses of classification trees include difficulty in capturing additive structure and instability (small changes in the training data can lead to very different trees) [**123**]. In order to reduce overfitting, trees can be regularised in several ways, such as limiting the depth of the tree, setting a minimum number of instances that must be available at a node so that it can be partitioned, or setting a maximum number of leaf nodes [**225**].

A random forest [**24**] consists of a collection of trees learned with randomised selection of samples and predictors. Each tree is learned from a random subset of available training samples. In addition, each node splitting rule can only have its decision variable chosen from a random subset of predictors (these subsets are sampled independently for each node of each tree). The decision of the whole forest is taken as the vote among all its trees.

Typically the size of the set of predictors randomly chosen at each node to search for an optimal (with respect to this random subset) partition is $p' = \sqrt{p}$. If the number of predictors is very large and $\sqrt{p} \ll p$, it may happen that the most valuable predictors are rarely selected and performance is poor. In this case it may be beneficial to increase p'.

The technique of randomly sampling "new" datasets from the training data is known in statistics as bootstrapping [**123**]. When outputs of models built on multiple bootstrap samples are used in combination, this is referred to as bootstrap averaging, or "bagging" for short. In the case of classification, averaging means a majority vote.

Random forests rely on bootstrapping and in addition they select from randomised subsets of features when searching for best node partitions [**123**]. Those two methods work in tandem to decorrelate the trees, increasing diversity and reducing the risk of overfitting which is a problem when only a single decision tree is used. [**123**].

Increasing the number of trees in a random forest is generally beneficial in terms of accuracy, but there is a point where adding more trees only has a small effect. The problem of selecting a near-optimal forest size was addressed by Latinne *et al.* [**157**] who propose a methodology where the McNemar test is used to decide when to stop adding more trees. Oshiro *et al.* [**199**] tested several forest sizes on a collection of machine learning problems, giving an indication of how much forest size has to be increased for a significant difference in performance to be observed.

3.2. Dimensionality reduction

When fitting high-dimensional models, as is the case with imaging data, model overfitting becomes a concern, as discussed in sub-section 3.1.2. As an alternative to regularisation methods discussed there, another approach to minimising overfitting is to reduce feature spaces to more relevant sets and build models from reduced features. Guyon [**116**], Saeys [**223**] and Bolon-Canedo [**20**] provide systematic overviews of these methods and their relative merits. In the following, a summary of algorithms relevant to this dissertation is presented.

3.2.1. Feature selection. Feature selection is a process of filtering the feature set to retain only a subset, while keeping the features in their original form. At its simplest, this can be done by examining the features individually and deciding whether they contain useful infomation about the prediction target. More sophisticated methods consider correlations between features or reduce the feature set with feedback about the performance of the classification model. The latter can be done either by an explicit mathematical formulation of the model cost function or algorithmically. Feature selection methods can be divided into three groups: filter, wrapper and embedded [**116, 223, 20**].

Filter methods. These methods process the features without any feedback from the classification algorithm [**20**]. A simple, mass univariate approach of this type

consists of applying a separate statistical test to each feature to decide whether the feature has different distributions depending on the label of the image. The ANOVA method [**113**], where the F-statistic is computed for each feature and used to rank the features, allowing a subset of most relevant features to be selected, is an example.

Saeys *et al.* [**223**] discuss several filter-type feature selection methods (alongside wrapper and embedded methods), including univariate statistical tests, correlation-based feature selection (CFS) [**120**], fast correlation-based feature selection (FCBF) [**274**] and the Markov blanket filter (MBF) [**149**]. Bolon-Canedo *et al.* [**20**] compare the performance of a large number of filter methods (alongside wrapper and embedded methods), including CFS [**120**], the consistency-based filter [**65**], the INTERACT algorithm [**282**], information gain [**121**], ReliefF [**150**], minimum redundancy maximum relevance (mRMR) [**204**] and the $\mathcal{M}_d$ filter [**227**]. The main advantage of filter methods is their efficiency and scalability [**223**]. The main drawback is that this approach ignores interaction with the classifier [**223**].

Wrapper methods. This subset consists of algorithms that select features based on performance of the classification algorithm. It is usually not feasible to fit the classifier to all possible combinations of features so a greedy search strategy is often used, either by forward selection (starting with an empty set and iteratively adding the most promising features) or by backward elimination (starting with all features and iteratively eliminating the least useful ones) [**116**]. Saeys *et al.* [**223**] list several more sophisticated methods, including "plus q take-away r", beam search, simulated annealing, randomised hill climbing, genetic algorithms and estimation of distribution algorithms.

Embedded methods. This subset comprises methods designed to perform feature selection as part of the classifier fitting process. In particular, it includes regularisation methods based on adding a sparsity-inducing norm as a term in the classifier's cost function.

For example, sparsity-inducing ℓ_1 regularisation can be applied to logistic regression [**102**] or substituted for the standard ℓ_2 regularisation in support vector machines [**285**]. ℓ_0 regularisation, which suffers from poor tractability, can be approximated with methods such as the one proposed by Westion *et al.* [**265**]. Since these methods select features simultaneously while fitting the model, they can detect interactions between features. However, one disadvantage is that ℓ_0 or ℓ_1 regularisation can lead to overly simplistic models, where only one feature from each set of correlated features is selected, mislading the analyst into concluding that its correlates are unimportant. One way of addressing this problem is to use mixed ℓ_1-ℓ_2 regularisation (Elastic Net) [**288**].

Random Forests can also be used for feature selection by computing feature importance scores during the fitting process. For each individual tree, the out-of-bag (OOB) instances, which were not used to build that particular tree, are classified and classification accuracy is recorded. To compute the importance of a specific feature, the values of this feature are randomly permuted across the OOB samples, the OOB samples are classified again and accuracy is recorded. The decrease in accuracy due to random permutation, averaged over all trees, is a measure of feature importance (variable importance) [**24, 123**].

The problems with regularisation methods can also be addressed to some extent with stability selection, proposed by Meinshausen and Buhlmann [**191**]. Stability selection fits a regularised model many times with randomisation (achieved with repeated sub-sampling of the data), and features are selected based on how often they appear in the resulting models [**191**]. Stability selection offers some interesting theoretical guarantees [**191**]. The main problem is the increased computational cost of fitting a large number of randomised models.

3.2.2. Feature agglomeration. Feature agglomeration is an alternative to feature selection. It relies on clustering correlated features together and then transforming each cluster into an agglomerate feature. Several such methods have emerged

in the field on natural language processing, where they were applied to bag-of-words models [**75**]. Any standard clustering algorithm (*e.g.* K-means or hierarchical clustering [**123**]) can in principle be adapted for feature agglomeration by simply transposing the data matrix before feeding it to the algorithm, so that features are clustered instead of samples. Once clusters of features are identified, feature agglomeration merges each cluster into a single feature with a pooling operation, such as summation or averaging (averaging was used in this dissertation).

In the context of image classification, if the pixels (or voxels) are thought of as features then image segmentation can be thought of as a type of feature clustering. This means that for image classification tasks the library of feature clustering methods can be extended with suitable segmentation algorithms.

K-means clustering. One of the possible formulations of the clustering problem is minimum sum-of-squares clustering (MSSC) [**6**], which is defined by the following cost function [**123**].

$$W(C) = \sum_{k=1}^{K} N_k \sum_{C(i)=k} \|\mathbf{x}_i - \bar{\mathbf{x}}_k\|_2^2 \tag{3.2.1}$$

where $C(i)$ is the cluster assignment function that returns the cluster to which i-th sample is assigned, N_k is the number of samples in k-th cluster and $\bar{\mathbf{x}}_k$ is the mean of the k-th cluster.

The K-means clustering algorithm, which finds a local minimum of 3.2.1, is initialised with a set of points that are an initial guess of cluster means, often chosen as randomly selected samples from training data [**123**]. Then the algorithm alternates between two steps:

(1) Compute the mean of each cluster
(2) Reassign each observation to the nearest cluster (using Euclidean distance).

This algorithm converges to a local minimum of the cost function 3.2.1, but the result may differ from the global minimum [**123**]. Therefore, starting K-means

repeatedly with different random initialisations and choosing the solution with the lowest cost function is recommended [**123**]. Finding a global minimum directly is infeasible, since the problem of minimising the cost function 3.2.1 is known to be NP-hard [**6**].

Choosing a suitable value for K is also an important part of the problem. It can be addressed by plotting the optimal value of the cost function for successive values of K and looking for a "kink" point where further increasing K gives only a small reduction in the cost function [**123**].

Agglomerative clustering. Agglomerative clustering starts with each observation as its own cluster and iteratively searches for a pair of clusters that are the most beneficial to merge according to some predefined criterion and merges them. With each iteration the number of clusters decreases by one and the process results in a tree-like hierarchy of clusters, with individual observations as leaves and the whole data set combined at the root (figure 3.2.1) [**123**]. In this dissertation the criterion used for selecting pairs of clusters to merge is minimising the increase in the sum of within-cluster variances as a result of merging these clusters, as originally proposed by Ward [**142**].

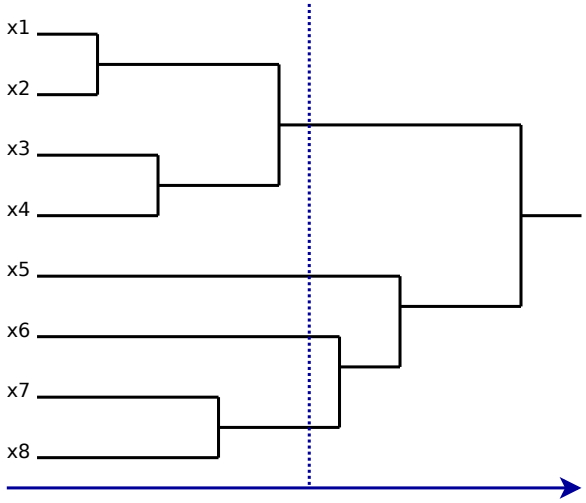


Figure 3.2.1: Agglomerative clustering. Initially each feature is considered a separate cluster (left) and clusters are iteratively merged (moving towards the right). The dotted vertical line indicates the solution for four clusters.

When feature agglomeration algorithms are applied to imaging problems, one may want to avoid clustering together voxels that are spatially distant from one another. To address this issue, spatial distances between clusters should be considered as well as their statistical similarities, making the algorithm spatially aware. Spatial neighbourhood constraints may also help to reduce the problems associated with multiple testing: if for each cluster only its neighbours are considered for merging, then false cluster assignments would seem less likely than when evaluating pairs of clusters from the whole image. A spatially aware Ward-type feature agglomeration algorithm was proposed by Michel *et al.* for fMRI analysis[**192**].

Chapter 5 describes two feature agglomeration algorithms in detail: one based on agglomerative clustering with spatial constraints (a simplified version of [**192**]) and one based on SLIC, which is an image segmentation algorithm. They are both evaluated, in combination with classification algorithms, on an image classification task.

3.2.3. Principal component analysis and manifold learning. Principal Component Analysis (PCA) computes a transformation that represents the data in a new basis. The vectors of this new basis are ordered in a sequence such that the first one is the direction of maximum sample variance in the feature space and each subsequent vector is the direction of maximum sample variance "subject to being orthogonal to the earlier ones" [**123**].

In PCA each instance $\mathbf{x}_i \in \mathbb{R}^p, i = 1, \ldots, n$ is represented by a vector $\lambda_i \in \mathbb{R}^q$ such that an approximation to $\mathbf{x}_i$ is constructed as [**123**]

$$\tilde{\mathbf{x}}_i = \bar{\mathbf{x}} + \mathbf{V}_q \lambda_i.$$

where $\bar{\mathbf{x}}$ is the mean of the instances [**123**]:

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i.$$

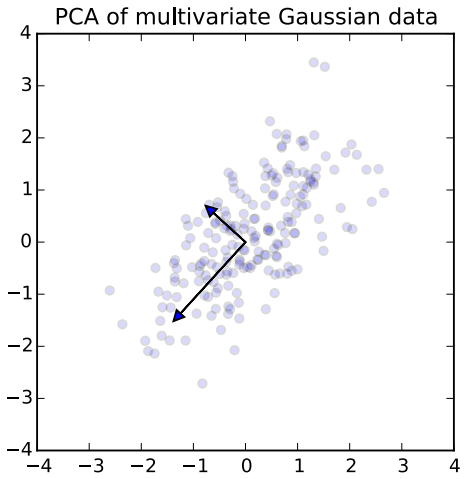


Figure 3.2.2: PCA example. Principal component analysis of data generated from a multivariate Gaussian distribution. The arrows are the two principal components.

The matrix $\mathbf{V}_q$ and the vector λ_i are computed as follows. First the data are mean-centered and arranged in a matrix $\mathbf{X}_c \in \mathbb{R}^{n \times p}$ such that each row of $\mathbf{X}_c$ is $\mathbf{x}_i^T - \bar{\mathbf{x}}$. Then the singular value decomposition (SVD) of $\mathbf{X}_c$ is computed [**123**]:

$$\mathbf{X}_c = \mathbf{U}\mathbf{D}\mathbf{V}^T.$$

Then for each q, the matrix $\mathbf{V}_q$ is constructed from the first q columns of $\mathbf{V}$. The vectors λ_i are the rows of the matrix $\mathbf{U}_q\mathbf{D}_q \in \mathbb{R}^{n \times q}$ [**123**].

Manifold learning [**253**] is a family of non-linear dimesionality reduction methods. Manifold learning algorithms reproduce in a low-dimensional space the local structure of the data in the original high-dimensional space.

For example, the Isomap algorithm [**245**] first constructs a neighbourhood graph G, connecting each point to its k nearest neighbours in the high-dimensional space, with edge weights representing pairwise distances. Then it computes pairwise distances for all points using a shortest-path algorithm over G. Finally, it optimises the distribution of data points in the low-dimensional space so that pairwise Euclidean distances are an optimal approximation of the shortest paths over G. A variety of manifold learning algorithms is reviewed and compared by van der Maaten in [**253**],

including Isomap, Laplacian Eigenmaps [**17, 18**] and Locally Linear Embedding (LLE) [**219**], among other methods.

3.3. Randomised projections

The Johnson-Lindenstrauss lemma [**140**] shows that a set of vectors can be embedded into a space of sufficiently large dimension in a way such that pairwise distances between these vectors are only distorted to a very small degree.

THEOREM 6. *(Johnson and Lindenstrauss* [**140**]*, citing from Baraniuk, Davenport, DeVore and Wakin* [**15**]*).*

"Let $\epsilon \in (0,1)$ *be given. For every set* Q *of* $\#(Q)$ *points in* $\mathbb{R}^N$*, if* n *is a positive integer such that* $n > n_0 = O\left(ln\left(\#(Q)\right)/\epsilon^2\right)$*, there exists a Lipschitz mapping* $f : \mathbb{R}^N \mapsto \mathbb{R}^n$ *such that*

$$(1-\epsilon)\left\|\mathbf{u}-\mathbf{v}\right\|_{\ell_2^N}^2 \leq \left\|f(\mathbf{u})-f(\mathbf{v})\right\|_{\ell_2^n}^2 \leq (1+\epsilon)\left\|\mathbf{u}-\mathbf{v}\right\|_{\ell_2^N}^2$$

for all $\mathbf{u}, \mathbf{v} \in Q$*."*

The notation $f(n) = O(g(n))$ is defined as follows.

DEFINITION 7. Big-O notation (citing from Constantinides [**58**]).

"Let f and g be functions from the set of integers or the set of reals to the set of reals. The function $f(x)$ is $O(g(x))$ if and only if

$$\exists c \in \mathbb{R}^+ \exists k \in \mathbb{R}^+ \forall x \left((x > k) \Rightarrow \left(|f(x)| \leq c\,|g(x)|\right)\right)."$$

A Lipschitz function is defined as follows.

DEFINITION 8. Lipschitz function (citing from [**127**])

"A function $f : A \to \mathbb{R}^m$, $A \subset \mathbb{R}^n$, is said to be L-Lipschitz, $L \geq 0$, if

$$|f(a) - f(b)| \leq L|a - b|$$

for every pair of points $a, b \in A$. We also say that a function is *Lipschitz* if it is L-Lipschitz for some L"

The Johnson-Lindenstrauss lemma proves the existence of a mapping but it does not say how it can be constructed. Several researchers have proposed constructions of randomised projections that can be shown to have similar bounds on distortion [**101, 134, 2, 64, 15**]. These contributions also include more precise bounds on n. In particular, Dasgupta and Gupta [**64**] derived the bound $n \geq 4\ln(\#(Q))/(\epsilon^2/2 - \epsilon^3/3)$ for matrices that implement projections onto random subspaces.

The main disadvantage of projections with unstructured random matrices is that the computation of such projections is still $O(nN)$ for each input vector. This can be improved with the fast Johnson-Lindenstrauss transform (FJLT) proposed by Ailon and Chazelle [**4**] and further simplified by Ailon and Liberty [**5**]. This transform can be written as

$$f(\mathbf{x}) = \Phi D\mathbf{x}$$

where the rows of Φ are drawn uniformly at random from a Hadamard matrix and D is a diagonal matrix with each diagonal element drawn from the set $\{-1, 1\}$ with a uniform probability distribution (D is generated only once and used for all input vectors).

The main advantage of this method over PCA is that each feature vector can be processed individually, without inspecting any other instances. Only the overall number of instances has to be known in advance, in order to decide a suitable dimension for the reduced feature space. The computational complexity of the fast Johnson-Lindenstrauss transform is dominated by the Hadamard transform, which

can be computed in $O(N \log N)$ time using an algorithm similar to the Fast Fourier Transform.

3.4. Measuring machine learning performance

3.4.1. Performance measures. When building machine learning models, it is important to be able to estimate their performance, compare them and select the best performing ones. For supervised learning models this is typically done by estimating the expected error when the model is used to predict its target variable for new (unseen) data.

For classification models the intuitive performace figure is *classification accuracy*:

$$Acc = \frac{1}{n} \sum_{i=1}^{n} I\left(y_i = g\left(\mathbf{x}_i\right)\right)$$

but in some cases accuracy can be miselading. Specifically, this happens when one label is far more common than the others. In this case, a classifier that always predicts the most common label can have a very high accuracy.

Restricting the discussion to binary classification with labels "0" and "1", two additional measures that are often quoted are *sensitivity* and *specificity*. In order to define them, let us first introduce the notions of a true positive (TP), true negative (TN), false positive (FP) and false negative (FN) as shown in table 3.4.1. A true positive occurs when the true label is "1" and "1" is predicted. A true negative occurs when the true label is "0" and "0" is predicted. A false positive is the case when "0" is the true label but "1" is predicted. Finally, a false negative is the case when "1" is the true label but "0" is predicted.

Table 3.4.1: Definition of true positive (TP), true negative (TN), false positive (FP) and false negative (FN)

		Predicted label	
		0	1
True label	0	TN	FP
	1	FN	TP

Sensitivity measures how good the classifier is at identifying positives as such:

$$Sens = \frac{TP}{TP + FN}$$

Specificity measures how good it is at identifying negatives as such:

$$Spec = \frac{TN}{TN + FP}$$

Some classifiers, such as logistic regression, can output estimated class probabilities instead of labels. Given these probabilities, it is possible to adjust the probability threshold for classifying a sample as positive. Each of the possible thresholds is associated with a pair of values for sensitivity and specificity of the resulting classifier. If sensitivity is plotted against one minus specificity, the resulting curve is known as the receiver operating characteristic (ROC). The area under the ROC curve (AUC) can be calculated and is also a common measure of classifier performance [**100**]. An example is shown in figure 3.4.1.

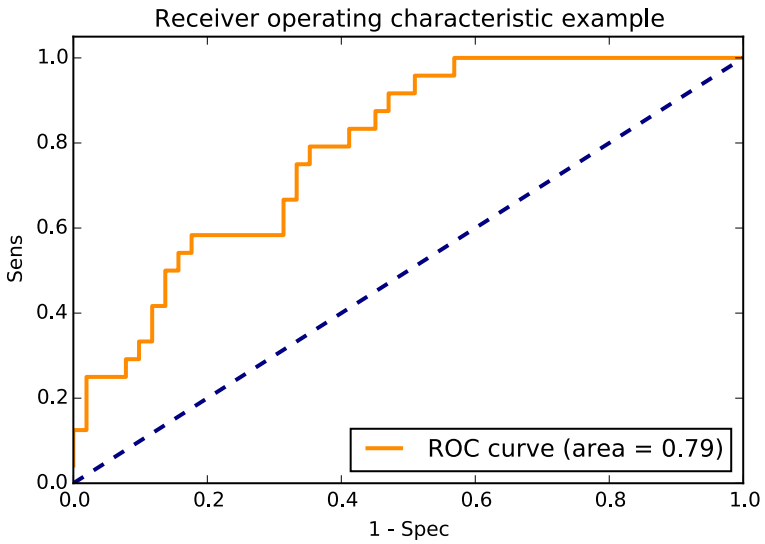


Figure 3.4.1: Receiver operating characteristic (ROC) curve. (Image generated with Scikit-learn example code.)

3.4.2. Data partitioning and cross-validation. Due to the problem of overfitting discussed in sub-section 3.1.2, machine learning error measures computed on

training data tend to be over-optimistic compared to the performance attained on unseen data [**123**].

Therefore, in order to have a reasonable estimate of performance, it is necessary to set a proportion of the available data aside for testing, ensuring that this data is not used in the training process.

In some cases an estimate of performance is required in order to adjust model parameters. In this case a validation set is partitioned out of the training data. A collection of models learned from the training data are compared with regards to their performance on the validation set and possibly other metrics such as sparsity. The final model is selected based on this comparison. In this case it is important to recognise that validation error is also optimistic, and that a separate test set is required that is not available in training and valiadation [**123**].

Often the data available for training, validation and testing is limited. In this case it is important to use it in the most efficient way possible. The cross-validation technique provides a way of doing that. K-fold cross-validation requires partitioning the data into K equal-sized subsets (called folds). Each fold is then used as a validation set for a classifier or regressor learned from the combined remaining folds. This gives K performance figures which can be combined (*e.g.* by averaging) to yield a single estimate.

An alternative to cross-validation is to generate a set of independent random partitions of the data into training and test sets. The disadvantage of this method is that the test sets from different partitions are likely to overlap and some instances may appear in none of the test sets. This problem does not appear with cross-validation, where each instance is used for testing exactly once.

3.5. Conclusions

This chapter is a discussion of the machine learning algorithms that are relevant to this dissertation. It provides essential background on classification algorithms that are used in chapters 5 and 6, including logistic regression, support vector machines (SVM) and random forests. Several methods for dimensionality reduction, which are used extensively in chapter 5, are also dicussed. The distinction between dimensionality reduction methods based on feature selection, feature agglomeration, principal component analysis and manifold learning is highlighted. Different types of feature selection methods are compared, including filter, wrapper and embedd methods. Randomised projections, which are used in chapter 6, are discussed. Methods for evaluating performance of machine learning algorithms are also discussed, which is important because data partitioning and cross-validation will be used to estimate the performance of the proposed algorithms throughout this dissertation.

CHAPTER 4

Wavelet packet basis learning for compressed sensing

An earlier version of this chapter was presented as a workshop paper at the 2012 MICCAI Workshop on Sparsity Techniques in Medical Imaging [**217**]. *The writing has been revised and extended with additional data for inclusion in this dissertation.*ś

4.1. Introduction

Compressed sensing (CS), as discussed in chapter 2, provides a mathematical framework for reconstruction of signals sampled at sub-Nyquist rates, provided that those signals can be represented sparsely with either an orthogonal transform or a dictionary of signals [**40, 82, 35**]. CS led to recent advances in medical imaging, and in particular in magnetic resonance imaging (MRI), starting with the work of Lustig *et al.* on structural MRI [**183, 181, 182**]. Lustig *et al.* showed that CS can be used to acquire MR images in much shorter times by reducing the full set of phase encodes to a randomised subset and solving the resulting underdetermined inverse problem by minimising the ℓ_1 norm of the wavelet representation of the image, with an additional total variation term, subject to the Fourier representation of the image being consistent with the acquired phase encodes. They also showed that for MR angiography (imaging of blood vessels) compressed sensing can be used without representing images with the wavelet transform since these images are sparse in image domain.

Compressed sensing MRI (CS-MRI) was extended to dynamic MRI by relying on Fourier-sparsity of the temporal view of voxels [**184, 104**] and these methods were subsequently improved by incorporating motion prediction and correction algorithms

[**143, 252**]. CS-MRI was also combined with parallel MRI (using spatially sensitive RF coils) to further accelerate image acquisition [**169, 166, 200, 256**].

While in most imaging applications wavelets provide accurate sparse approximations of signals, there is also much interest in finding representations adapted to particular signals. Examples include patch-based dictionaries [**3, 213**], as well as dictionaries adapted specifically for compressed sensing [**84**]. Patch-based dictionaries were applied to CS-MRI for static [**19, 211, 212, 12, 175, 214, 236**] and dynamic CS-MRI [**11, 167, 168, 261, 31, 260**].

Similarly, the adaptive signal representation framework of wavelet packets [**55, 56**] was applied to CS by Peyre [**206, 207**]. While patch-based dictionaries and wavelet packets are similar in that both approaches adapt a set of atoms to efficiently represent training signals, they are also very different in other aspects. In particular, wavelet packets provide a basis for the whole image, rather than small patches. They also naturally have a multiscale structure and representation coefficients can be computed efficiently using Mallat's mutiresolution analysis (MRA) algorithm [**188**].

The work in [**206, 207**] explores finding the best basis while reconstructing the undersampled signal, *i.e.* without learning from prior examples. In contrast, the algorithm proposed in this work learns an adapted basis from a collection of example images. The proposed method is based on well-known principles and algorithms for wavelet packets and approximation in bases [**55, 56, 210, 266, 187**].

The main contribution of this work consists of designing a basis search cost function that includes the criteria that are important in compressed sensing. A suitable algorithm is also selected for optimising this cost function and the proposed method is tested on images from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [**197**] database.

4.2. Compressed sensing

To reiterate from chapter 2, a general model of a compressed sensing system can be written as

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{z} \tag{4.2.1}$$

where $\mathbf{x}$ is the sparse representation vector of the signal, $\mathbf{A}$ is the information operator that transforms signals from the sparse representations to the physical measurements, and $\mathbf{z}$ represents random noise inherent in instrumentation. The key property of compressed sensing is that $\mathbf{A}$ has more columns than rows, *i.e.* the number of measurements in $\mathbf{y}$ is less than the number of components of the sparse representation vector $\mathbf{x}$. In compressed sensing MRI, typically $\mathbf{A} = \mathbf{RFD}$, where $\mathbf{D}$ is a basis (or dictionary) that is used to sparsely represent the image, $\mathbf{F}$ is a multidimensional Fourier transform that models the MRI acquisition process, and $\mathbf{R}$ is a matrix that selects a subset of measurements. The estimate $\hat{\mathbf{x}}$ of $\mathbf{x}$ is computed by solving the problem [**39, 35**]

$$\min_{\tilde{\mathbf{x}} \in \mathbb{R}^n} \|\tilde{\mathbf{x}}\|_1 \quad s.t. \quad \|\mathbf{y} - \mathbf{A}\tilde{\mathbf{x}}\|_2 \leq \varepsilon \tag{4.2.2}$$

where ε^2 is an upper limit on the noise energy $\|\mathbf{z}\|_2^2$. If $\mathbf{A}$ satisfies the restricted isometry property [**37**] with an appropriate isometry constant, then the solution $\hat{\mathbf{x}}$ to the problem (4.2.2) is such that [**39, 35**]

$$\|\hat{\mathbf{x}} - \mathbf{x}\|_2 \leq C_0 \frac{\|\mathbf{x} - \mathbf{x}_s\|_1}{\sqrt{s}} + C_1 \varepsilon \tag{4.2.3}$$

for some constants C_0 and C_1, where $\mathbf{x}_s$ is the best s-sparse approximation of $\mathbf{x}$ (best approximation with s non-zero components).

One of the main design goals in compressed sensing is to find a dictionary that leads to the sparsest representations possible, so that accurate approximations can be constructed with small s. At the same time, equation (4.2.3) implies that the modelling error induced by neglecting small coefficients will degrade the quality of

reconstruction. The objective of this chapter is to design an algorithm for finding a wavelet packet basis that minimises the approximation error for a specified level of sparsity.

4.3. Wavelet packets

Wavelet packets [**55, 56**] are an extension of wavelets. To reiterate from chapter 2, a single-level discrete wavelet transform consists of filtering the source signal with an orthogonal filter bank, followed by downsampling of the resultant sequences. A multi-resolution analysis (wavelet decomposition) then consists of recursively applying this process on the lowpass branch of the filter bank, producing a tree-like structure. With wavelet packets, the filter bank can be applied to the high-pass branches as well, which means that a variety of filter trees can be built (an example wavelet packet tree is shown in figure 4.3.1). All possible choices of which branches to decompose further give rise to a large set of admissible trees [**187**], or in other words "a library of bases" [**79**]. The leaves of each admissible tree define a complete set of basis vectors.

Wavelet packets bases are a superset of wavelet bases. Therefore, with a careful choice of basis from a wavelet packet library, one may be able to find sparser approximations than with the standard wavelet basis. The Coifman-Wickerhauser (CW) basis selection algorithm [**56**] can efficiently find an optimal tree, in the sense that the associated basis minimises a cost function with respect to the resultant representation vector, provided that the cost function satisfies the criterion of an *additive information cost function*:

Definition 9. Additive information cost function (quoting from [**56**]):

> "A map $\mathcal{M}$ from sequences $\{x_i\}$ to $\mathbb{R}$ is called an *additive information cost function* if $\mathcal{M}(0) = 0$ and $\mathcal{M}(\{x_i\}) = \sum_i \mathcal{M}(x_i)$."

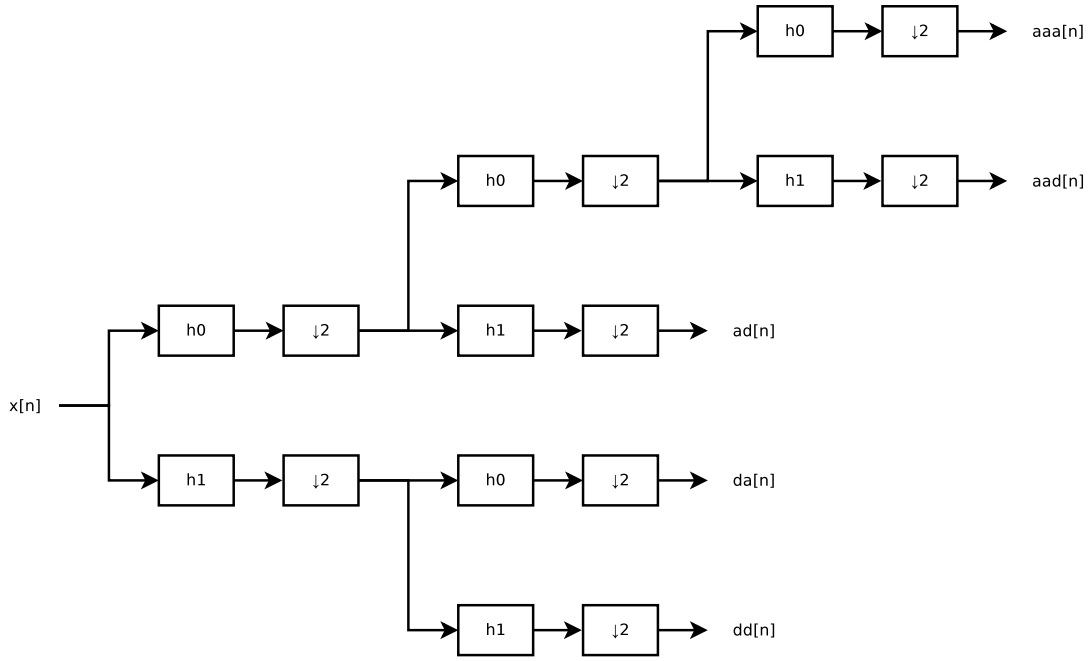


Figure 4.3.1: A wavelet packet tree. This particular tree differs from the standard wavelet transform by applying an additional iteration of wavelet decomposition to the high-pass output of the first pair of filters.

The Coifman-Wickerhauser basis selection algorithm consists of two major steps. (1) The input signal is decomposed into its wavelet packet tree and the data vector at each node is recorded. (2) The information cost associated with each node is computed as the minimum of two values: the information cost of its coefficient vector and the sum of information costs of its children (computed recursively using the same rule). If a node's coefficient vector has lower information cost than the sum of information costs of its children, it becomes a leaf node. Otherwise, its children are included in the optimal tree [**56**].

4.4. Optimised wavelet packet bases and compressed sensing

4.4.1. Single signal case. The analysis in this section is based on [**187**] (pp. 450-452, 611-614), adapted for approximation error measured in ℓ_1 norm.

Given a vector (signal) $\mathbf{y} \in \mathbb{R}^n$, basis search can be expressed as minimising the following Lagrangian with respect to the wavelet packet basis $\mathbf{B}_{wp}$ and the vector

of approximation coefficients $\mathbf{x}_\Lambda \in \mathbb{R}^n$.

$$\mathcal{L}\left(y, \mathbf{B}_{wp}, \tau, x_\Lambda\right) = \left\|\mathbf{B}_{wp}^T\mathbf{y} - \mathbf{x}_\Lambda\right\|_1 + \tau \left\|\mathbf{x}_\Lambda\right\|_0 \tag{4.4.1}$$

where τ is the Lagrangian multiplier. $\|.\|_0$ denotes the ℓ_0 pseudo-norm and $\mathbf{A}^T$ denotes the transpose of matrix $\mathbf{A}$ (we consider real-valued images only). A predefined pair of filters is assumed (low-pass and high-pass) that is used at each branching in the wavelet packet tree.

The decision to minimise the transform domain ℓ_1 error (instead of the more common ℓ_2 error) is based on equation (4.2.3), where $\|\mathbf{x} - \mathbf{x}_s\|_1$ appears in the numerator of one of the terms of the upper bound on the compressed sensing reconstruction error.

Consider the parameter τ in eq. (4.4.1). Since $\mathbf{B}_{wp}$ is orthonormal, each component of $\mathbf{B}_{wp}^T\mathbf{y}$ can be treated independently when searching for an optimal approximation. If a component's ℓ_1 norm is less than τ then the value of the cost function can be reduced by setting that component to zero because the reduction in the sparsity term $\tau \left\|\mathbf{x}_\Lambda\right\|_0$ will outweigh the increase in the error term $\left\|\mathbf{B}_{wp}^T\mathbf{y} - \mathbf{x}_\Lambda\right\|_1$. Therefore, τ is also the threshold at which the entries of $\mathbf{x} = \mathbf{B}^T\mathbf{y}$ should be set to zero in $\mathbf{x}_\Lambda$ (see [**187**] pp. 612-613 for a more formal argument with error defined in ℓ_2 norm). Therefore, equation (4.4.1) can be written in alternative form

$$\mathcal{L}\left(\mathbf{y}, \mathbf{B}_{wp}, \tau\right) = \sum_i \inf\left(\left|\mathbf{x}\left[i\right]\right|, \tau\right) \quad \text{where} \quad \mathbf{x} = \mathbf{B}^T\mathbf{y} \tag{4.4.2}$$

where $\mathbf{x}\left[i\right]$ denotes the i-th component of the vector $\mathbf{x}$.

If τ is fixed, the cost function in eq. (4.4.2) is a valid additive information cost function, so the Coifman-Wickerhauser algorithm can be used for best basis search.

The form of $\mathcal{L}$ in eq. (4.4.2) is independent of $\mathbf{x}_\Lambda$. This is a result of the fact that $\mathbf{B}_{wp}$ is an orthonormal basis: by fixing τ and selecting $\mathbf{B}_{wp}$, we implicitly assign $\mathbf{x}_\Lambda = \rho_\tau\left(\mathbf{B}_{wp}^T\mathbf{y}\right)$ where $\rho_\tau\left(\mathbf{a}\right)$ is an operator that sets to zero all components of the vector $\mathbf{a}$ that are smaller than τ.

The discussion above is concerned with optimal approximations for any given coefficient threshold τ but this threshold is difficult to choose. It is more practical to specify either the number of coefficients or the approximation error. A closely related problem of finding best wavelet packet bases in rate-distortion sense for source coding applications was considered in [**210**]. In the following, a similar approach is applied to the compressed sensing basis search problem.

Consider two solutions to minimising $\mathcal{L}$, with respective thresholds $\tau_1 < \tau_2$. The basis $\mathbf{B}_1$, optimised for $\tau = \tau_1$, will favour accuracy at the cost of sparsity, while the trade-off will move in the opposite direction for the basis $\mathbf{B}_2$, optimised for $\tau = \tau_2$. Since both bases are optimal at their respective thresholds, we have $\left\|\mathbf{B}_1^T\mathbf{y} - \rho_{\tau_1}\left(\mathbf{B}_1^T\mathbf{y}\right)\right\|_1 \leq \left\|\mathbf{B}_2^T\mathbf{y} - \rho_{\tau_2}\left(\mathbf{B}_2^T\mathbf{y}\right)\right\|_1$ and $\left\|\rho_{\tau_1}\left(\mathbf{B}_1^T\mathbf{y}\right)\right\|_0 \geq \left\|\rho_{\tau_2}\left(\mathbf{B}_2^T\mathbf{y}\right)\right\|_0$. So if for any value of τ the approximation is not sparse enough, τ can be increased and basis search repeated to improve sparsity at the cost of accuracy. Similarly, if the approximation is not accurate enough, accuracy can be improved at the cost of sparsity by reducing τ and repeating the basis search. Therefore, the optimal threshold for a required level of sparsity or approximation error can be found by bisection search.

4.4.2. Extension to multiple signals. In compressed sensing MRI applications, the signal to be sparsely approximated is not fully known, since we only have partial Fourier data. Instead, a collection of images of similar anatomy may be available that can be used as a substitute in the basis search process. In this context, the objective is to find a basis that minimises the expected value of the information cost over the training set. Therefore, the cost function is the mean information cost of the training data, *i.e.*

(4.4.3)
$$\bar{\mathcal{L}}\left(\{\mathbf{y}_i\}_{i=1,\ldots,N}, \mathbf{B}_{wp}, \tau\right) = \frac{1}{N}\sum_{i=1}^{N}\left(\left\|\mathbf{B}_{wp}^T\mathbf{y}_i - \rho_\tau\left(\mathbf{B}_{wp}^T\mathbf{y}_i\right)\right\|_1 + \tau\left\|\rho_\tau\left(\mathbf{B}_{wp}^T\mathbf{y}_i\right)\right\|_0\right)$$

where $\mathbf{y}_i$ are the training samples and N is the number of signals in the training set. Since $\rho_\tau\left(\mathbf{B}_{wp}^T\mathbf{y}_i\right) = (\mathbf{x}_\Lambda)_i$, this is the mean of the single-image cost function (4.4.1) over all training samples. This cost function can be optimised by decomposing each signal in a full wavelet packet tree, constructing a joint tree [**266**] where each node is assigned a cost by taking the average of the same node over all individual image trees, and then applying the Coifman-Wickerhauser algorithm to the joint tree to find an optimal basis [**266**].

Note that the same threshold τ is used for wavelet packet decompositions of all images. It is therefore necessary to ensure that images are pre-processed so that one value of τ is suitable for all of them. A simple approach to solving this problem is to scale image intensities so that *e.g.* the 99^{th} percentile of image histograms matches across the whole dataset (scaling images to match their maximum intensities is somewhat less robust to outliers).

4.4.3. Compressed sensing reconstruction. The proposed method finds a sparsifying basis that can be used with standard compressed sensing reconstruction algorithms. In this work it is integrated with the SparseMRI software [**181**]. The main change is replacing the wavelet transform with the optimised wavelet packet tree transform that was learned from the training data.

4.5. Experiments

Experiments were conducted on a data set consisting of 826 MR images from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database [**197, 135**] (`http://www.loni.ucla.edu/ADNI`). Details of the image acquisition process were described by Jack *et al.* [**135**]. These images had some pre-processing applied by ADNI, including GradWarp, B1 non-uniformity correction and N3 intensity correction, depending on the scanner manufacturer. Some further processing was applied to these images, as described in the following. The images were resliced to

1mm voxel size using the tools from J. Shen's NIFTI toolbox, version 2011-09-21 (`http://www.rotman-baycrest.on.ca/~jimmy/NIfTI/`), and then either cropped or zero-padded to $256 \times 256 \times 256$ size. Image intensities were then scaled to make the 99^{th} percentile of all image histograms match. The experiments were then done on 2D slices of images pre-processed in this way (one slice from each image), but the basis search method can be extended to three-dimensional images. Computations were done using MATLAB (The MathWorks, Natick, MA).

4.5.1. Approximation of brain MR images. The aim of this experiment was to measure the trade-off between sparsity and accuracy for an adapted wavelet packet basis and compare it to a wavelet basis. Performance was tested with ten-fold cross-validation. For each fold, a basis was trained to best approximate the out-of-fold data with a specific fraction of coefficients (values of 25%, 20%, 10%, 5%, 2% and 1% were tested), and the in-fold data was used for testing.

The results were compared to approximations of the same sparsity generated with a standard wavelet basis. Both wavelet and wavelet packet decompositions were done to four levels with the db4 filter bank from the Daubechies family of wavelets.

Cross-validation produced ten estimates of approximation accuracy (one for each fold). These estimates were then averaged to give a single figure for each sparsity setting. Results are presented in figure 4.5.1.

These results show that compared to wavelets, adapted wavelet packet representations can improve the accuracy of sparse approximations over a wide range of sparsities. The basis adapted for 30% sparsity still gives lower ℓ_1 approximation error at 2% of coefficients, compared to wavelets.

Suprisingly, it can also be observed that the wavelet packet representation adapted for 30% (or 20%) sparsity also gives more accurate approximations at sparsities as low as 2% than the wavelet packets adapted for those specific sparsities. This

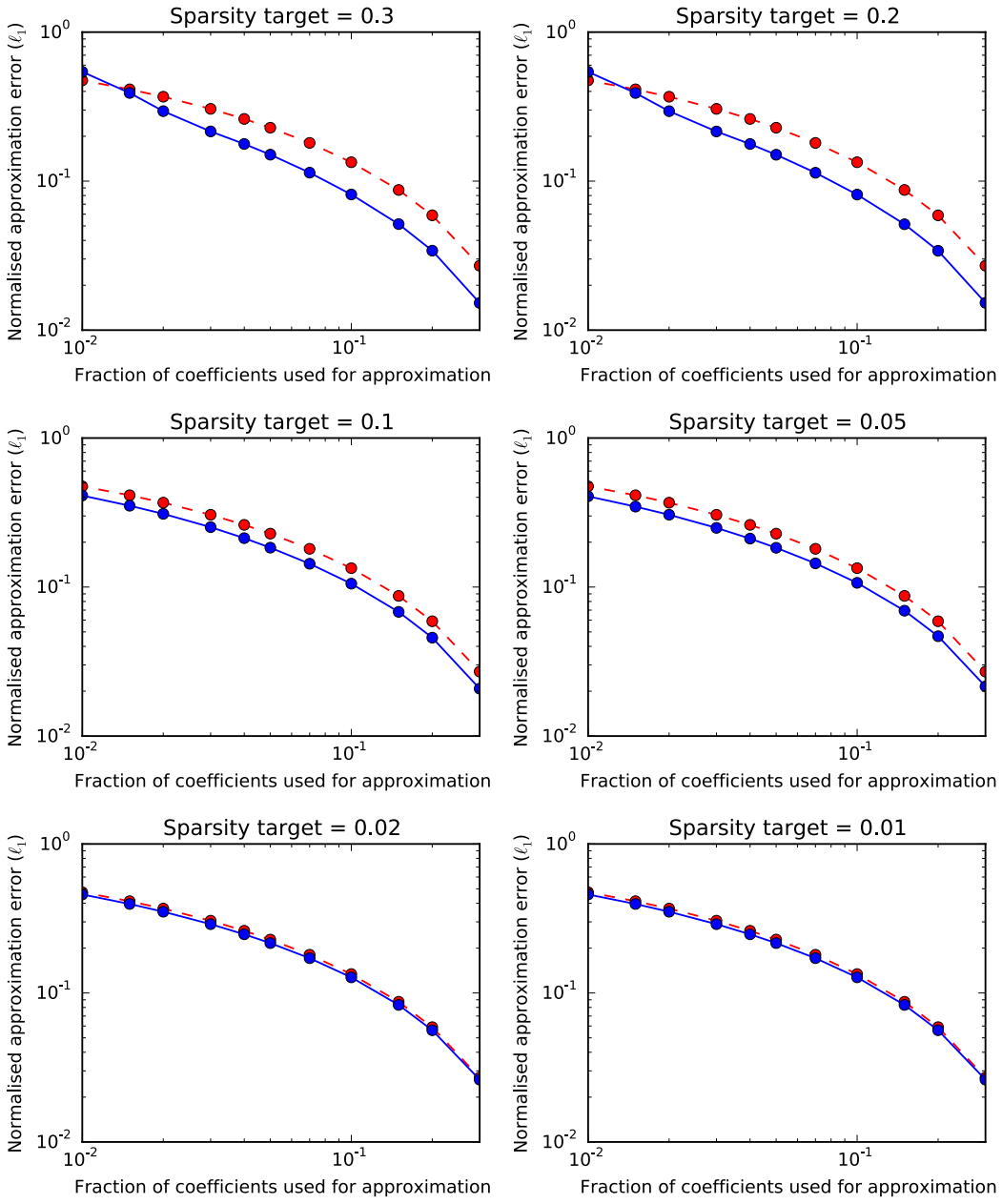


Figure 4.5.1: Approximation of brain MR images with wavelet packets (solid blue) and wavelets (dashed red). *N.b.* the wavelet basis is the same in all of the cases so the dashed red lines are also the same.

could be a sign of overfitting but it could also indicate an error in the software implementation of the proposed method.

Interestingly, for each individual value of target sparsity it was observed that in the ten-fold cross-validation process the same basis was learned every time. This

appears to indicate that the proposed method adapts the basis to a specific task, which was brain imaging in this case.

4.5.2. Compressed sensing reconstruction of brain MR images. In this experiment, images from fold 1 were reconstructed using the basis learned from folds 2 to 10. The decision to reconstruct only the images from fold 1 was made because of the long time required to compute each reconstruction. Undersampling mask generation and CS reconstruction were done using SparseMRI V0.2 (`http://www.stanford.edu/~mlustig/SparseMRI.html`). Some changes were made to the original software: the k-space mask generation code was modified to enable lower sampling densities and the k-space density compensation step was omitted in the reconstruction process.

The method of [**181**] uses a combination of wavelet domain sparsity and total variation (TV) to regularise the reconstruction. The weights of transform domain sparsity and TV were left at their default values in the SparseMRI package (0.05 and 0.02 respectively) for both wavelet and wavelet packet experiments.

An example image with its wavelet and wavelet packet reconstructions is displayed in figure 4.5.2. Table 4.5.1 presents the peak signal-to-noise ratio (PSNR) over a range of configurations. PSNR is computed with the formula $10 \cdot \log_{10}\left(\frac{(\max(\mathbf{x}))^2}{\frac{1}{n}\|\hat{\mathbf{x}}-\mathbf{x}\|_2^2}\right)$ where $\mathbf{x}$ is the reference image (represented as a vector of voxel intensities), $\hat{\mathbf{x}}$ is the reconstruction (also represented as a vector of voxel intensities), n is the number of components of $\mathbf{x}$, and $\max(\mathbf{x})$ denotes the largest component of the vector $\mathbf{x}$.

These results show that substituting the wavelet transform with a pre-trained wavelet packet transform in compressed sensing reconstruction with sparsity and total variation regularisation leads to a small improvement in PSNR. This is a somewhat suprising result given the clear advantage that wavelet packets had in the sparse approximation experiments.

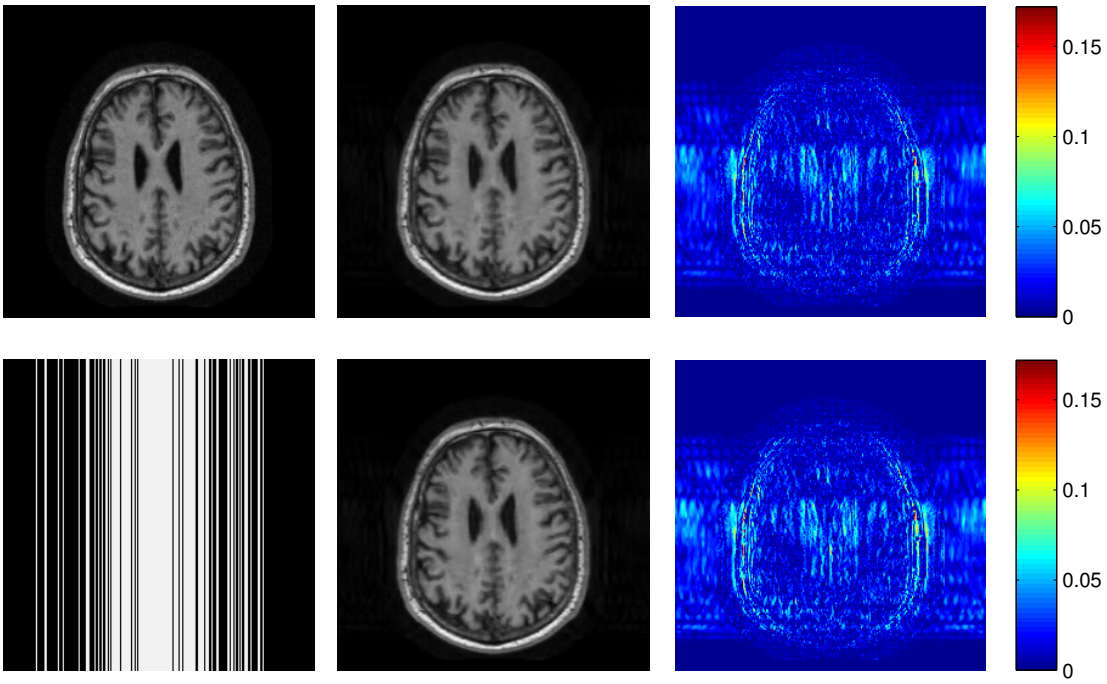


Figure 4.5.2: Example reconstructions with wavelets and wavelet packets. Top row (left to right): original image, wavelet reconstruction from 40% k-space sampling and the associated error map. Bottom row (left to right): k-space mask for 40% sampling (sampled frequencies in white), wavelet packet reconstruction from 40% sampling and the associated error map.

Table 4.5.1: Compressed sensing reconstruction PSNR with wavelets (W) and wavelet packets (P). The PSNR below are means over the 83 images in fold 1. m/n is the ratio of the number of k-space samples to the number of pixels in the image. $s_n^\star$ is the target sparsity for which the wavelet packet basis was optimised.

m/n	W [dB]	P ($s_n^\star = 0.1$) [dB]	P ($s_n^\star = 0.05$) [dB]	P ($s_n^\star = 0.02$) [dB]	P ($s_n^\star = 0.01$) [dB]	P ($s_n^\star = 0.005$) [dB]
0.4	34.0999	**34.3187**	34.0599	33.8154	33.7118	33.7118
0.35	31.4394	**31.5990**	31.3284	31.1503	31.0973	31.0973
0.3	27.4338	**27.5222**	27.4048	27.2641	27.2068	27.2068
0.25	24.5557	**24.6713**	24.5758	24.4548	24.4156	24.4156
0.2	23.7575	**23.8077**	23.7507	23.6528	23.6355	23.6355

These results also seem to show that improved sparsity is not enough to achieve a visible improvement in the quality of compressed sensing MRI reconstruction. Indeed, early research in compressed sensing emphasised the importance of another factor: incoherence between the sparse representation basis and the sensing basis

[**41**]. Coherence between two n-dimensional bases $\mathbf{\Phi}$ and $\mathbf{\Psi}$ is defined as [**34**]

$$\mu\left(\mathbf{\Phi}, \mathbf{\Psi}\right) = \sqrt{n} \cdot \max_{1 \leq k,j \leq n} |\langle \phi_k, \psi_j \rangle|$$

where ϕ_k are basis vectors from $\mathbf{\Phi}$ and ψ_j are basis vectors from $\mathbf{\Psi}$. Importantly, the sensing basis in CS-MRI is the Fourier basis. The longer waveforms of the wavelet packet representation could be more coherent with the Fourier basis, which may affect CS reconstruction quality. The standard wavelet basis also includes long waveforms in the low-frequency bands but the variable-density sampling patterns with dense sampling in the low-frequency region could be the factor that alleviates the problem in this case by providing more complete information about the lower frequencies. Explaining this problem is possible topic for future work.

4.6. Conclusions

A method was proposed for learning a wavelet packet basis from a set of images for compressed sensing applications. The core of the proposed method is the Coifman-Wickerhauser algorithm and the main contribution of this work is designing a cost function that balances sparsity and ℓ_1 approximation error while also being compatible with this algorithm. The performance of the proposed method was evaluated in two tasks: approximation and compressed sensing reconstruction of unseen brain MR images. The results show that a wavelet packet basis learned from example images can yield more accurate sparse approximations of unseen brain images than a standard wavelet basis. However, despite this significant improvement, in reconstruction of brain images from partial k-space data the difference between the learned basis and a wavelet basis is small. Further work will be required to explain this but coherence between the wavelet packet basis used for sparse approximation and the Fourier sensing basis is likely to be a contributing factor. The next step after that would be to modify the cost function to include a term which accounts for this coherence.

The experiments conducted in this study were focused on 2D slices rather than full 3D images. 3D reconstructions can be done by reconstructing all slices individually and then combining them into a complete 3D image. Alternatively, the wavelet packet framework and the proposed method can be extended to 3D signals. The latter approach is likely to be preferable [**181, 182**].

The method proposed in this chapter learns the sparse approximation basis from a set of example images. This means that it can adapt to different tasks. The examples shown in this chapter are focused on brain imaging, which is the overall topic of this dissertation, but other applications could be an interesting topic for future work.

Using example images to learn a sparse approximation basis for reconstruction can also be viewed as biasing the solution to the reconstruction problem towards the types of images that are commonly seen in a specific application. This is somewhat similar to the methods used in CT reconstruction, where regularisation is used to bias the solution to the type of images seen in CT, for example by penalising total variation [**270**].

CHAPTER 5

Sparse classification of AD with FDG-PET images

5.1. Introduction

Medical imaging enables *in vivo* study of brain structure and function in Alzheimer's disease. The Alzheimer's Disease Neuroimaging Initiative (ADNI) [**197**] imaged a large number of elderly participants with MRI and PET, creating a database that can be used by the neuroimaging community to evaluate image analysis tools. The focus of this chapter is on machine learning techniques that can be used in the context of dementia to distinguish AD patients from normal controls and to predict for patients with mild cognitive impairment whether they are likely to progress to AD. These techniques rely on machine learning algorithms with feature vectors produced by image analysis algorithms. FDG-PET images can be used to derive features by studying intensities voxel-wise [**128, 118, 129, 224, 110, 130, 138, 111**], averaging them within anatomical regions [**280, 278, 279, 171, 47**] or mapping them onto the cortical surface [**193, 235, 275**].

Due to large dimensionality of voxel-wise image data, a concern naturally arises when voxel-wise features are used that the large number of features could lead to classifier overfitting. Therefore, it may be beneficial to apply feature selection with the aim of reducing the risk of overfitting and improving prediction performance, with the additional benefit that models with less variables are easier to visualise and interpret [**116**]. Feature selection can be data-driven or based on prior knowledge of areas affected by disease based on previous studies [**50**].

A study by Salas-Gonzalez *et al.* [**224**] found that feature selection with additional dimension reduction using factor analysis gives very good accuracy, up to 95% in

distinguishing between AD patients and normal controls using FDG-PET images from the ADNI database. Other methods using feature selection have also been reported to give state-of-the-art results on both MR and FDG-PET data [**130, 278, 279, 139, 170, 287, 286, 241, 171**]. However, a study by Chu *et al.* [**50**] of MR images available from ADNI concluded that automatic feature selection does not improve classification accuracy in classification using voxel-wise grey matter density features, although extraction of pre-defined anatomical structures can be of some benefit. They noted that it may be the case that most features are weakly informative, rather than non-informative. They also remarked that multivariate patterns may have made it more difficult to rank features.

In contrast to feature selection, feature agglomeration performs clustering of correlated features and combines them in groups to produce a smaller set of features. This approach has been studied far less than feature selection, but it did receive some attention from the neuroimaging community. Fan *et al.* [**95**] used supervised watershed segmentation with additional feature selection to build classifiers for MR images, and their algorithm was subsequently applied to distinguishing MCI patients from normal subjects using a combination of structural MRI and ^{15}O-water-PET [**94**]. Michel *et al.* [**192**] used supervised hierarchical clustering with spatial connectivity constraints to build features for analysis of fMRI data.

The aim of this chapter is to compare several classification and dimensionality reduction algorithms applied to classification of FDG-PET images available from ADNI. The scope of this comparison includes feature selection as well feature agglomeration, combined with classifiers including linear methods (linear SVM and logistic regression) as well as the non-linear random forest classifier.

Among many existing feature selection methods, mass univariate F-tests (ANOVA) and ℓ_1 regularisation were evaluated in this chapter. Two algorithms for feature agglomeration were also evaluated: a method based on Ward clustering with spatial neighbourhood constraints and another method based on simple linear iterative

clustering (SLIC) supervoxels [**1**]. SLIC is an image segmentation algorithm, but this chapter presents a potentially novel way of using SLIC for feature agglomeration in image-based classification.

5.2. Background

5.2.1. Classification algorithms. Classification algorithms are discussed in chapter 3, but there are additional concerns that apply to high-dimensional data. When choosing an algorithm for an image-based classification problem, it has to be taken into account that in a typical medical imaging scenario the number of voxels is much larger than the number of images available for training, and this may still be the case after applying feature selection. Simple algorithms with strong regularisation are often chosen for problems of this type [**123**] and therefore this chapter includes two linear classifiers: SVM and regularised logistic regression, with SVM being very common in Alzheimer's disease classification studies [**60**]. In addition, the non-linear random forest classifier was also evaluated. Random forests compensate for the risk of overfitting in decision trees by combining a large number of trees through bootstrap and additionally randomly restricting the choice of features for each branching point in these trees [**24, 123**]. Random forests were previously applied to multi-modal classification with FDG-PET and MRI data [**110**].

5.2.2. Dimensionality reduction algorithms. Dimensionality reduction is discussed from an overall perspective in chapter 3. In the following, the algorithms that are used in this chapter are discussed in detail.

5.2.2.1. *Mass univariate F-test.* The mass univariate F-test method is a filter-type feature selection algorithm which relies on conducting a separate hypothesis test for each of the features. In order for this test to be computed, the group memberships of individual samples have to be known (in this chapter the two groups are AD patients and healthy controls). The null hypothesis of each of these tests is

that the mean of the feature is the same for all groups, meaning that the feature is not informative in a classification setting. The F-statistic is computed as the ratio of between-group variability to within-group variability [**113**]:

$$F = \frac{s_b^2}{s_w^2}$$
$$s_b^2 = \frac{\sum_i n_i \left(\bar{x}_i - \bar{x}\right)^2}{m-1}$$
$$s_w^2 = \frac{\sum_{ij} \left(x_{ij} - \bar{x}_i\right)^2}{n-m}$$

where x_{ij} is the j-th observation in i-th group, $\bar{x}_i$ is the sample mean of the i-th group, $\bar{x}$ is the sample mean of all data, n_i is the number of observations in i-th group, n is the total number of observations and m is the number of groups.

There are several ways of selecting a subset of features based on their individual values of the F-statistic. One relies on computing the associated p-value for each feature's F-statistic and selecting features that have p-values below a specific threshold, which can be adjusted to control the false positive rate, familywise error or false discovery rate at a specified level. An alternative way is to choose a specified number of features (or a specified proportion of the total number of features) from the start of their sequence sorted by p-value increasing. Therefore, all these variations on the mass univariate F-test method require the value of a parameter to be chosen.

The univariate F-test can also be referred to as one-way ANOVA, where ANOVA stands for "analysis of variance". If there are only two groups being compared, one-way ANOVA is essentially equivalent to the two-tailed, two-sample t-test [**218**].

5.2.2.2. *Ward feature agglomeration.* Ward feature agglomeration relies on Ward clustering [**142**], a type of agglomerative clustering algorithm (agglomerative clustering was discussed in Chapter 3). The spatially aware Ward feature agglomeration algorithm used in this chapter was proposed by Michel *et al.* for fMRI applications [**192**]. They also added an additional modification where the Ward tree is pruned in

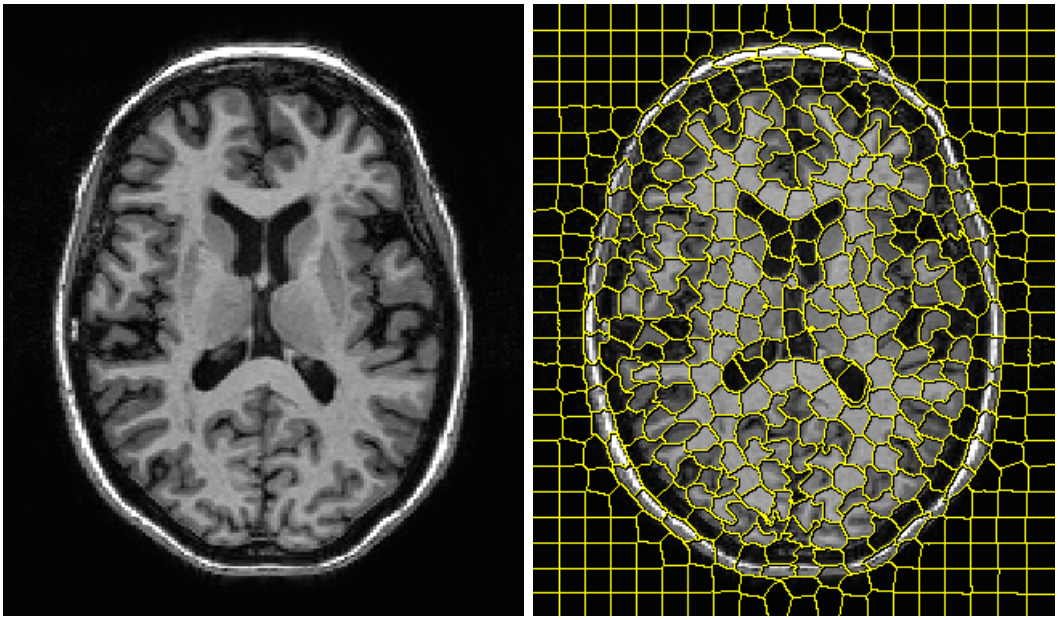
a supervised way, but the algorithm used in this chapter is unsupervised and does not include this modification.

5.2.2.3. *Feature agglomeration with SLIC supervoxels.* Simple linear iterative clustering (SLIC) was proposed as a method for segmenting images into superpixels (or supervoxels in three dimensions) by Achanta *et al.* [**1**]. It was soon applied to 3D electron microscopy, where supervoxels were used as part of a segmentation algorithm for mitochondria [**180**].

SLIC is based on the K-means clustering algorithm, which was described in Chapter 3, with some modifications described below.

In SLIC, each pixel (or voxel) is assigned a feature vector which is a concatenation of colour parameters (colour vector) and spatial coordinates. The cluster centres are initialised with feature vectors obtained by sampling the image on a regular grid. These intial values can then be adjusted by searching within a 3×3 spatial neighbourhood of each of the selected points for pixels with the lowest gradient (this step was omitted in the implementation used for this chapter). Then the algorithm runs a modified K-means loop on the feature vectors until convergence. The modifications consist of introducing a spatial window when searching for pixels to include in each cluster (which makes the algorithm much more efficient) and also introducing a compactness factor that balances the influence of spatial distance and colour distance [**1**]. Additional post-processing may be applied to ensure spatial connectivity within each superpixel [**1**], but it was not applied in this study. If the input images are RGB colour images, they are transformed into the CIELAB colour space before running SLIC [**1**].

A subsequent update to SLIC, named "SLICO" (or "slic-zero"), replaces the global compactness parameter with a value adjusted adaptively for each individual supervoxel, although an initial compactness setting is still required for the first iteration (`http://ivrl.epfl.ch/research/superpixels`). This improved version of SLIC



(a) Original image **(b)** SLICO

Figure 5.2.1: SLICO applied to a 2D slice from an MR image of a human brain.

is used in this chapter. Figure 5.2.1 shows an example of SLICO (without the spatial connectivity post-processing step) applied to a brain image segmentation task.

In this chapter, SLICO was adapted for feature agglomeration as follows. First, for each voxel the colour vector was replaced with the vector of image intensities at this specific voxel for all images. After SLICO segmentation, each supervoxel was merged into a single feature by averaging its constituent voxels.

5.2.3. Image registration. The work presented in this chapter is based on images that underwent a process of spatial normalisation and intensity adjustment as part of the work by Gray *et al.* [**112, 109**]. The following brief description of image registration is intended to provide the background for section 5.3, which describes the image normalisation process.

Before applying machine learning algorithms in a voxel-wise fashion, it is important to ensure that the images are spatially aligned to ensure that a voxel indexed with a particular set of coordinates represents the same anatomical location in all images. This is achieved by applying a spatial transformation that is estimated with an

image registration algorithm. Image registration is the task of mapping the points in one image to the respective points in another image. In the context of medical imaging, image transformations are often required to include an affine component as well as a non-rigid component to account for deformations due to motion and anatomical differences between patients. The free-form deformation algorithm [**221**] is an example of a non-rigid registration method.

Image registration can be based on identifying salient features (keypoints) in the source and target images and matching them, or alternatively it can seek to align images in a way that ensures a close match of pixel or voxel intensities between the transformed source image and the target image [**243**]. Image cross-correlation is a simple measure of similarity for intensity-based methods but normalised mutual information is more commonly used in practice due to its advantages of overlap invariance and being able to match images from different modalities [**238, 221**].

Medical image registration is discussed in detail by Hajnal *et al.* [**119**] and a recent review by Sotiras *et al.* is available in [**237**].

5.3. Image normalisation

The FDG-PET images studied in this chapter originate from the ADNI database and were processed into a homogenous data set by Gray *et al.* for the study in [**110**]. The following summary describes the processing steps applied. Full details are available in [**109**].

The images available from ADNI were acquired under several slightly different FDG-PET protocols (30-minute static, 30-minute dynamic or 60-minute dynamic). The dynamic scans were first converted to a 30-minute static format by aligning (with rigid registration) every frame of each dynamic sequence to its first frame and averaging [**109**]. This was done using the IRTK package (`https://www.doc.ic.ac.uk/`

`~dr/software/`) with normalised mutual information as a measure of image similarity. For the 60-minute sequences only the final 30 minutes were used. The reasoning behind this is that a large majority of scans were 30-minute dynamic scans starting approximately 30 minutes after tracer injection, whereas the 60-minute dynamic scans started simultaneously with tracer injection [**109**].

These FDG-PET images were then registered (using affine registration) to their respective MR images using the tools from IRTK [**109**].

SPM5 "Segment" module was then used to linearly and non-linearly register the MR images to the MNI brain template image and the resultant transformation parameters were used to transform FDG-PET images from their individual MRI spaces to the common MNI space, completing the process of spatial alignment [**109**].

The MNI-space FDG-PET images were then smoothed with scanner-specific Gaussian kernels, reducing them to a common spatial resolution of 8mm full-width-at-half-maximum (FWHM), which was followed by another smoothing step, using a Gaussian kernel with 8mm FWHM [**109**]. The aim of this smoothing step was to reduce noise (which is a common problem in PET), and additionally reduce the influence of any misalignment due to imperfections in the registration process [**109**].

This was followed by intensity normalisation, using the reference cluster method of Yakushev *et al.* [**271**], to compensate for the differences between subjects in overall radioactivity [**109**].

The processed dataset consisted of 287 images taken at baseline (a subject's first scan as part of ADNI), including 71 AD patients, 69 cognitive normal (CN) controls and 147 MCI patients. 73 of the MCI patients later converted to AD, so they were labeled as progressive MCI (pMCI). The remaining 74 MCI patients did not convert to AD so they were labeled as stable MCI (sMCI). The total number of subjects (287) is somewhat smaller than the 315 originally available from ADNI because of

Table 5.3.1: Numbers of subjects in AD, CN, pMCI and sMCI groups.

	AD	CN	pMCI	sMCI
Number of subjects	71	69	73	74

image quality issues in some cases and failure of particular processing steps in others. The detailed list of excluded images is available in [**109**].

For the experiments conducted specifically for this chapter, the images provided by Gray *et al.* [**109**] were finally re-sampled, from the original resolution of 2mm (image size 79 x 95 x 82), by a factor of 2 in each direction, yielding images with 4mm resolution. The re-sampling process did not apply any filtering since the images were already smoothed twice, including with an 8mm FWHM kernel. The aim of re-sampling was to reduce the dimensionality of machine learning data (by a factor of 8), thus enabling a more extensive set of parameter configurations to be covered in grid search.

5.4. Classification pipelines

This chapter evaluates the following five classification algorithms (described in detail in chapter 3).

(1) Linear Support Vector Machines (SVM) with ℓ_2 regularisation
(2) Linear Support Vector Machines (SVM) with ℓ_1 regularisation
(3) Logistic Regression with ℓ_2 regularisation
(4) Logistic Regression with ℓ_1 regularisation
(5) Random Forests (RF)

The regularisation parameter for logistic regression and SVM was adjusted by cross-validation in both ℓ_1 and ℓ_2 regularisation cases.

For random forests the parameter controlling the proportion of features randomly sampled to search for the best node partition was also adjusted by cross-validation. A common convention is to set the number of features sampled as the square root

of the total number of features, but in high-dimensional problems this would cause only a small proportion of features to be sampled for consideration in each node partition, due to the very large total number of features. Therefore, if the relevant features are sparse then there isf a risk of having node partitions where very few or no informative features appear in the subset available to choose from [**123**]. This concern was especially relevant in those cases where no dimensionality reduction was used, but for consistency the proportion of features sampled was chosen by cross-validation in all cases.

In general, large numbers of trees are preferable but there are limits with regards to what is computationally feasible. In this chapter the number of trees in each random forest was set to 1000 as a value that was manageable with the computational power available, considering that forests had to be built many times due to cross-validation requirements.

For each of the above classifiers, the following dimensionality reduction algorithms were evaluated as pre-processing steps.

(1) No dimensionality reduction.
(2) Feature selection with mass univariate F-test (one-way ANOVA).
(3) Feature agglomeration with Ward clustering, with a spatial neighbourhood constraint.
(4) Feature agglomeration with SLIC supervoxels, using the SLICO version of this method.

These dimensionality reduction methods have some parameters that were adjusted by cross-validation. These parameters are: for ANOVA the p-value threshold below which features are selected, for Ward agglomeration the number of clusters, and for SLIC the number of clusters as well.

Logistic Regression with ℓ_1 regularisation already incorporates feature selection, as the ℓ_1 norm favours sparse solutions. So in effect, four dimensionality reduction

methods were evaluated specifically for logistic regression. In addition, ℓ_1 regularised Logistic Regression is combined with each of ANOVA, Ward and SLIC methods, to see if there are any benefits to applying two types of dimensionality reduction methods in a sequence. Similarly, the SVM classifier is also evaluated in a form with ℓ_1 regularisation.

The experiments were done with Python using the Scikit-learn package (`http://scikit-learn.org/stable/`) [**203**] for machine learning algorithms. Scikit-image (`http://scikit-image.org/`) was used for SLIC, with modifications to enable the SLICO version that were later merged into the master branch (see `https://github.com/scikit-image/scikit-image/pull/864` for details). The Nibabel package (`http://nipy.org/nibabel/`) was used to read and write images.

5.5. Evaluation

All classifiers were trained on the data representing subjects diagnosed as Alzheimer's disease (AD) or cognititive normal (CN). The trained classifiers were then evaluated on two classification tasks:

(1) Classify a subject from the {AD, CN} pool as either AD or CN. We also refer to this task as "AD detection".
(2) Predict whether a subject from the MCI pool will convert to AD or remain stable (within a time horizon dependent on the data available).

The approach of using a classifier trained on AD and CN data for prediction of progression from MCI to AD has been tested before by other authors with good results [**235, 61**]. Alternative approaches include combining AD, CN and MCI subjects in a semi-supervised learning framework [**97, 277**] or using domain transfer learning [**48, 47**].

For each classification pipeline this experiment was repeated 100 times with a stratified random partition of the entire data set into 75% training data and 25% test

data (the partitioning included MCI subjects even though they were only used for testing). The parameters for each classification pipeline were selected according to best results in ten-fold cross-validation on the training subset only, followed by re-fit (with the selected parameters) to the whole training subset. In other words, the parameter tuning loops were nested inside the testing loops.

Classification results are summarised in the next section with the standard metrics of classification accuracy, sensitivity and specificity, as discussed in chapter 3.

5.6. Results

Table 5.6.1 displays means and standard errors of performance metrics: accuracy (Acc), sensitivity (Sens) and specificity (Spec) for the AD detection and MCI prediction tasks. Figure 5.6.1 shows box plots for classification accuracies.

Best accuracy (86.43%) on the AD vs CN task was achieved by the composition of ANOVA feature selection and a linear SVM classifier with ℓ_2 regularisation. However, the box plots show that the variations due to different train-test partitions are of similar magnitude to the differences between particular methods. A similar observation holds for MCI prediction accuracies, but the most accurate method in this case is SLIC-based feature agglomeration followed by a logistic regression classifier with ℓ_2 regularisation (69.38% accuracy).

Table 5.6.2 displays the numbers of features output by the first stage of each pipeline and used by the final classifier (*n.b.* those two numbers may differ due to some of the classifiers having their own embedded feature selection, specifically random forests and the ones with ℓ_1 regularisation). The number of features had quartiles that were scattered over a very wide range, which meant that plots were difficult to interpret, so these numbers are only presented in tables.

Table 5.6.1: Classification performance. The performance metrics presented include accuracy (Acc), sensitivity (Sens) and specificity (Spec) for two tasks: AD/HC and pMCI/sMCI. These figures are presented as *mean*(*standard error*). Note that for both tasks the classifiers were trained to distinguish between AD and HC subjects.

	Algorithm	AD/HC			pMCI/sMCI		
		Acc	Sens	Spec	Acc	Sens	Spec
Embedded	Linear SVM (L2)	.8534 (.0058)	.8428 (.0095)	.8647 (.0083)	.6689 (.0076)	.7025 (.0105)	.6433 (.0103)
	Linear SVM (L1)	.8477 (.0056)	.8344 (.0093)	.8618 (.0091)	.6646 (.0072)	.6663 (.0107)	.6633 (.0100)
	Logistic (L2)	.8549 (.0054)	.8300 (.0097)	**.8812** (.0079)	.6897 (.0073)	.7031 (.0112)	.6795 (.0093)
	Logistic (L1)	.8563 (.0059)	.8533 (.0093)	.8594 (.0086)	.6724 (.0070)	.6650 (.0103)	.6781 (.0094)
	Random Forest	.8586 (.0061)	.8694 (.0083)	.8471 (.0093)	.6505 (.0070)	**.7288** (.0107)	.5910 (.0092)
Filter	ANOVA → Linear SVM (L2)	**.8643** (.0050)	.8556 (.0088)	.8735 (.0079)	.6559 (.0072)	.6975 (.0116)	.6243 (.0096)
	ANOVA → Linear SVM (L1)	.8429 (.0057)	.8244 (.0095)	.8624 (.0094)	.6603 (.0074)	.6519 (.0110)	.6667 (.0093)
	ANOVA → Logistic (L2)	.8500 (.0063)	.8267 (.0105)	.8747 (.0083)	.6773 (.0073)	.6644 (.0129)	**.6871** (.0091)
	ANOVA → Logistic (L1)	.8474 (.0055)	.8367 (.0089)	.8588 (.0087)	.6692 (.0063)	.6787 (.0109)	.6619 (.0084)
	ANOVA → Random Forest	.8511 (.0060)	.8606 (.0085)	.8412 (.0095)	.6505 (.0070)	.7212 (.0105)	.5967 (.0094)
Clustering	Ward → Linear SVM (L2)	.8306 (.0063)	.8250 (.0101)	.8365 (.0095)	.6611 (.0078)	.7094 (.0118)	.6243 (.0101)
	Ward → Linear SVM (L1)	.8394 (.0063)	.8294 (.0107)	.8500 (.0085)	.6708 (.0076)	.6913 (.0117)	.6552 (.0096)
	Ward → Logistic (L2)	.8409 (.0060)	.8244 (.0100)	.8582 (.0087)	.6754 (.0067)	.7075 (.0117)	.6510 (.0088)
	Ward → Logistic (L1)	.8417 (.0057)	.8283 (.0086)	.8559 (.0085)	.6659 (.0074)	.6875 (.0113)	.6495 (.0109)
	Ward → Random Forest	.8531 (.0058)	.8694 (.0081)	.8359 (.0097)	.6441 (.0070)	.7188 (.0100)	.5871 (.0100)
	SLIC → Linear SVM (L2)	.8334 (.0065)	.8200 (.0104)	.8476 (.0087)	.6743 (.0078)	.6950 (.0110)	.6586 (.0106)
	SLIC → Linear SVM (L1)	.8569 (.0060)	.8444 (.0100)	.8700 (.0089)	.6738 (.0072)	.6819 (.0099)	.6676 (.0101)
	SLIC → Logistic (L2)	.8474 (.0069)	.8300 (.0109)	.8659 (.0097)	**.6938** (.0072)	.7087 (.0102)	.6824 (.0096)
	SLIC → Logistic (L1)	.8551 (.0059)	.8478 (.0097)	.8629 (.0085)	.6705 (.0077)	.6825 (.0104)	.6614 (.0101)
	SLIC → Random Forest	.8606 (.0066)	**.8733** (.0084)	.8471 (.0099)	.6422 (.0074)	.7150 (.0112)	.5867 (.0090)

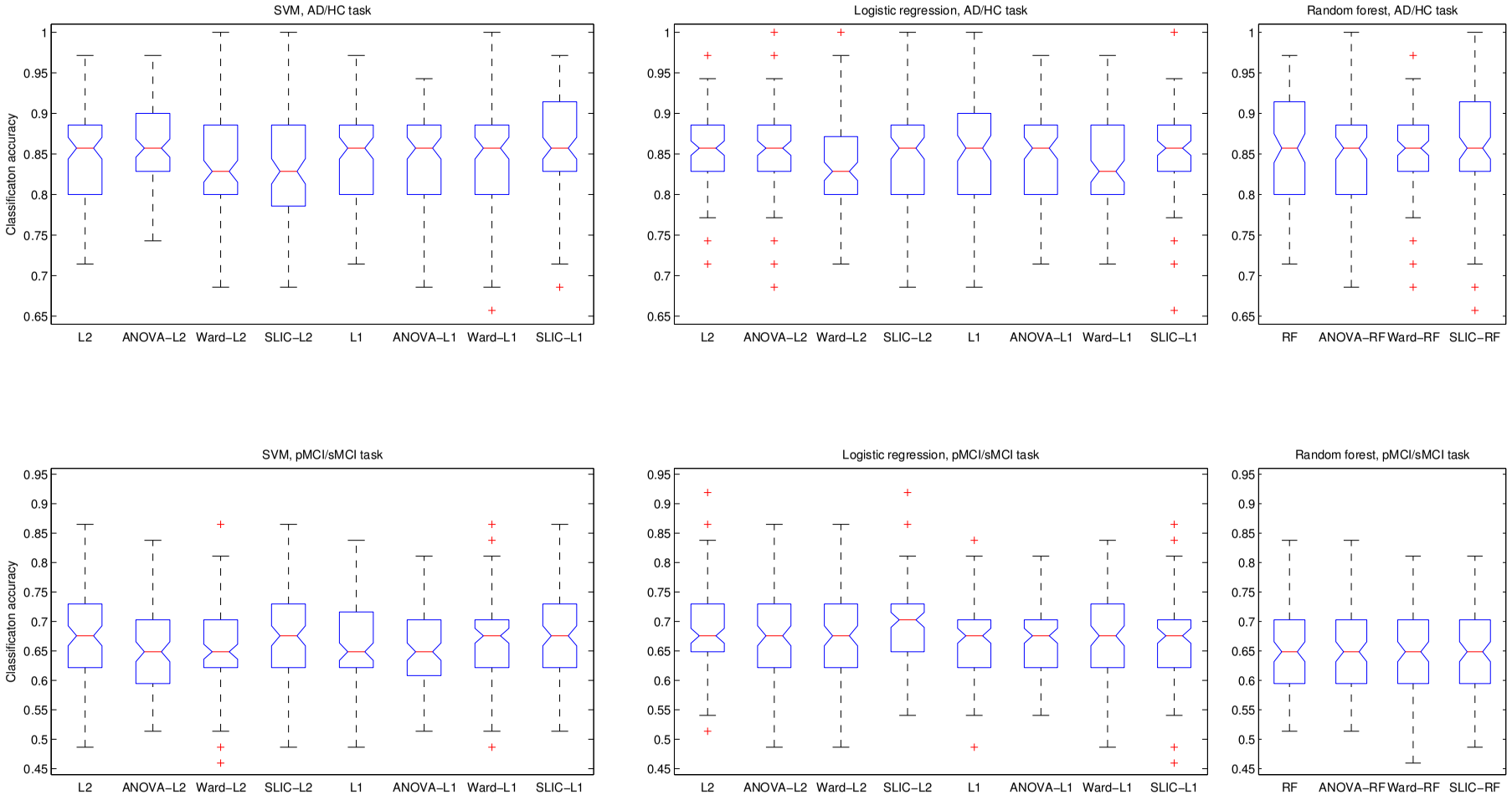


Figure 5.6.1: Classification accuracy in AD/CN and sMCI/pMCI classification tasks. For SVM and logistic regression the labels under box plots indicate the dimensionality reduction method and the type of regularisation (L2 refers to ℓ_2 and L1 to ℓ_1). For random forests the labels indicate the dimensionality reduction method. Note that for both tasks the classifiers were trained to distinguish between AD and HC subjects.

Table 5.6.2: Number of features selected by each algorithm, quoting the first, second and third quartile. "After first stage" refers to the number of features output by the filtering or clustering method used. "Used by classifier" refers to the number of features actually used by the classification algorithm in the last stage of the classification pipeline. The total number of features available is 43725. The classifiers were trained to distinguish between AD and HC subjects.

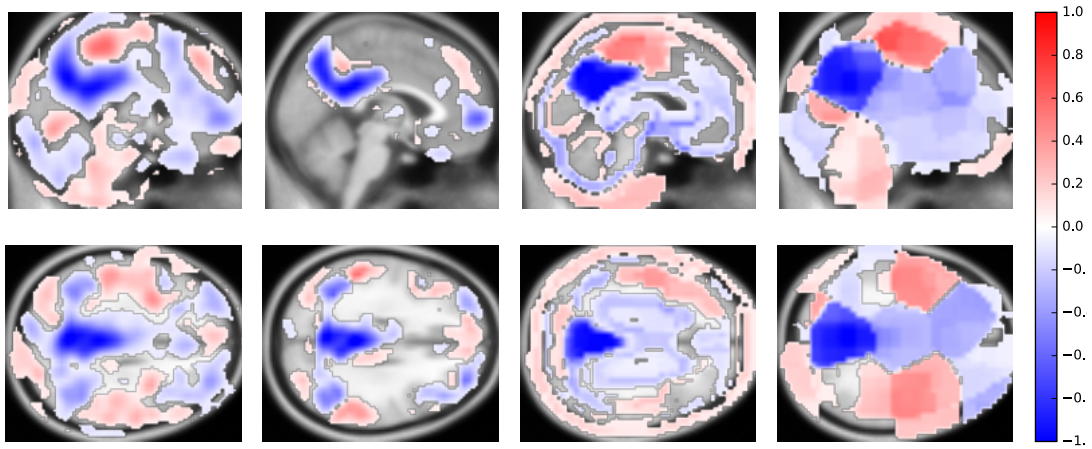
	Algorithm	After first stage			Used by classifier		
		Q_1	Q_2	Q_3	Q_1	Q_2	Q_3
None or embedded	Linear SVM (L2)		*n/a*		43725	43725	43725
	Linear SVM (L1)		*n/a*		56	143.5	1402
	Logistic (L2)		*n/a*		43725	43725	43725
	Logistic (L1)		*n/a*		29	59.5	1168
	Random Forest		*n/a*		3302	3749	4283
Filter	ANOVA → Linear SVM (L2)	8868	14870	21072	8868	14870	21072
	ANOVA → Linear SVM (L1)	9002	16099	24202	51.5	146.5	798
	ANOVA → Logistic (L2)	11860	20715	36614	11860	20715	36614
	ANOVA → Logistic (L1)	7493	15007	24398	31	50	335
	ANOVA → Random Forest	11182	23706	33295	3002	3509.5	4229.5
Clustering	Ward → Linear SVM (L2)	56	316	562	56	316	562
	Ward → Linear SVM (L1)	100	247	1389	23.5	35	58
	Ward → Logistic (L2)	215	316	1000	215	316	1000
	Ward → Logistic (L1)	178	562	2470	18	28.5	42
	Ward → Random Forest	562	1778	3162	560	1438.5	2032
	SLIC → Linear SVM (L2)	493	883	883	493	883	883
	SLIC → Linear SVM (L1)	870	883	1957	28	39	57
	SLIC → Logistic (L2)	876.5	883	1957	876.5	883	1957
	SLIC → Logistic (L1)	870	883	1957	19	28	39
	SLIC → Random Forest	5986	5986	5986	1906.5	2237.5	2513.5

Dimensionality reduction appears to have the most pronounced effect for ℓ_1-regularised classifiers, where the number of features is less than 150 in all cases. ANOVA appears to select (in the median) about one third to a half of the total number of features. The differences in the number of features between ANOVA with different classifiers are probably due to cross-validation selecting different ANOVA significance thresholds. This is expected, as each classifier may achieve its optimum performance with different fature subsets.

Classification pipelines with Ward or SLIC feature agglomeration appear to settle on much smaller numbers of features than those with ANOVA feature selection. In addition, Ward agglomeration settles on less features than SLIC agglomeration in most cases. It is worth remembering that these specific numbers of clusters were selected through cross-validation, so they are chosen for best performance.

Linear classification pipelines can be visualised by displaying images of classifier weights assigned to respective voxels or clusters of voxels. Visualisations of linear classifiers with different dimensionality reduction pre-processing steps are shown in figures 5.6.2, 5.6.3, 5.6.4 and 5.6.5. Random forests are not suitable for this type of representation but feature importances (described in chapter 3) can be displayed instead, showing which voxels are relevant. However, feature importances do not indicate whether a feature is positively or negatively correlated with AD. Methods using random forests with different dimensionality reduction pre-processing steps are visualised in figure 5.6.6. In all cases the following steps were followed to generate the images.

Voxel weights were averaged over the 100 random training-test partitions of the data. Since the classifiers were trained on downsampled (4mm resolution) images, the results were interpolated to 2mm resolution to match the template. Some voxel weights had very large values, so the intensities were clipped as follows to ensure better contrast. The individual clipping threshold for each image was chosen as either the 0.01^{th} or the 99.99^{th}percentile of the intensity distribution, whichever



(a) No dim. red. **(b)** ANOVA **(c)** Ward **(d)** SLIC

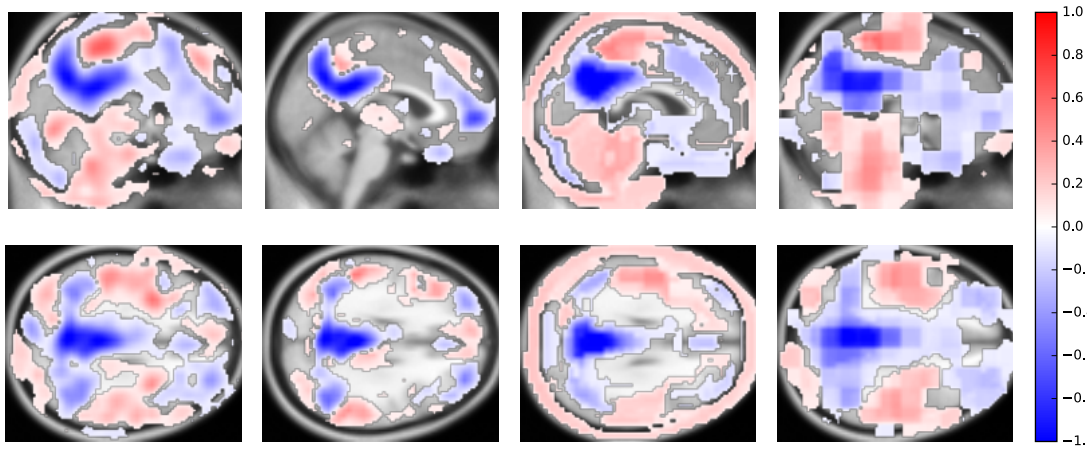
Figure 5.6.2: Voxel weights for SVM with ℓ_2 regularisation and different dimensionality reduction pre-processing steps, averaged over 100 random training/test partions of the data.

had larger absolute value. Values outside the range between this threshold and its negative were clipped. The problem of very large classifier weights was only an issue in case of sparse classifiers but clipping was applied to maps of all linear classifiers for consistency.

Voxel weights were then rescaled in each image individualy to the $[-1, 1]$ range and mapped onto a colormap. Voxels in the $[-0.05, 0.05]$ range were set to transparent. Finally, these images were overlaid on top of the MNI152lin anatomical template.

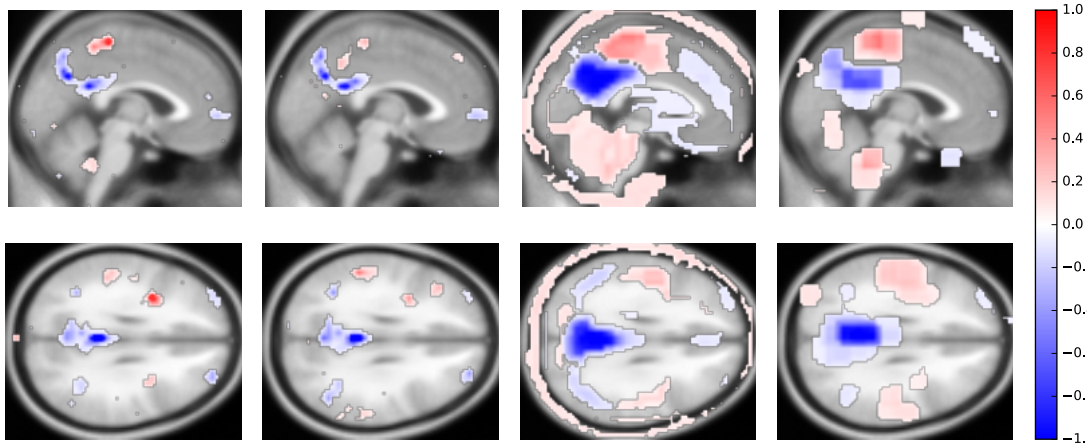
5.7. Discussion

5.7.1. Classification accuracy. The main notable feature of these results is that, surprisingly, there is little difference in terms of performance between very different algorithms. As each subplot in figure 5.6.1 shows, the variation due to applying different dimensionality reduction methods is relatively small when compared to the variation of results over the 100 random dataset partitions for each specific classification pipeline. None of the algorithms appear to particularly stand out from the rest. Furthermore, using the linear algorithms in their ℓ_2-regularised forms gave



(a) No dim. red. **(b)** ANOVA **(c)** Ward **(d)** SLIC

Figure 5.6.3: Voxel weights for logistic regression with ℓ_2 regularisation and different dimensionality reduction pre-processing steps, averaged over 100 random training/test partions of the data.

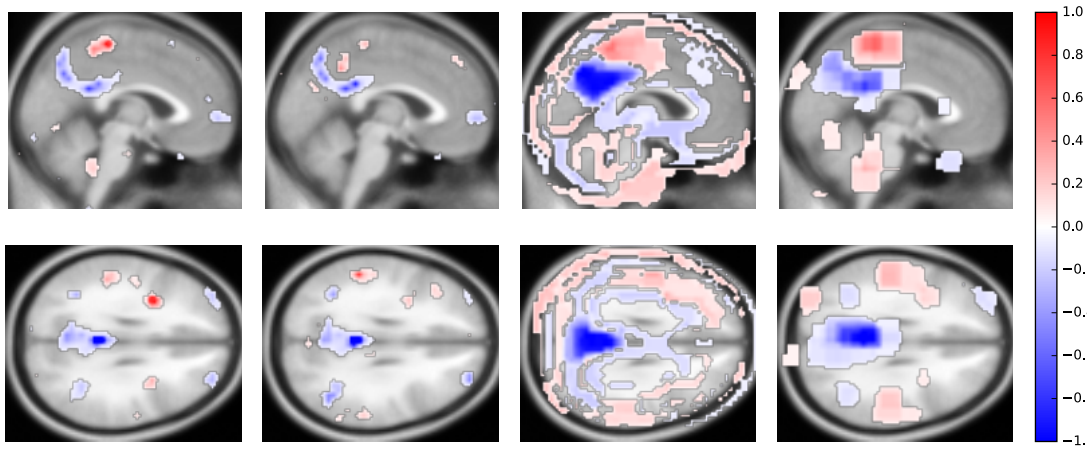


(a) No dim. red. **(b)** ANOVA **(c)** Ward **(d)** SLIC

Figure 5.6.4: Voxel weights for SVM with ℓ_1 regularisation and different dimensionality reduction pre-processing steps, averaged over 100 random training/test partions of the data.

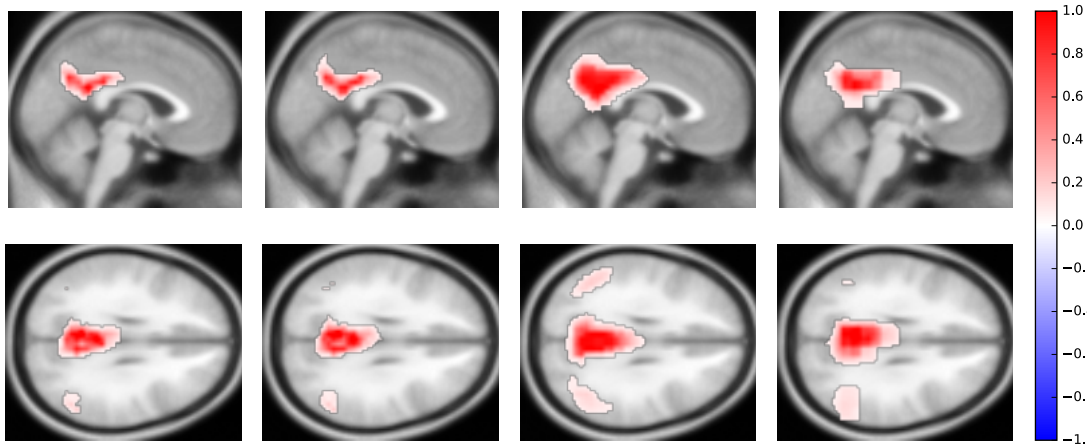
results which appear to be just as good as when using ℓ_1 regularisation or any of the other dimensionality reduction methods evaluated in this chapter.

The lack of benefit from using sparse ℓ_1 methods could indicate that FDG-PET image classification with voxel-wise features is not intrinsically a sparse problem. Perhaps this is due to correlation between neighbouring voxels in the images, which



(a) No dim. red. **(b)** ANOVA **(c)** Ward **(d)** SLIC

Figure 5.6.5: Voxel weights for logistic regression with ℓ_1 regularisation and different dimensionality reduction pre-processing steps, averaged over 100 random training/test partions of the data.



(a) No dim. red. **(b)** ANOVA **(c)** Ward **(d)** SLIC

Figure 5.6.6: Feature importances for random forest classifiers with different dimensionality reduction pre-processing steps, averaged over 100 random training/test partitions of the data. *N.b.* feature importances only quantify the importance of a feature, without indicating if a large value makes the image more likely to be classified as AD or normal.

distributes information across regions. In this case, ℓ_1-regularised methods, which aim to remove redundant features, would have the disadvantage of not being able to use redundancy to reduce sensitivity to noise. The images used in this chapter were carefully processed, with multiple smoothing steps that suppress noise, which could

explain why sparse models did not show degraded performance (due to excess noise) either. Smooth and low-noise images could also be an explanation of why feature agglomeration produced no improvement over classifiers on their own: first, the noise-suppressing properties of feature aglomeration have no advantage, and second, due to high autocorrelation within the images, neighbouring voxels are likely to be treated by the classifier similarly anyway.

Another reason for performance parity between pipelines with and without feature selection could be that superfluous features do little or no harm to the accuracy of the classifier, presumably due to being assigned relatively small weights under the standard ℓ_2 regularisation.

One subtle pattern that can be deduced from figure 5.6.1 is that using random forests as the final classifier gives overall slightly better performance on the AD/CN task than the linear classifiers (SVM and logistic regression). However, when the classifiers trained on AD/CN data are used to classify MCI subjects into pMCI and sMCI categories, pipelines based on random forests are slightly less accurate than those based on linear classifiers. This could indicate that random forests are more agile in learning a particular task but perhaps their learning is not as easily transferable to similar tasks as in the case of linear classifiers. Nevertheless, those differences are still relatively small.

The Ward and SLIC feature clustering algorithms perform similarly but SLIC seems to be slightly better in most pipelines. This could be due to SLIC's preference for more compact clusters, which can be thought of as prior knowledge of probable cluster shapes.

Table 5.7.1 lists AD/HC classification and pMCI/sMCI prediction accuracies reported in several journal publications using ADNI data where FDG-PET was used either alone or with other modalities. In comparison, the best results achieved in this work are 86.43% accuracy for the AD/HC task (achieved with ANOVA feature

selection and an SVM classifier) and 69.38% for the pMCI/sMCI task (achieved with SLIC feature agglomeration and a logistic regression classifier with ℓ_2 regularisation, trained on the AD/HC data).

The accuracy achieved in this chapter on the AD/HC task is lower than the state-of-the-art while the accuracy for the pMCI/sMCI task is comparable to the more sophisticated multi-modal method of Young *et al.* [**272**] and higher than some of the other methods. However, one should also be careful when comparing the performance figures reported by different studies, as there are differences in the subsets of ADNI subjects used and evaluation methodologies. Studies of algorithms for prediction of conversion from MCI to AD also differ in the time horizon defined for conversion [**272**].

Most of the methods listed in table 5.7.1 combine data from multiple modalities. In contrast, the methods presented in this chapter rely mainly on FDG-PET, with MR images only used to help with spatial alignment of FDG-PET images. This means that the results obtained for the methods evaluated in this chapter are not directly comparable with most of the results presented in table 5.7.1. Including features derived from MR images as well as genetic and CSF biomarkers would probably improve the results of this chapter. However, two of the methods listed in table 5.7.1 are based on FDG-PET: the one proposed by Gray *et al.* [**112**] and the one proposed by Salas-Gonzalez *et al.* [**224**].

The method proposed by Gray *et al.* [**112**] is based on regional features, measuring the strength of the FDG-PET signal per mm^3 over individual anatomical brain regions. The intensities are normalised with a reference cluster method. An SVM classifier with a radial basis function kernel is then applied to these features. Importantly, their dataset is an earlier version of the one used to generate the results in this chapter and the testing methodology is also similar, making it easier to compare the two methods. The classification accuracies presented in this chapter are higher

Table 5.7.1: Classification accuracies in other studies using ADNI FDG-PET images, with or without other modalities.

Method	Modalities	Number of subjects: AD, HC, MCI (pMCI, sMCI)	Classification accuracy (AD/HC)	Conversion prediction accuracy (pMCI/sMCI)
Cheng *et al.* [**47**]	MRI, FDG-PET, CSF	51, 52, 99 (43, 56)	-	79.4%
Gray *et al.* [**112**]	FDG-PET (MRI used for image alignment)	71, 69, 147 (62, 85)	81.6%	56.4%
Gray *et al.* [**111**]	MRI, FDG-PET, CSF, APOE	37, 35, 75 (34, 41)	89.0%	58.0%
Hinrichs *et al.* [**130**]	MRI, FDG-PET, CSF, NPSE, APOE	48, 66, 119	92.4%	-
Liu et al. [**171**]	MRI, FDG-PET	51, 52, 99 (43, 56)	94.37%	67.83%
Salas-Gonzalez *et al.* [**224**]	FDG-PET	53, 52, 114	92% (leave-one-out CV) 87% (two-fold CV) 95.2% (best parameters)	-
Young *et al.* [**272**]	MRI, FDG-PET, APOE	63, 73, 143 (47, 96)	-	69.9%
Zhang *et al.* [**278**]	MRI, FDG-PET, CSF	45, 50, 91 (43, 48)	93.3%	73.9%
Zhang *et al.* [**279**]	MRI, FDG-PET	0, 0, 88 (38, 50)	-	78.4%

than the ones in reported in [**112**], showing an improvement from treating voxels as individual features, rather than averaging them within regions.

Salas-Gonzalez *et al.* [**224**] used a smaller subset of ADNI FDG-PET images and reported accuracies of 87% with two-fold cross-validation and 92% with leave-one-out cross-validation. Their result from two-fold cross-validation is close to the ones obtained in this chapter. They included voxel selection using the t-test method in their classification pipeline and additionally they combined and transformed voxel-wise features into a smaller feature set with factor analysis. They also report a 95.2% accuracy figure which appears to be the highest test accuracy achieved by adjusting the number of factors, but generally one should choose parameters by nested cross-validation, rather than doing it in this way, in order to avoid overfitting.

Methods for AD classification discussed in the literature are complex combinations of image processing, feature extraction and machine learning steps. Therefore, it is difficult to distinguish whether the performance of a particular method is due to accurate spatial normalisation, well-designed features, effective machine learning algorithms or a combination of these factors.

5.7.2. Spatial distribution of features. Several observations can be made based on the visualisations of individual classifiers. First, there are only small differences between SVM and logistic regression when these classifiers are used either directly or with ANOVA feature selection step preceding them. This pattern appears both in case of ℓ_2 and ℓ_1 regularisation. However, this is not surprising, as logistic regression and SVM cost functions can both be written in a $(loss)+(regularisation)$ format. For logistic regression the loss term is the binomial deviance ($\log\left[1+e^{-yf(x)}\right]$). For SVM the loss term is the hinge loss ($\left[1-yf\left(x\right)\right]_+$ with $\left[.\right]_+$ denoting the positive part). These functions have similar tails, while there is a difference close to the classification boundary [**123**].

The visual representations of classification pipelines using Ward feature agglomeration appear broadly similar across classification algorithms, although they are somewhat sparser for ℓ_1-regularised models. Meanwhile, for SLIC the appearance of the weight maps seems to vary, particularly between ℓ_2 and ℓ_1 regularisation.

With ℓ_1 regularisation, logistic regression and SVM are very sparse when trained on either raw voxels or ANOVA selected voxels. With only on a small number of voxels, these classifiers may be sufficiently accurate, but they are not as interpretable as the more dense ℓ_2-regularised models. In addition, they could be more sensitive to unusual anatomy not seen in training data, or to noisy images. It may be possible to combine the benefits of both ℓ_2 and ℓ_1 approaches by applying the elastic net regularisation [**288, 123**] which combines ℓ_2 and ℓ_1 terms. This could be an interesting topic for future work.

For all four linear classifiers, the weight maps are noticeably sparser when ANOVA feature selection is used compared to when these classifiers are applied to all voxels. This pattern is particularly visible in the case of ℓ_2-regularised classifiers. ANOVA feature selection in conjunction with ℓ_1 regularisation gives the sparsest feature maps of all the linear models, with similar maps for logistic regression and SVM.

When a random forest classifier is the only stage in the pipeline, or when a random forest is preceded by ANOVA feature selection, the distribution of feature importances appears to be fairly sparse. When feature agglomeration is applied before a random forest, the extent of relevant regions seems to be larger, particularly in the case of Ward agglomeration.

The voxels located in the precuneus seem to be important to all classifiers. This is consistent with the findings of other studies which found that the precuneus is one of the brain regions associated with hypometabolism in AD and amnestic MCI [**155, 205, 195**].

5.8. Conclusions

This chapter evaluated a number of dimensionality reduction methods (ℓ_1 regularisation, ANOVA, Ward agglomeration and SLIC agglomeration) in conjunction with several commonly used classification algorithms (SVM, logistic regression and random forests), applied to the task of classifying FDG-PET images as belonging either to AD patients or cognitively normal (CN) controls. The classifiers trained on AD and CN images were also used for prediction of whether MCI patients would progress to AD, based on FDG-PET images of their brains. It was found that there was no substantial benefit in terms of classification accuracy to applying the dimensionality reduction methods included in the comparisons. However, the dimensionality reduction methods evaluated in this chapter were rather simple, and more sophisticated methods may give better results.

A recent study by Chu *et al.* [**50**] investigated feature selection in the context of AD classification using MR images from ADNI. Their comparison included feature selection based on mass univariate t-tests, recursive feature elimination (RFE), predefined regions of interest (ROI) based on prior knowledge of regions affected by AD and selection of ROIs based on averaged absolute t-values within them. They concluded that using prior knowledge to select relevant brain regions improved classification accuracy but data-driven methods did not bring improvements or even made accuracies worse. In contrast, this chapter focused on FDG-PET images, only considered data-driven methods and did not include recursive feature elimination, while including ℓ_1 regularisation and two feature agglomeration algorithms.

However, other algorithms proposed in the literature often include feature selection and achieve good results [**130, 278, 279, 139, 170, 287, 286, 241, 171**]. This could indicate that feature selection is a useful component but not sufficient by itself to achieve state-of-the-art accuracy. Many of these methods also rely on features derived from multiple modalities, whereas the methods evaluated in this chapter are based mainly on FDG-PET features with MR images only used to help with

alignment of the FDG-PET images. Perhaps augmenting the input data with MRI features would improve the results. It could also be the case that the relatively simple dimensionality reduction methods evaluated in this chapter are not as effective as some of the more complex algorithms proposed in the literature on AD detection.

It is a strength of this study that not only the average accuracies of various classification methods, but also the variations of these averages across randomised partitions into training and test data were quantified. This enabled comparisons between different methods to assess whether any differences are substantial relative to the random variations.

The clustering algorithms used by the feature agglomeration methods evaluated in this chapter are unsupervised. Perhaps supervised clustering, such as the algorithm proposed by Fan *et al.* [**95**] as part of a method for MRI-based classification or the algorithm proposed by Michel *et al.* [**192**] for fMRI studies (which was only evaluated in this chapter in unsupervised form) would give better results. In addition, feature agglomeration was done using the arithmetic mean as the pooling function. For the ℓ_2-regularised linear classifiers one large coefficient is more "expensive" than a set of proportionally smaller coefficients distributed over a set of correlated features, so the features that were agglomerated may have lost their weight in deciding the classifier output. Therefore, an alternative pooling may be preferable, such as summation or summation followed by division by the square root of the cluster size.

This chapter evaluates both ℓ_1 and ℓ_2 regularisations but not the elastic net which combines both of them. It may be an intersting topic for future work to extend this evaluation to the elastic net [**288, 123**].

Finally, the spatial normalisation methods used to prepare data for this study require MR images to align the FDG-PET images. This is a weakness relative to methods that only require FDG-PET images, such as [**224**].

CHAPTER 6

Detection of Alzheimer's disease with scattering networks

6.1. Introduction

As already discussed in chapter 5, image-based classification is one of widely studied problems in medical imaging. A key challenge when developing an image classsification algorithm is to choose an effective feature representation. In this chapter scattering networks [**26, 190, 27**] are applied as a feature representation to image-based classification of Alzheimer's disease. We start with an overview of feature representations for AD detection with MR images. Then scattering networks are discussed in detail and an extension for 3D volumetric images is proposed. The challenge posed by high dimensionality of scattering representations is addressed with the fast Johnson-Lindenstrauss transform [**5**]. The proposed methods are evaluated in AD detection and prediction tasks on the ADNI data set.

6.2. Feature represenations for structural MRI

In their comparative evaluation, Cuignet *et al.* [**60**] distinguished three types of feature representations for AD detection with structural MR images: voxel-based, vertex-based and ROI-based.

Voxel-based features are defined on the level of individual voxels. Methods based on features of this type for structural MRI analysis often rely on voxel-based morphometry (VBM) [**9, 108, 68, 228**] to map the concentration of grey matter across the brain. VBM in its basic form starts with aligning all images to a template, in order to remove large-scale brain shape differences. This is followed by tissue segmentation, which partitions the brain volume into grey matter (GM), white matter (WM)

and cerebrospinal fluid (CSF). The GM probability map is then smoothed, which compensates for registration errors and makes the data distribution closer to Gaussian. The grey matter concentration map generated in this way is then analysed using statistical methods [**9**]. Optimised VBM is a more advanced protocol that adds several improvements, including more careful preparation of a study-specific template, spatial normalisation that is optimised for GM and WM, and a modulation step at the end (but before smoothing) that corrects GM and WM concentration for expansion or contraction due to spatial normalisation [**108**]. Voxel-wise VBM features are used in many Alzheimer's disease detection and prediction methods [**131, 94, 93, 257, 129, 194, 195, 60, 67, 97, 130, 50, 272, 174, 239**]. VBM can also be used as an intermediate feature representation. For example, Fan *et al.* [**95**] use VBM-like analysis to compute voxel-wise features and then apply a sophisticated dimensionality reduction method to obtain features which are subsequently used for clasification. They also assess the effectiveness of their method by training a classifer to distinguish schizophrenia patients from controls based on brain MR images.

Tensor-based morphometry (TBM) [**52**] can be used instead of VBM. TBM also requires images to be coregistered, but subsequently it relies only on examining the deformation fields that bring images into (non-rigid) alignment with a template. The partial derivatives of the deformation field in three directions are used to construct a Jacobian matrix field and the determinants of these matrices represent local contraction or dilation [**9**]. TBM is used in several AD classification and prediction methods proposed in the literature [**244, 130, 148, 138**].

Vertex-based features can be computed from structural MR images by mapping cortical thickness onto the cortical surface represented as a vertex mesh [**62, 99, 98**]. Features of this type are used in a number of image-based AD detection methods [**164, 13, 208, 73, 60, 49, 90**].

ROI-based features encompass features derived by measuring the volume or shape of specific anatomical structures [**57, 53, 106**]. In Alzheimer's disease classification the hippocampus is of particular interest, since it is known to be affected by atrophy early in AD [**136**]. Hippocampal volume and shape can be computed by segmenting an image, with shape represented as a decomposition into spherical harmonics [**106**]. Ewers *et al.* [**91**] investiagated volumes of the hippocampus, the entorhinal cortex and several other biomarkers as predictors of conversion from MCI to AD dementia. Longitudinal analysis of hippocampal atrophy was also used by Wolz *et al.* for AD detection and prediction of progression from MCI [**267**].

Several authors have proposed features such as GM volumes computed within a number of ROIs defined by anatomical segmentation [**280, 277, 278, 279, 176, 240, 171, 241, 287**]. Suk and Shen [**240**] developed this idea further by using ROI features as the input to a stacked autoencoder (a type of artificial neural network), which learns more advanced features that are then used for classification.

Recently several feature representations based on image patches were also proposed. Patches can be derived from image intensities directly [**59, 249**] or from intermediate features such as tissue density maps [**173, 174, 239**]. Patches can also be used to learn more advanced features with deep learning [**239**].

Wavelet methods for structural brain image analysis were also proposed in the literature. Lao *et al.* [**156**] proposed a wavelet representation of VBM-type images and applied this method in several settings: simulated atrophy images, classification between male and female brains, and classification of brain images into age groups. Canales-Rodriguez *et al.* [**33**] proposed a wavelet-based modification of the VBM pipeline where statistical analysis is done in the wavelet domain and applied it to a data set where images of one group were modified to simulate cortical thinning. Hackmack *et al.* [**117**] developed a method based on the magnitude representation of the dual-tree complex wavelet transform [**147, 226**] for detection of multiple sclerosis (MS) using structural MR images. They also validated their method using images of

AD patients and healthy controls. The three-dimensional directional wavelets used in their work are conceptually similar to those used in this chapter. Chaplot *et al.* [**44**] proposed classifying brain images using the discrete wavelet transform (DWT) as a feature representation and the support vector machine (SVM) as a classifier.

Finally, some AD detection algorithms proposed in the literature combine different types of features into hybrid models [**177, 268, 264, 275**].

6.3. Scattering networks

Scattering networks were recently introduced as a new wavelet technique for signal processing [**26, 190, 27**]. A scattering network can be represented as a tree of filter banks, as shown in figure 6.3.1. Each filter bank consists of a low-pass filter and several band-pass filters. The output of each low-pass filter constitutes part of the network's output, and the complex muduli of the outputs of the band-pass filters are processed recursively by further filter banks. This construction is continued until a defined maximum depth of the tree is reached. Filter banks in scattering networks are typically based on complex Morlet wavelets (figure 6.3.2).

Scattering networks are particularly suitable for defining feature representations in medical image-based machine learning problems because a network can be designed with the properties of translation invariance and Lipshitz continuity with respect to deformations [**27**]. Rotation invariance can be added with a more complex network structure [**230, 232**] and scaling invariance can be included as well [**231**].

Scattering networks have been applied successfuly in image classfication setings, with state of the art results for handwritten digits [**26, 27**] as well as textures [**26, 230, 27, 231, 232**]. Other applications include audio classification [**7, 8**] and classification of physiological signals [**51**]. A scattering transform was also defined on graphs using Haar wavelets [**46**].

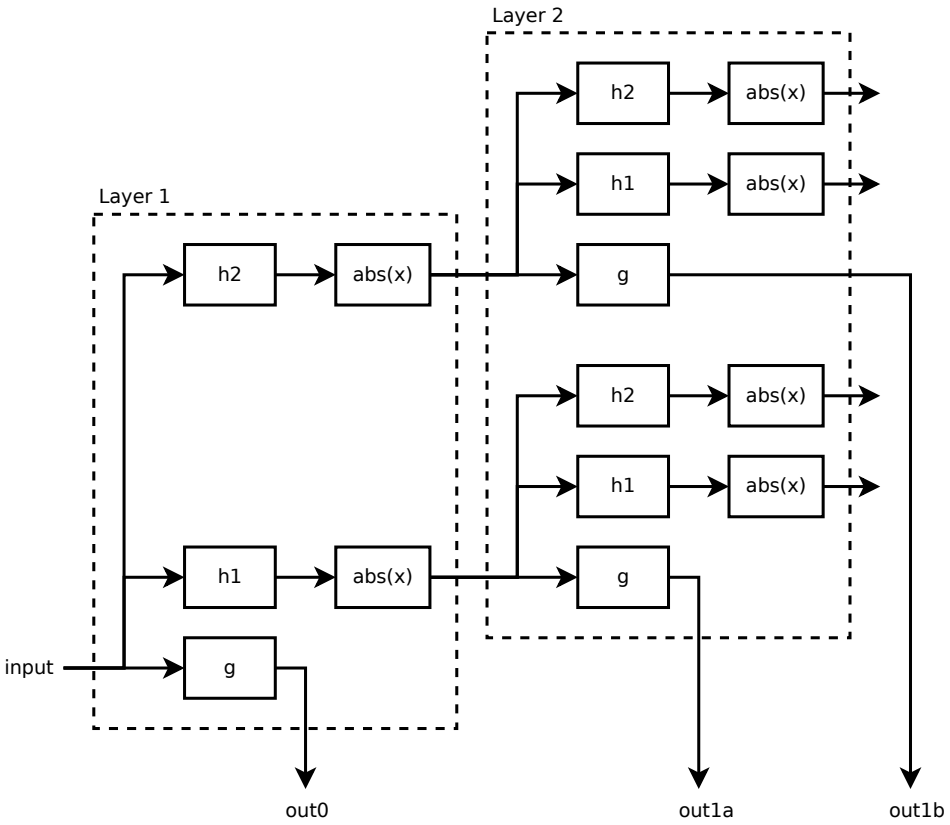


Figure 6.3.1: A simple scattering network. Each filter bank consists of the filters g (low-pass), h0 and h1 (both band-pass), which separate the image into different spatial frequency bands. The low-pass signal is part of the network's output. The band-pass signals are further processed with the complex modulus operator, producing inputs to the next network layer. The number of layers is a pre-defined parameter of the network.

In the context of medical imaging, scattering networks are particularly interesting because they construct feature representations that change continuously with image deformations [**27**]. This means that shapes of anatomical structures can be represented implicitly, without the need for complex shape modeling.

Scattering networks for tomographic images. Scattering networks have been designed so far to process one-dimensional signals [**7, 8, 51**] and two-dimensional images [**26, 230, 27, 231, 232**]. However, working with tomographic images requires a scattering network that can process three-dimensional images. This type of network can be defined by extension from the two-dimensional algorithm of Bruna and Mallat [**26, 27**], which is done in the following.

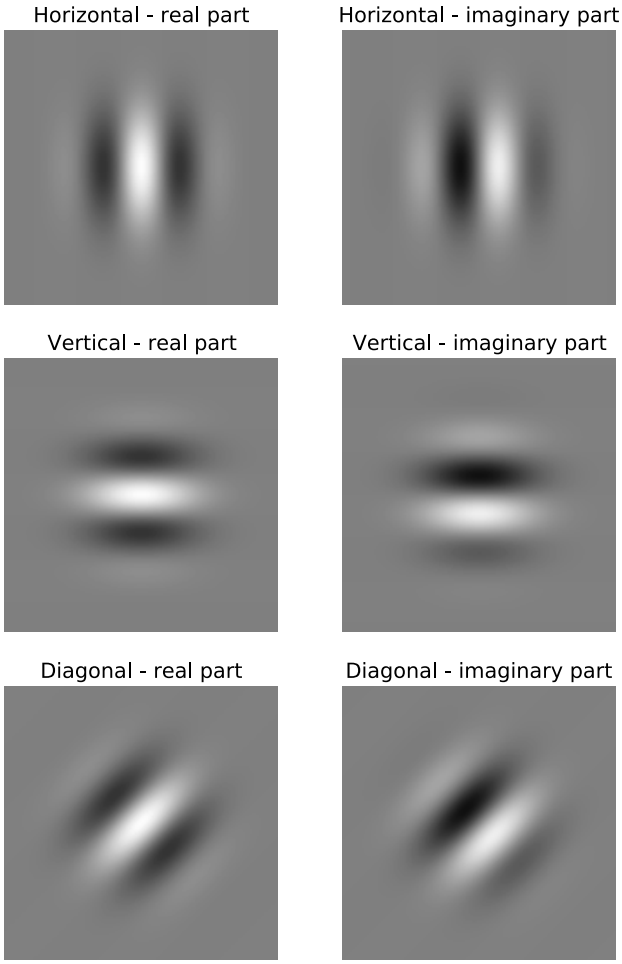


Figure 6.3.2: Two-dimensional Morlet wavelets with three spatial orientations: horizontal, vertical and diagonal.

In the 2D case the filter banks are usually built with Morlet wavelets, *i.e.* complex exponentials modulated with a Gaussian window and then mean-subtracted to meet the wavelet admissibility criteria. The filters are directional, with their orientations distributed evenly on a circle. The forms of the ϕ (low-pass) and ψ_k (band-pass, with $0 \leq k < K$) filters can be written in a vector format, as a function of $\mathbf{u} = (x, y)^T$:

$$\phi(\mathbf{u}) = C_\phi e^{-\left\|\mathbf{u}^T \boldsymbol{\Sigma}^{-1} \mathbf{u}\right\|^2}$$

$$\psi_k(\mathbf{u}) = C_{\psi_k} \left(e^{i\mathbf{u}\cdot\xi_k} - \beta_k\right) e^{-\left\|\mathbf{u}^T \boldsymbol{\Sigma}_k^{-1} \mathbf{u}\right\|^2}, \quad 0 \leq k < K$$

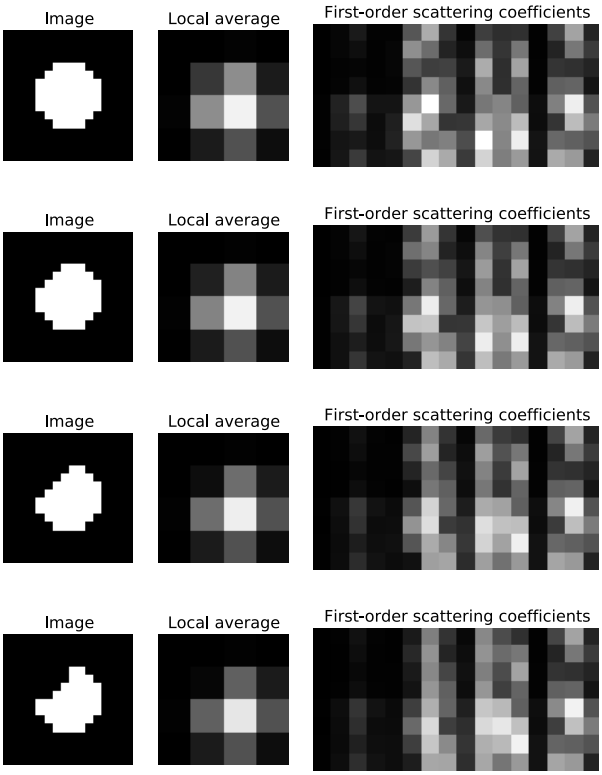


Figure 6.3.3: Scattering and deformation example. As the circle is deformed, the local averages and first-order scattering coefficients change in a gradual way. (The first-order scattering coefficients were re-arranged into a matrix for display and do not correspond spatially to their respective images.)

where C_ϕ and C_{ψ_k} are normalisation constants and β_k is adjusted to ensure that the sum (integral) of $\psi_k(\mathbf{u})$ over its domain is zero[1]. The parameters $\mathbf{\Sigma}$ and $\mathbf{\Sigma}_k$ (matrices, controlling shapes and spatial extents od the Gaussian envelopes) and ξ_k (vector, peak frequency different for each filter in the bank) have to be chosen as part of the design process.

With a vector formulation in 2D, the 3D version is a straightforward extension: re-define $\mathbf{u} = (x, y, z)^T$ and choose $\mathbf{\Sigma}$ and $(\mathbf{\Sigma}_k, \xi_k), k \in 0 \leq k < K$.

[1] *N.b.* the imaginary part of a complex exponential is odd and the windowing Gaussian is even, so the imaginary part of the sum (integral) vanishes. This leaves the real part which is annihalated with the right choice of β.

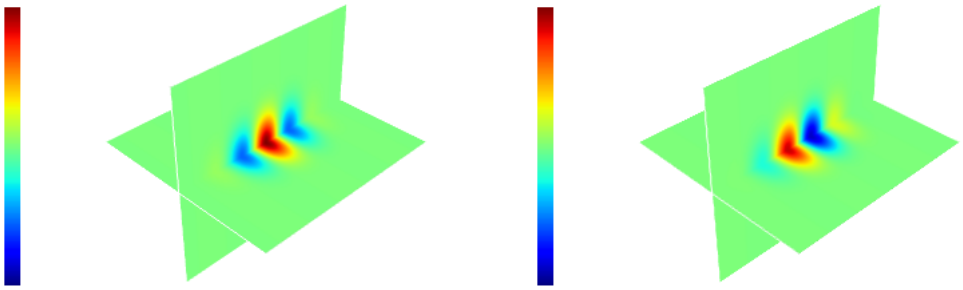


Figure 6.3.4: Three-dimensional Morlet wavelet (left: real part, right: imaginary part). Two orthogonal cross-sections are shown in each plot.

While it is straightforward to choose any number of filter orientations equidistributed on a circle, with a sphere (in the 3D case) this is more difficult. One option is to use the vertices or faces of Platonic solids, although this only works for up to 20 points (dodecahedron, with 20 vertices, is the Platonic solid with the most vertices). Alternatively, one can use one of several algorithms (*e.g.*[**72**]) that produce points approximately evenly distributed on a sphere. However, defining a large number of orientations for the filters is not practical because the computational complexity of the network grows quickly with the number of orientations (the total number of filters in the network is $K^{\bar{m}}$). Therefore, filters oriented towards vertices of Platonic solids were used in this work. One of the platonic solids (the regular icosahedron) is shown in figure 6.3.5.

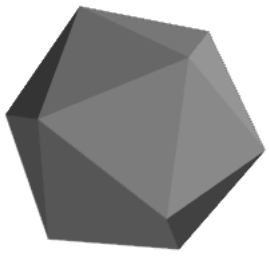


Figure 6.3.5: Regular icosahedron, one of the Platonic solids.

With the filter banks defined as above, the scattering network is constructed according to the standard architecture (figure 6.3.1).

It is worth noting that, as long as only real-valued signals are considered, pairs of filters oriented towards opposite points on the unit circle (2D) or unit sphere (3D) give responses with values that are complex conjugates of each other (same as with Fourier transforms). Therefore, computational requirements can be reduced by computing the response of only one filter from each pair.

6.4. Fast classifier training on high-dimensional data

Image descriptions generated with scattering networks can capture subtle characteristcs of shape and texture, but the generated output is extremely high-dimensional. This is particularly the case for 3D tomographic images where a large number of directional filters is required to cover the unit sphere. When the number of features excedes the number of samples in the data set by orders of magnitude, the risk of overfitting is an important consideration. This means that simple and highly regularised algorithms are preferred [**123**]. Therefore, logistic regression with ℓ_2 regularisation was chosen for the experiments conducted in this study.

Machine learning algorithms applied directly to scattering features require large amounts of computer memory and time for training. With a linear classification algorithm it is natural to consider a linear dimensionality reduction step to make classifier training faster and more memory-efficient. Intuitively, as long as the pairwise distances between samples in the reduced feature space are close to those in the original feature space, the distribution of the data is approximately preserved and the classifiers trained on native and dimensionality-reduced feature vectors should be in close agreement.

In the experiments described in this chapter the fast Johnson-Lindenstrauss transform (FJLT) proposed by Ailon and Liberty [**5**] was used for dimensionality reduction. In order to apply this transform to feature vectors with dimensions that are not powers of two, the Hadamard transform in FJLT was substituted with the discrete cosine transform (DCT). It is important to note that, to the author's best knowledge, the construction based on DCT has not been verified mathematically in the same way as the Hadamard formulation, although a similar construction based on the Fourier transform was addressed by Krahmer and Ward [**151**].

In order to compare the Hadamard-based FJLT and DCT-based FJLT, a simple experiment was done with sets of vectors generated from Gaussian and exponential distributions, with each component independent and identically distributed (i.i.d). For each distribution a dataset D consisting of 1000 randomly generated vectors in 65536 dimensions was generated. These vectors were then reduced to a lower dimensionality with both Hadamard-based FJLT and DCT-based FJLT. The maximum approximation error was then computed as

$$\max_{\mathbf{x},\mathbf{y}\in D} \frac{|\|\mathbf{x}_t - \mathbf{y}_t\|_2 - \|\mathbf{x} - \mathbf{y}\|_2|}{\|\mathbf{x} - \mathbf{y}\|_2}$$

where the vectors $\mathbf{x}, \mathbf{y}$ are in original space and $\mathbf{x}_t, \mathbf{y}_t$ are their transformed versions. The results are shown in figure 6.4.1.

It appears that in this small experiment the Hadamard-based FJLT and DCT-based FJLT have nearly identical performance. However, it is also important to note that real data is unlikely to have independent identically distributed components and indeed it is likely that the samples will be correlated with each other. Therefore, this small experiment is not a sufficient replacement for a thorough mathematical proof.

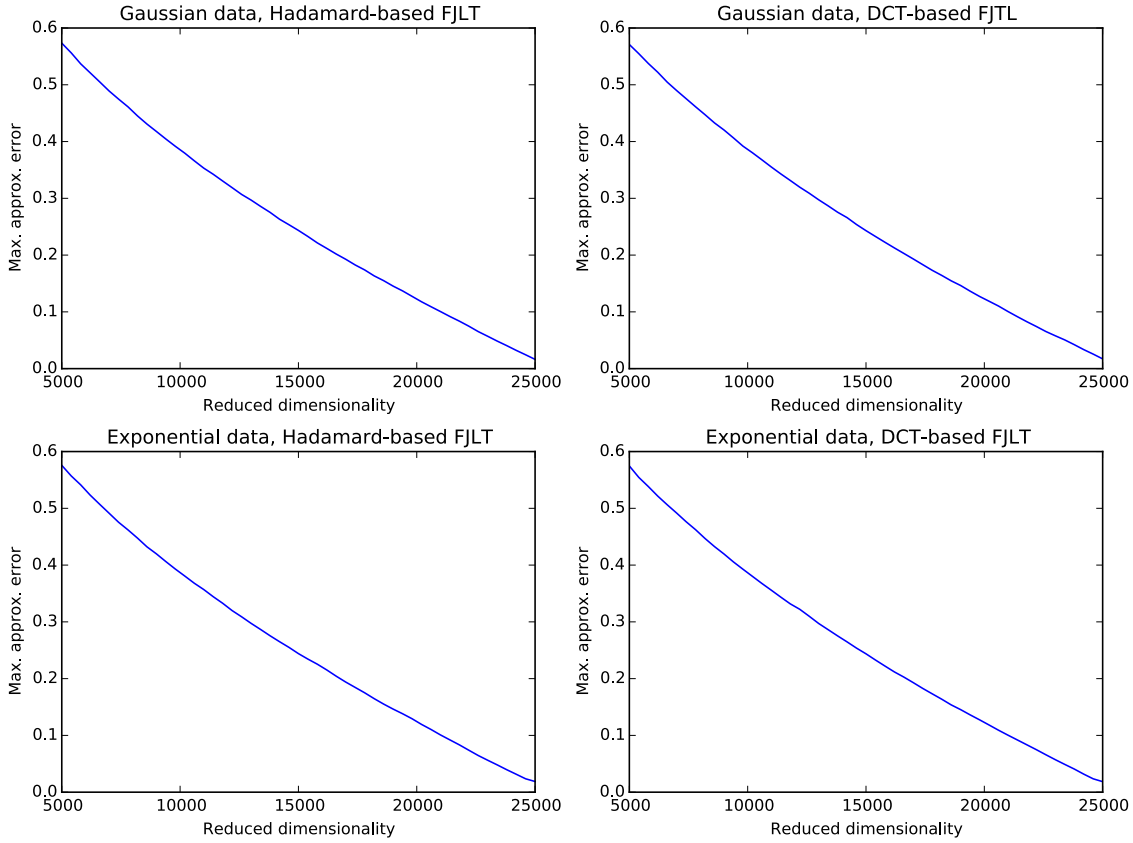


Figure 6.4.1: Hadamard-based FJLT and DCT-based FJLT. Both transformations were computed for two data sets (Gaussian and exponential vectors of i.i.d. random variables) for a range of target dimensionalities.

It is worth noting that there are several alternatives to the dimensionality-reduced classification method outlined above. Cannings and Samworth [**43**] recently proposed a classification algorithm that combines the decisions produced by an arbitrary base classifier applied to many random projections of the data, with the set of random projections pruned to retain only those that produce the smallest estimate of test error. Fern and Brodley [**96**] earlier proposed a clustering algorithm that combines the results of many runs of a base clustering algorithm, each computed on a random projection of the data. Dasgupta and Freund [**63**] proposed an algorithm for hierarchical vector quantization which also relies on random projections. Compressive classification, where the data from a compressive sensor is used directly for classification was also proposed in compressed sensing literature [**124, 69, 83, 215**] and those methods can be used for efficient classififcation of high-dimensional signals

by simulating the compressive sensor with a random projection. The technique of applying machine learning algorithms to data that was dimesionality-reduced with random projections was also used to address problems including texture classification [**172**] and object tracking [**281**]. In addition, there are similarities between the proposed compressive classification method and the algorithms used for approximating kernel expansions [**209, 159**], where randomised projections are used for a similar purpose.

6.5. Experiments

The data used to evaluate scattering representations for image-based classification consisted of the following two data sets derived from the ADNI database.

(1) MRI T1-weighted intensity images, taken at baseline.
(2) TBM Jacobian determinant maps of deformation fields generated by registering 24-month follow-up MR T1-weighted images to baseline MR T1-weighted images of respective subjects.

Both data sets originate from projects conducted by other researchers [**114, 259, 138**]. In the following, more detail is given on the processing steps that were used to generate them.

6.5.1. MR T1 intensities. The data used in these experiments consists of baseline MR T1-weighted images from the ADNI database. These images were originally processed by Guerrero *et al.* for the work presented in [**114**]. Processing applied to images available from ADNI by Guerrero *et al.* consisted of brain extraction with "pincram" (pyramidal intra-cranial masking, similar to [**126**]) and affine alignment to the MNI152 template. These images were then divided into different categories as follows.

Patients who reverted from AD to MCI or from MCI to cognitive normal (CN) at any point were excluded from the analysis because they did not fall clearly into any of the categories described in the following, and additionally because these patients may have been cases of misdiagnosis. The subjects who were diagnosed as AD or CN at baseline were labeled as such. Among the MCI patients, those who were diagnosed as MCI at month 36 were labeled as stable MCI (sMCI) and those who were diagnosed as AD at month 36 were labeled as progressive MCI (pMCI). Patients who were diagnosed with MCI at baseline but had no month 36 scan available were excluded. Finally, there were several patients with baseline scans missing in the data set and these patients were also excluded. The final data set consisted of 759 subjects, including 254 AD, 277 CN, 113 pMCI and 115 sMCI.

To apply scattering networks to this data, scattering representations were computed using Morlet wavelets and then their dimensionality was reduced to 25000 using the FJLT with DCT as the core transform. Key parameters of the scattering network (depth, number of scales in the filter banks, and number of directional axes to tune the filters to) were varied to evaluate their importance for classification performance.

Voxel intensity features were used as a benchmark for evaluating scattering features. Only the values of voxels falling within a brain mask were extracted from the images in this case. The dimensionality of these feature vectors was still impractical to work with, so it was reduced to 25000 in the same way as it was done for the scattering features.

Scattering features as well as voxel features can be normalised to zero mean and unit standard deviation before being dimensionality-reduced. Variants with and without normalisation were both evaluated.

6.5.2. Jacobian determinant maps. This data set was adapted from the work of Vounou *et al.* [**259**] and Janousova *et al.* [**138**] and the following summary is based on the information provided in [**259**]. The data set consists of 510 subjects

for whom both baseline and 24 month follow-up MR images were available from ADNI by October 2010. The follow-up images were non-rigidly registered to their respective baseline scans using a B-spline registration algorithm [**221**]. This was done in a coarse-to-fine progression (20mm control point spacing, then 10mm, 5mm and finally 2.5mm). The Jacobians of the resultant deformation fields quantify the voxel-wise expansion or contraction between the baseline and follow-up scan [**22**]. Baseline scans were also aligned with the MNI152 template using non-rigid registration with 10mm control point spacing, giving deformation fields that were used to warp the Jacobian maps to a common space. This data set was available complete with labels distinguishing AD, CN, pMCI and sMCI subjects. It consisted of 510 subjects, including 105 AD, 165 CN, 117 pMCI and 123 sMCI.

Voxel-wise Jacobian determinant maps were processed with scattering networks in the same way as described above for MR T1-weighted intensity inages. As a benchmark, voxel-wise Jacobian determinants were also used directly as features, with FJLT dimensionality reduction to 25000 dimensions.

Table 6.5.1: Summary statistics of the data sets.

Dataset	Number of subjects				
	AD	CN	pMCI	sMCI	Total
MR T1 intensity	254	277	113	115	759
Jacobian	105	165	117	123	510

Both types of features were evaluated with and without feature normalisation (to zero mean and unit standard deviation) before the dimensionality reduction step.

6.5.3. Classification. Logistic regression with ℓ_2 regularisation was used as the classification algorithm in all experiments reported in this chapter. FJLT approximately preserves pair-wise distances between data points, so a linear classifier is suitable for classifying data that was dimensionality-reduced with this method.

The work of Chu *et al.* [**50**] has shown that feature selection does not significantly improve classification accuracy of AD with MR images. This is consistent with the

results obtained in Chapter 5 of this dissertation for FDG-PET images. Therefore, no feature selection was applied in the experiments done (with MR images) for this chapter.

6.5.4. Cross-validation. The set of AD and CN subjects was randomly partitioned into 75% training data and 25% test data, in a stratified way. Training data was used to train the classifier and test data was used for evaluation. This process was repeated 100 times to give 100 classifiers with their respective performance figures on held-out test data. Each one of those 100 classifiers was then also tested for separating pMCI patients from sMCI patients in the MCI subset, based on their baseline scans.

6.6. Results

The two data sets (MRI T1 intensity and Jacobian) with two classifier options (with and without normalisation) resulted in four sets of performance figures. Each of these sets of results further consists of results for voxel-wise features and scattering features with different scattering network parameters.

Four performance measures were estimated for each configuration, both for AD vs CN classification and for pMCI vs sMCI classification: accuracy, sensitivity, specificity and the area under the ROC curve (AUC). The sample means of all these metrics estimated from the 100 randomised repetitions of the experiments are presented in tables. In addition, classification accuracy is visualised with box plots.

Table 6.6.1 and figure 6.6.1 show results for MR T1-weighted images *without* feature normalisation. In this case scattering features clearly improve on voxel-wise features, both in AD vs CN classification and in MCI prediction, regardless of the particular setting of scattering parameters. Average accuracy in AD vs CN classification for all scattering configurations is 68.45%, compared to 60.81% for voxel-wise features. In MCI prediction, average accuracy for all scattering configurations is 57.68%,

compared to 54.57% for voxel-wise features. Best mean AD vs CN accuracy is 69.26%, achieved with scattering features with $D = 2$, $J = 2$ and $L = 6$. Best mean MCI prediction accuracy is 58.36%, achieved with scattering features with $D = 2$, $J = 2$ and $L = 10$. It appears that the particular choice of values for D, J and L has only a small effect on the results.

Table 6.6.2 and figure 6.6.2 show results for MR T1-weighted images *with* feature normalisation. It appears that adding feature normalisation makes the results worse, particularly for the scattering feature representations. However, scattering features still have higher performance figures than voxel-wise features. Average accuracy in AD vs CN classification for all scattering configurations is 65.25%, compared to 60.29% for voxel-wise features. In MCI prediction, average accuracy for all scattering configurations is 56.58%, compared to 53.60% for voxel-wise features. Best mean AD vs CN accuracy is 66.97%, achieved with $D = 2$, $J = 2$ and $L = 6$. Best mean MCI prediction accuracy is 57.71%, achieved with $D = 1$, $J = 1$ and $L = 3$.

Table 6.6.3 and figure 6.6.3 show results for Jacobian images *without* feature normalisation. Average accuracy in AD vs CN classification for all scattering configurations is 77.43%, compared to 72.26% for voxel-wise features. In MCI prediction, average accuracy for all scattering configurations is 61.95%, compared to 59.95% for voxel-wise features. Best mean AD vs CN accuracy is 77.87%, achieved with scattering features with $D = 2$, $J = 2$ and $L = 3$. Best mean MCI prediction accuracy is 62.92%, achieved with scattering features with $D = 2$, $J = 2$ and $L = 10$. The particular choice of D, J and L appears to have only a small effect on the results, especially in AD/CN classification.

Table 6.6.4 and figure 6.6.4 show results for Jacobian images *with* feature normalisation. Average accuracy in AD vs CN classification for all scattering configurations is 76.70%, compared to 71.53% for voxel-wise features. In MCI prediction, average accuracy for all scattering configurations is 62.74%, compared to 61.88% for voxel-wise features. In this case normalisation seems to make classification accuracies

Table 6.6.1: Classification results for MRI T1 data *without feature normalisation*. Means are given with standard errors in parentheses.

Method	AD vs CN				pMCI vs sMCI			
	Acc [%]	Sens [%]	Spec [%]	AUC	Acc [%]	Sens [%]	Spec [%]	AUC
Voxels	60.81 (0.36)	55.02 (0.74)	66.19 (0.57)	0.6303 (0.0043)	54.57 (0.21)	45.35 (0.51)	63.63 (0.34)	0.5500 (0.0020)
D1_J1_L3	67.97 (0.36)	64.33 (0.59)	71.35 (0.53)	0.7241 (0.0036)	57.49 (0.17)	49.40 (0.31)	65.44 (0.25)	0.5901 (0.0012)
D1_J1_L6	68.03 (0.38)	64.19 (0.63)	71.59 (0.53)	0.7222 (0.0038)	57.32 (0.15)	49.18 (0.30)	65.31 (0.26)	0.5983 (0.0013)
D1_J1_L10	68.23 (0.41)	64.33 (0.70)	71.84 (0.51)	0.7207 (0.0037)	57.47 (0.17)	50.30 (0.30)	64.52 (0.30)	0.5946 (0.0012)
D1_J2_L3	68.18 (0.42)	64.20 (0.71)	71.87 (0.59)	0.7249 (0.0040)	57.98 (0.19)	49.91 (0.33)	65.90 (0.27)	0.5971 (0.0014)
D1_J2_L6	68.17 (0.43)	64.70 (0.69)	71.39 (0.57)	0.7272 (0.0039)	57.60 (0.20)	50.38 (0.33)	64.70 (0.26)	0.5929 (0.0013)
D1_J2_L10	68.39 (0.40)	65.06 (0.63)	71.48 (0.55)	0.7284 (0.0037)	58.19 (0.18)	50.65 (0.30)	65.59 (0.27)	0.5978 (0.0013)
D2_J1_L3	68.59 (0.41)	64.33 (0.67)	72.54 (0.55)	0.7246 (0.0038)	57.04 (0.18)	46.88 (0.34)	**67.03 (0.23)**	0.5975 (0.0013)
D2_J1_L6	68.92 (0.34)	65.48 (0.61)	72.12 (0.48)	0.7267 (0.0033)	57.63 (0.16)	49.79 (0.32)	65.34 (0.26)	0.5940 (0.0012)
D2_J1_L10	68.27 (0.40)	64.98 (0.68)	71.32 (0.53)	0.7246 (0.0037)	57.33 (0.16)	48.26 (0.31)	66.24 (0.28)	0.6006 (0.0013)
D2_J2_L3	68.71 (0.38)	**65.84 (0.67)**	71.38 (0.56)	0.7312 (0.0036)	57.98 (0.18)	**50.93 (0.34)**	64.90 (0.23)	0.5994 (0.0012)
D2_J2_L6	**69.26** (0.40)	65.41 (0.65)	**72.84** (0.54)	**0.7358 (0.0036)**	57.72 (0.16)	49.93 (0.33)	65.38 (0.22)	0.5999 (0.0011)
D2_J2_L10	68.66 (0.40)	65.47 (0.66)	71.62 (0.56)	0.7304 (0.0037)	**58.36 (0.17)**	49.86 (0.30)	66.70 (0.25)	**0.6014 (0.0014)**

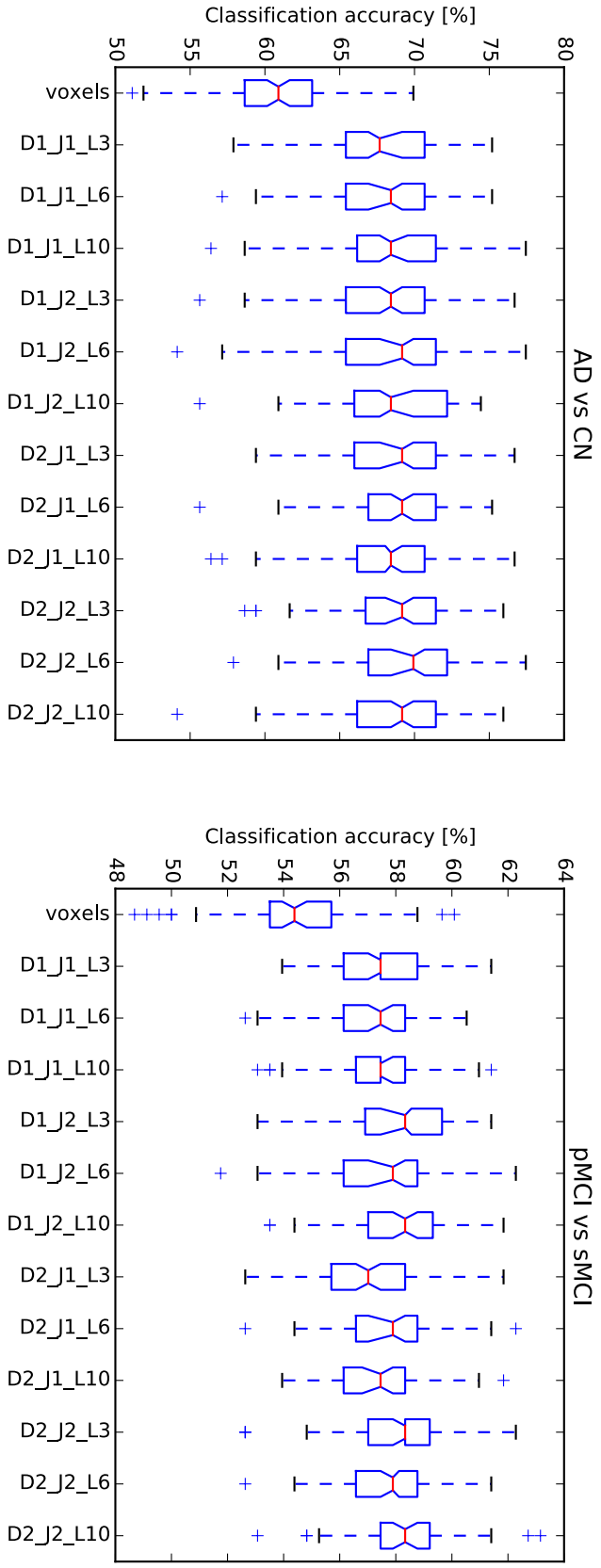


Figure 6.6.1: Classification accuracies for MRI T1 data *without feature normalisation.*

Table 6.6.2: Classification results for MRI T1 data *with normalised features*. Means are given with standard errors in parentheses.

Method	AD vs CN				pMCI vs sMCI			
	Acc [%]	Sens [%]	Spec [%]	AUC	Acc [%]	Sens [%]	Spec [%]	AUC
Voxels	60.29 (0.37)	53.30 (0.65)	66.78 (0.69)	0.6311 (0.0046)	53.60 (0.21)	40.94 (0.37)	**66.03 (0.31)**	0.5571 (0.0018)
D1_J1_L3	63.80 (0.35)	59.91 (0.58)	67.42 (0.60)	0.6786 (0.0037)	**57.71 (0.19)**	**51.48 (0.41)**	63.83 (0.29)	0.5961 (0.0014)
D1_J1_L6	64.29 (0.33)	61.55 (0.54)	66.84 (0.54)	0.6972 (0.0036)	56.45 (0.18)	48.12 (0.36)	64.63 (0.37)	0.5800 (0.0013)
D1_J1_L10	64.52 (0.34)	61.52 (0.59)	67.30 (0.49)	0.6877 (0.0033)	55.15 (0.17)	46.35 (0.28)	63.79 (0.33)	0.5755 (0.0012)
D1_J2_L3	65.08 (0.34)	60.36 (0.58)	**69.46 (0.52)**	0.6961 (0.0031)	56.74 (0.16)	48.05 (0.34)	65.27 (0.26)	0.5815 (0.0015)
D1_J2_L6	65.38 (0.40)	62.39 (0.59)	68.16 (0.54)	0.7009 (0.0036)	56.61 (0.17)	47.68 (0.31)	65.37 (0.27)	0.5889 (0.0015)
D1_J2_L10	65.41 (0.33)	61.77 (0.54)	68.80 (0.55)	0.7058 (0.0034)	55.21 (0.14)	45.26 (0.26)	65.00 (0.27)	0.5747 (0.0013)
D2_J1_L3	65.41 (0.33)	63.28 (0.57)	67.39 (0.56)	0.6996 (0.0036)	57.62 (0.18)	49.60 (0.34)	65.50 (0.22)	**0.6060 (0.0012)**
D2_J1_L6	65.27 (0.37)	63.94 (0.53)	66.51 (0.56)	0.7092 (0.0038)	57.67 (0.16)	49.88 (0.30)	65.32 (0.27)	0.6014 (0.0014)
D2_J1_L10	65.50 (0.38)	65.12 (0.56)	65.86 (0.57)	0.7102 (0.0038)	57.11 (0.16)	48.92 (0.31)	65.17 (0.23)	0.5988 (0.0013)
D2_J2_L3	66.43 (0.32)	64.36 (0.55)	68.35 (0.55)	0.7197 (0.0030)	55.67 (0.20)	46.24 (0.33)	64.94 (0.29)	0.5869 (0.0014)
D2_J2_L6	**66.97 (0.37)**	**65.14 (0.56)**	68.67 (0.59)	**0.7218 (0.0036)**	56.91 (0.17)	49.50 (0.26)	64.19 (0.28)	0.5915 (0.0012)
D2_J2_L10	64.92 (0.41)	63.09 (0.61)	66.62 (0.61)	0.7007 (0.0038)	56.16 (0.17)	47.02 (0.31)	65.14 (0.26)	0.5856 (0.0015)

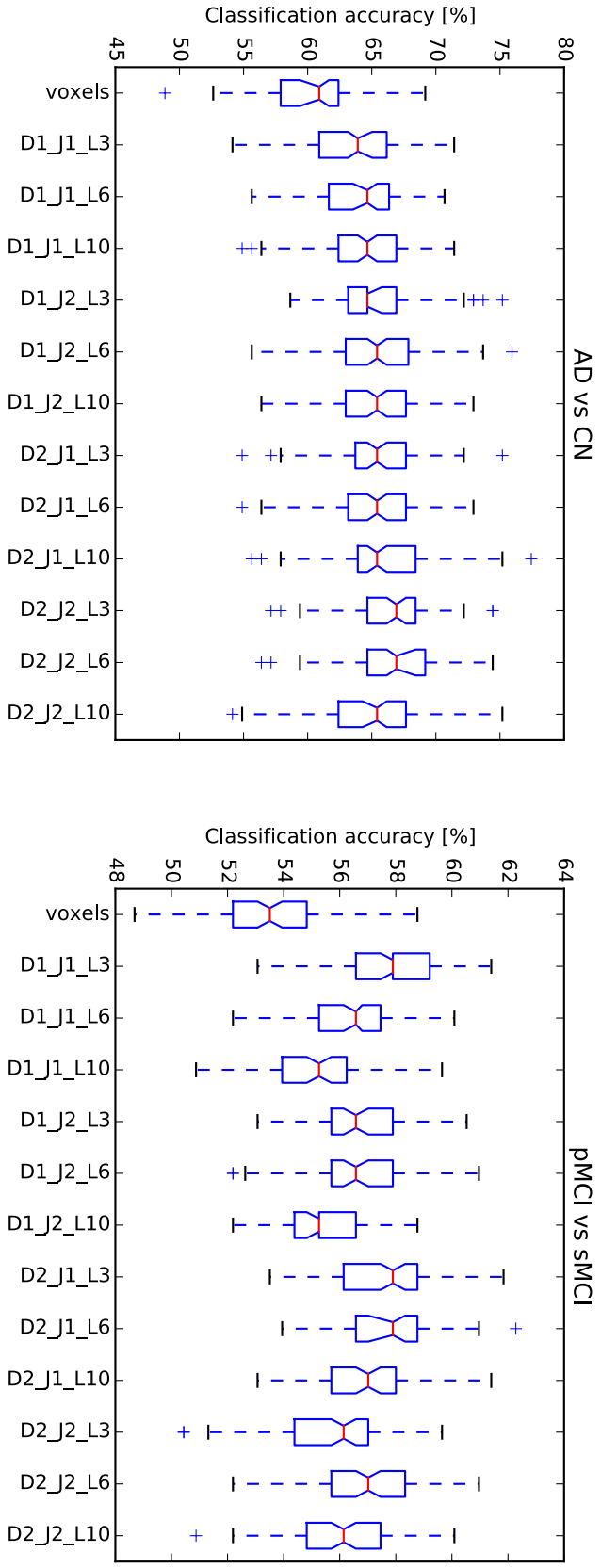


Figure 6.6.2: Classification accuracies for MRI T1 data *with normalised features*.

Table 6.6.3: Classification results for MRI Jacobian data *without feature normalisation*. Means are given with standard errors in parentheses.

Method	AD vs CN				pMCI vs sMCI			
	Acc [%]	Sens [%]	Spec [%]	AUC	Acc [%]	Sens [%]	Spec [%]	AUC
Voxels	72.26 (0.42)	46.77 (0.90)	88.05 (0.60)	0.7442 (0.0051)	59.95 (0.17)	36.72 (0.58)	82.05 (0.33)	0.6635 (0.0015)
D1_J1_L3	76.96 (0.38)	55.77 (0.77)	90.07 (0.54)	0.8345 (0.0039)	61.78 (0.17)	40.50 (0.36)	82.02 (0.27)	0.6675 (0.0014)
D1_J1_L6	77.26 (0.40)	56.08 (0.81)	90.38 (0.55)	**0.8453 (0.0039)**	61.21 (0.18)	39.68 (0.36)	81.69 (0.25)	0.6651 (0.0013)
D1_J1_L10	77.66 (0.38)	56.54 (0.80)	90.74 (0.50)	0.8433 (0.0042)	61.90 (0.19)	41.20 (0.38)	81.59 (0.25)	0.6712 (0.0014)
D1_J2_L3	77.04 (0.38)	56.69 (0.82)	89.64 (0.54)	0.8296 (0.0041)	62.28 (0.18)	41.96 (0.42)	81.61 (0.27)	0.6708 (0.0014)
D1_J2_L6	77.72 (0.36)	57.85 (0.83)	90.02 (0.51)	0.8360 (0.0039)	62.55 (0.17)	43.09 (0.38)	81.07 (0.27)	0.6712 (0.0014)
D1_J2_L10	77.37 (0.34)	56.92 (0.76)	90.02 (0.52)	0.8297 (0.0040)	62.38 (0.18)	42.36 (0.45)	81.42 (0.30)	0.6722 (0.0014)
D2_J1_L3	77.53 (0.37)	56.12 (0.75)	**90.79 (0.50)**	0.8436 (0.0039)	61.27 (0.16)	40.64 (0.36)	80.89 (0.26)	0.6658 (0.0013)
D2_J1_L6	77.35 (0.36)	56.19 (0.83)	90.45 (0.48)	0.8395 (0.0041)	61.75 (0.17)	40.95 (0.37)	81.53 (0.27)	0.6629 (0.0015)
D2_J1_L10	77.04 (0.37)	56.00 (0.84)	90.07 (0.49)	0.8433 (0.0040)	61.39 (0.17)	39.59 (0.37)	**82.12 (0.25)**	0.6670 (0.0013)
D2_J2_L3	**77.87 (0.38)**	58.38 (0.82)	89.93 (0.51)	0.8361 (0.0041)	62.35 (0.18)	42.80 (0.39)	80.95 (0.24)	0.6733 (0.0015)
D2_J2_L6	77.56 (0.37)	57.42 (0.79)	90.02 (0.54)	0.8341 (0.0040)	61.64 (0.18)	41.84 (0.39)	80.47 (0.27)	0.6681 (0.0015)
D2_J2_L10	77.81 (0.35)	**57.92 (0.76)**	90.12 (0.49)	0.8392 (0.0038)	**62.92 (0.17)**	**43.92 (0.36)**	80.98 (0.23)	**0.6755 (0.0014)**

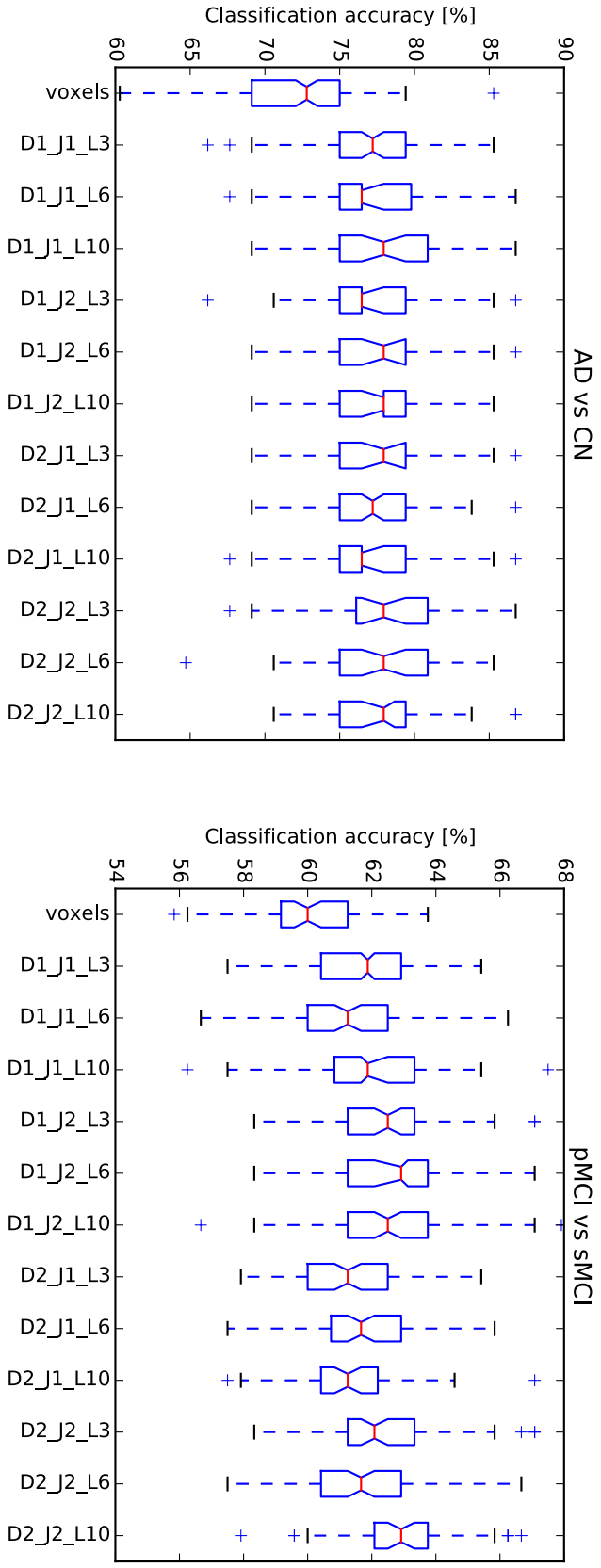


Figure 6.6.3: Classification accuracies for MRI Jacobian data *without feature normalisation.*

more widely distributed, both in AD vs CN classification and MCI prediction. Best mean AD/CN classification accuracy is 79.59%, achieved with $D = 1$, $J = 1$ and $L = 3$ (the simplest scattering network), and best mean MCI prediction accuracy is 65.50%, with the same scattering network. Simpler scattering networks seem to perform better than the ones with more filters or more scattering layers.

Relative to MR T1-weighted images, Jacobian images give much better results, both in case of scattering features and voxel-wise features. However, these two datasets are not directly comparable because they have different numbers of subjects in individual groups. In particular, the Jacobian data set has an uneven balance of AD and CN subjects (105 AD and 165 CN), which would give a classifier that always predicts CN expected accuracy of 61.11% on the AD/CN task.

6.7. Discussion

The experimental results presented in the previous section show that transforming images with scattering networks can improve classification performance. Improvements were observed for spatially normalised T1-weighted images as well as longitudinal Jacobian maps. Normalising scattering features led to slightly worse AD vs CN classification accuracy for both data sets. When the classifiers trained on AD and CN data were applied to prediction of MCI progression, the classification pipelines including the feature normalisation step performed slightly worse on T1-weighted data and slightly better on Jacobian data.

The particular choice of parameters for the scattering network (D, J and L) seems to have a limited effect on classification accuracy. It is surprising that networks with one layer are so close in performance to networks with two layers. This seems to indicate that most of the relevant features are computed by the first layer. One could argue that the features computed by the second layer are disadvantaged because their amplitide is small relative to the features from the first layer. In that case, the effect should have been eliminated by normalising the features, but in most

Table 6.6.4: Classification results for MRI Jacobian data *with normalised features*. Means are given with standard errors in parentheses.

Method	AD vs CN				pMCI vs sMCI			
	Acc [%]	Sens [%]	Spec [%]	AUC	Acc [%]	Sens [%]	Spec [%]	AUC
Voxels	71.53 (0.40)	47.04 (0.94)	86.69 (0.61)	0.7369 (0.0052)	61.88 (0.20)	38.29 (0.50)	**84.31 (0.42)**	0.6642 (0.0021)
D1_J1_L3	**79.59 (0.43)**	63.62 (0.89)	**89.48 (0.57)**	**0.8596 (0.0035)**	**65.50 (0.19)**	**48.20 (0.56)**	81.96 (0.28)	0.7157 (0.0008)
D1_J1_L6	77.65 (0.44)	60.73 (0.99)	88.12 (0.63)	0.8362 (0.0035)	64.31 (0.23)	45.14 (0.74)	82.55 (0.37)	0.7148 (0.0010)
D1_J1_L10	78.01 (0.44)	**63.85 (1.06)**	86.79 (0.57)	0.8391 (0.0038)	63.77 (0.23)	43.94 (0.72)	82.64 (0.35)	0.7131 (0.0011)
D1_J2_L3	79.15 (0.43)	63.19 (0.92)	89.02 (0.57)	0.8566 (0.0034)	63.58 (0.19)	44.24 (0.56)	81.97 (0.28)	**0.7216 (0.0010)**
D1_J2_L6	77.19 (0.42)	62.19 (0.94)	86.48 (0.57)	0.8286 (0.0036)	62.09 (0.17)	42.21 (0.52)	81.00 (0.27)	0.6990 (0.0012)
D1_J2_L10	76.57 (0.36)	60.54 (0.97)	86.50 (0.55)	0.8249 (0.0038)	62.44 (0.19)	42.57 (0.58)	81.34 (0.30)	0.7025 (0.0013)
D2_J1_L3	77.34 (0.49)	61.00 (1.12)	87.45 (0.62)	0.8324 (0.0045)	63.54 (0.24)	45.85 (0.86)	80.37 (0.45)	0.7006 (0.0012)
D2_J1_L6	74.74 (0.45)	60.42 (1.14)	83.60 (0.72)	0.8085 (0.0041)	61.49 (0.23)	41.02 (0.89)	80.96 (0.51)	0.6779 (0.0013)
D2_J1_L10	73.62 (0.45)	60.96 (1.12)	81.45 (0.82)	0.8039 (0.0040)	62.42 (0.24)	44.13 (0.94)	79.81 (0.55)	0.6864 (0.0013)
D2_J2_L3	76.72 (0.42)	60.65 (0.95)	86.67 (0.60)	0.8325 (0.0039)	62.31 (0.16)	42.63 (0.61)	81.02 (0.44)	0.6841 (0.0011)
D2_J2_L6	75.46 (0.43)	59.00 (0.98)	85.64 (0.68)	0.8143 (0.0039)	60.30 (0.17)	39.87 (0.70)	79.74 (0.49)	0.6727 (0.0013)
D2_J2_L10	74.37 (0.39)	60.19 (1.05)	83.14 (0.60)	0.8002 (0.0039)	61.15 (0.19)	41.99 (0.70)	79.37 (0.41)	0.6741 (0.0013)

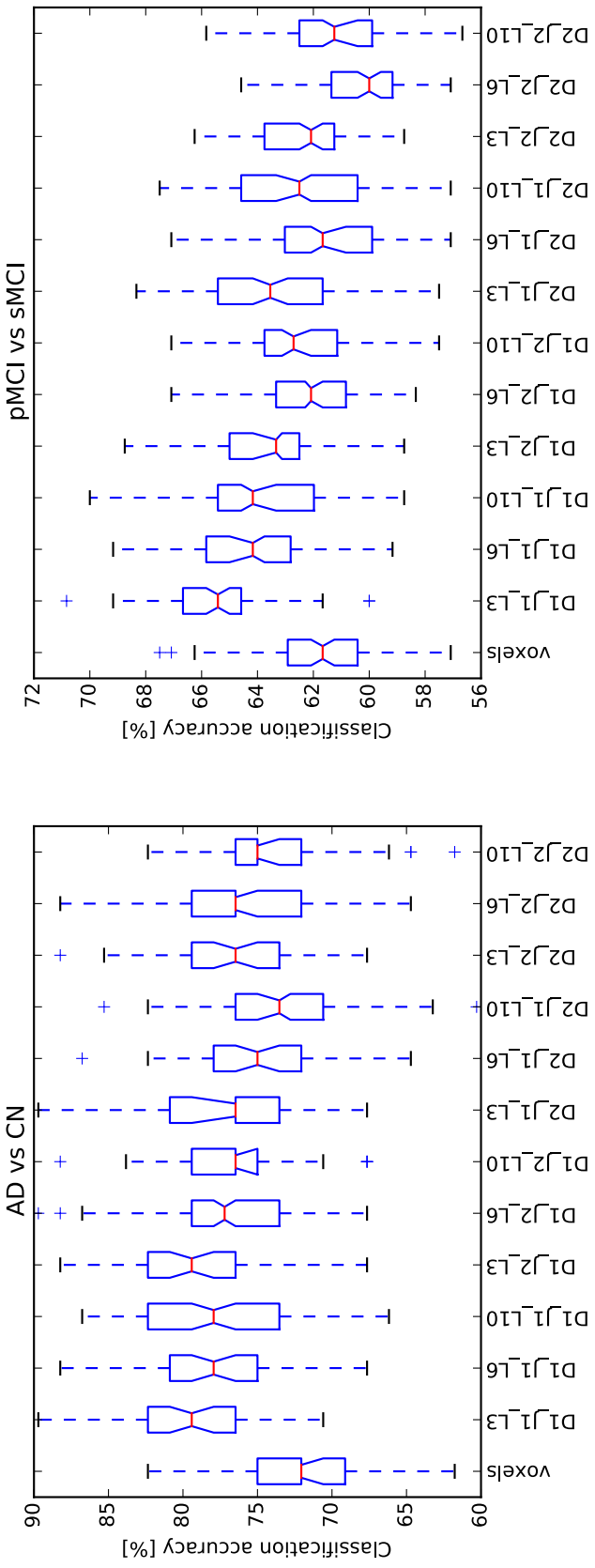


Figure 6.6.4: Classification accuracies for MRI Jacobian data *with normalised features.*

cases there was no benefit from applying such normalisation, as the experiments have shown. Insensitivity to J and L is more difficult to comment on, since it is difficult to say how these values should be chosen. However, it is an advantage for an algorithm if it is not particularly sensitive to the parameters. It is also important to note that these results are from just two closely related data sets. It may be the case that, for different types of data, more layers could be beneficial or specific values of J and L could be substantially better than others.

Table 6.7.1 shows results from selected ADNI classification studies found in the literature. The imaging modalities and types of features used in each study are listed in addition to performance metrics for AD/CN classification and MCI prediction.

The affinely aligned MR T1-weighted images used in this chapter were co-registered by Guerrero *et al.* for their work in [**114**], so it is natural to compare our results to theirs. They achieved 89% AD vs CN classification accuracy and 73% MCI prediction accuracy, compared to 68.45% for AD vs CN and 57.68% for MCI in this chapter (averaged results for all scattering networks). Their work combined affinely aligned images with non-rigidly aligned images, using manifold learning for dimesionality reduction. They also combined data from different scanner field strengths (1.5T and 3T) and from different phases of the ADNI project: ADNI-1, ADNI-GO and ADNI-2. In contrast, this chapter focuses on affinely aligned images and only includes 1.5T images from ADNI-1. Furthermore, the evaluation methodology is different in this chapter, using a set of random partitions into training and test data, which is in contrast to the single partition into a training subset and a test subset used by Guerrero *et al.* [**114**]. It would be an interesting topic for future work to apply the manifold learning methods studied in [**114**] to the scattering features studied in this chapter, to see if the improvements due to replacing voxel-wise features with scattering features would also be observed with their dimensionality reduction and classification algorithms.

Study	Modalities	Type of features for MRI	AD/CN				sMCI/pMCI			
			acc [%]	sens [%]	spec [%]	AUC	acc [%]	sens [%]	spec [%]	AUC
Hinrichs *et al.* [**130**]	MRI, PET, APOE, CSF	VBM	92.4	86.7	96.6	0.977	-	-	-	0.7667
Young *et al.* [**272**]	MRI, PET, APOE	VBM	-	-	-	-	69.9	78.7	65.6	0.795
Liu *et al.* [**174**]	MRI	VBM	92.0	90.9	93.0		0.723	-	-	
Suk *et al.* [**239**]	MRI, PET	VBM patches, deep learning	95.35	94.65	95.22	0.9877	75.92	48.04	95.23	0.7466
Koikkalainen *et al.* [**148**]	MRI	TBM	86.0	81	91	-	72.1	77	71	-
Vounou *et al.* [**259**]	MRI	Longitudinal TBM	90.3	87.5	92.1	-	82.1	81.5	82.9	-
Janousova *et al.* [**138**]	MRI	Longitudinal TBM	90.3	87.5	92.1	-	82.1	81.5	82.9	-
Cho *et al.* [**49**]	MRI	Cortical thickness	-	82	93	-	-	63	76	-
Eskildsen *et al.* [**90**]	MRI	Cortical thickness	84.5	79.4	88.9	0.905	69.9	55.2	73.1	0.635
Yun *et al.* [**275**]	MRI, PET	Cortical thickness, ROI	93.6	90.1	96.5	0.951	-	-	-	-
Westman *et al.* [**264**]	MRI, CSF	Cortical thickness, subcortical volumes	91.8	88.5	94.6	0.958	66.1	68.0	64.4	0.578
Zhang *et al.* [**278**]	MRI, PET	Multi-ROI	93.3	-	-	-	73.9	68.6	60.2	0.797
Zhang *et al.* [**279**]	MRI, PET	Multi-ROI	-	-	-	-	78.4	79.0	78.0	0.745
Zhang *et al.* [**280**]	MRI, PET, CSF	Multi-ROI	93.2	93	93.3	0.976	-	-	-	-
Liu *et al.* [**171**]	MRI, PET	Multi-ROI	94.37	94.71	94.04	0.9724	67.83	64.88	70.00	0.6957
Coupe *et al.* [**59**]	MRI	Patch-based	89	84	93	-	71	70	72	-
Tong *et al.* [**249**]	MRI	Patch-based	88.8	85.9	91.3	-	69.6	67.1	71.4	-
Wolz *et al.* [**268**]	MRI	Cortical thickness, TBM, ROI, manifold learning	89	93	85	-	68	67	69	-
Guerrero *et al.* [**114**]	MRI	Manifold learning	89	87	91	-	73	73	73	-

Table 6.7.1: Classification results in other studies of the ADNI data set.

The Jacobian data used in this chapter originates from the project of Vounou *et al.* [**259**] and Janousova *et al.* [**138**]. They achieved 90.3% accuracy for AD/CN classification and 82.1% accuracy for MCI prediction, compared to 77.43% and 61.95% respectively in this work. The main differences between their work and the method presented in this chapter are that first, they applied feature selection to voxel-wise data (no feature selection was used in this work) and second, they used a Gaussian kernel SVM classifier trained on selected features (in this work logistic regression was used for classification). It would be an interesting direction for future work to apply their feature selection and classification algorithm to scattering representations of longitudinal Jacobian images, potentially combining the benefits of both methods.

In comparison to other studies listed in table 6.7.1, the classification accuracies achieved in this work are lower. This seems to indicate that the simple approach of scattering feature transformation followed directly by classification is not sufficient. Perhaps combining scattering representations with feature selection, manifold learning or other methods proposed in the papers listed in table 6.7.1 would improve the results. In particular, it would be an interesting direction for future work to compute scattering representations of VBM maps and use them for classification. Additionally, it should be noted that the accuracy of the diagnostic consensus criteria is about 90% [**109**], so methods that exceed this accuracy may be suffering from the problem of overfitting the (imperfect) diagnosis made by a human.

Bruna and Mallat proposed a method for displaying scattering representations by partitioning a disk into regions, with each region coloured to represent one scattering network output [**27**]. Scattering networks are convolutional and produce a full set of outputs for each voxel so a grid of disks is required, although in practice the outputs are down-sampled and disks are displayed on a coarser grid. Extending this visualisation technique to three-dimensional tomographic images would require a 3D grid of balls, each divided into sectors, which would be difficult to read. Alternatively, it may be possible to adapt the techniques developed for visualisation of convolutional

neural networks [**162, 89, 158, 233, 276**] for scattering networks, but this is not a straightforward task. Therefore, developing methods for visualisation of classifiers built with 3D scattering networks remains a topic for future work.

6.8. Conclusions

An extension of scattering networks for analysis of three-dimensional tomographic images is proposed in this chapter. Since the feature vectors output by these networks are of very high dimensionality, a compressive classification method is also proposed which enables learning classifiers by computing a random projection of the data and training a linear classifier for the low-dimensional representations.

The proposed 3D scattering networks combined with compressive classification were evaluated on a medical image classification task. A classifier was trained to distinguish pre-processed brain images of Alzheimer's disease patients from those of cognitive normal (CN) controls. This classifier was then tested on held-out images of AD and CN subjects, and in addition it was evaluated as a method to predict whether mild cognitive impairment (MCI) patients would progress to AD. In both tasks scattering features improved classification accuracy compared to voxel-wise features, although these results are still short of state of the art reported in the literature. It may be possible to further improve the results obtained in this chapter by combining the proposed feature respresentations with more advanced machine learning methods, such as manifold learning or deep neural networks [**153, 234, 242, 125**]. Combining scattering networks with deep convolutional networks was already suggested by Bruna and Mallat [**27**].

While the proposed formulation of 3D scattering networks is an intuitive extension of the two-dimensional version developed by Bruna and Mallat [**27**], it still requires thorough theoretical validation. Similarly, the fast Johnson-Lindenstrauss transform based on the discrete cosine transform also requires mathematical proof. In addition, visualisation techniques for 3D scattering networks still have to be developed.

The proposed compressive classification method is also fairly simple, so there may be scope for improvement in this part, especially by considering elements of other compressive classification algorithms proposed in the literature, as discussed in section 6.4.

CHAPTER 7

Outlook

The contributions presented in this dissertation can be extended further or combined with methods proposed by other authors, forming part of the basis of future work. This chapter collects the ideas for extensions that were already discussed in the conclusions of individual chapters.

The method proposed in chapter 4 for learning an optimised basis for compressed sensing can be further developed to take into account the coherence between the learned basis and the sensing basis (which is the Fourier basis in case of MRI). This would address the weakness which was hypothesised as a possible reason for only small improvement over standard wavelets in the MRI reconstruction task, and possibly improve the performance of the proposed method.

The evaluation of sparse methods for AD classification based on FDG-PET images that was presented in chapter 5 can be improved by adding more algorithms, particularly some of the more sophisticated methods that were proposed by other authors. The feature clustering methods discussed in chapter 5 may be possible to improve by making these algorithms supervised, *i.e.* by including information about the labels of individual images. This would make it possible to cluster features not only based on their pairwise correlations, but also based on their correlations with the labels. There are several examples of supervised feature clustering for medical imaging in recent literature [**95, 192**].

The extension of scattering networks to three-dimensional images that was proposed in chapter 6 still requires thorough mathematical validation. Similarly, the fast Johnson-Lindenstrauss transform based on the discrete cosine transform also

requires a mathematical proof. The problem of displaying 3D scattering representations in a legible way also remains unsolved, with a possible source of inspiration in methods currently used to visualise neural networks [**162, 89, 158, 233, 276**]. The proposed application to AD classification based on MR images could also be improved with elements of the methods that are current state of the art in this field. The high-dimensional features produced by scattering networks can be analysed with manifold learning methods. Alternatively, they can be fed as inputs to complex machine learning models, such as deep convolutional networks which have been used to achieve remarkable progress in computer vision in recent years [**153, 234, 242, 125**]. Indeed, this is one of the possible directions suggested for scattering networks by Bruna and Mallat [**27**].

Finally, the question of whether compressed sensing MRI can be combined with scattering representations to improve detection of AD remains open. One could start the work in this direction by investigating if the types of artifacts specific to CS-MRI images have any adverse effects on classification with scattering representations. If more efficient MRI scans could be combined with computer-aided diagnosis, this would allow for more extensive screening for AD. A long-term objective would be to optimise image aquisition and reconstruction specifically for classification. Research is this direction would contribute to the field of application-driven medical imaging [**32, 107, 133, 70**].

CHAPTER 8

Summary and conclusion

This chapter summarises the main points of all the previous chapters of this dissertation and presents an overall conclusion.

8.1. Summary

Chapter 1 presented the general context of this dissertation by briefly introducing the physical principles of two relevant medical imaging modalities: magnetic resonance imaging (MRI) and positron emission tomography (PET). It also discussed Alzheimer's disease (AD) on a basic level and explained the role of MRI and FDG-PET in AD research. Finally, it discussed the contributions made by this dissertation and presented an outline of the rest of it.

Chapter 2 discussed a number of mathematical methods related to wavelets and sparsity, providing the mathematical foundations for the rest of this dissertation. It started with a definition of continuous wavelets, then discussed discrete wavelets and showed with an example how wavelet decompositions can be used to represent images in a sparse way. This was followed by a general discussion of sparse representations, with algorithms for finding sparse encodings in a given dictionary as well as for learning dictionaries adapted for sparsely representing a given data set. The mathematical foundations of compressed sensing (CS) were also discussed. Finally, this chapter also mentioned some applications of these methods to medical imaging problems.

Chapter 3 introduced the basics of machine learning and in particular the machine learning algorithms that are essential to subsequent chapters. A distinction was

made between supervised and unsupervised learning, with classification a particular example of supervised learning. Several classification algorithms were discussed next, including logistic regression, support vector machines (SVM) and random forests. The problem of overfitting was also highlighted and it was discussed how it can be addressed by regularisation (as is the case with regularised logistic regression) or averaging over a set of randomised models (as is the case with random forests). This was followed by a discussion of dimensionality reduction methods, including feature selection, feature agglomeration, principal component analysis (PCA) and manifold learning. Randomised projections were also mentioned. Finally, chapter 3 discussed common ways to measure the performance of classification algorithms, with definitions of standard performance indicators and a brief discussion of cross-validation.

The main contributions of this dissertation were presented in chapters 4, 5 and 6.

Chapter 4 presented a method for learning a wavelet packet basis for a set of images with an optimisation criterion selected specifically for compressed sensing. This method is an adaptation of the algorithm proposed by Coifman and Wickerhauser [**56**]. The presented method was tested in two tasks: sparse approximation of brain MR images and brain MR image reconstruction from compresssed sensing measurements. It was shown that the basis learned with the proposed method can approximate images more sparsely than a standard wavelet tree. In the compressed sensing reconstruction task the improvement over the standard wavelet representation was rather small, which was hypothesised to be due to increased coherence between the adapted wavelet packet basis and the Fourier basis, compared to the coherence between the standard wavelet basis and the Fourier basis.

Chapter 5 presented an evaluation of potential benefits of applying dimensionality reduction to Alzheimer's disease detection with machine learning based on FDG-PET images. A number of feature selection and feature clustering algorithms were each evaluated in combination with common classification algorithms. It was found

that there was no substantial benefit in terms of classification accuracy to applying the dimensionality reduction methods included in the comparisons. This finding is at odds with reports in the literature which proposed and successfully validated sparse methods for AD classification. It was hypothesised that this could be due to the relative simplicity of the methods studied in chapter 5 compared to the sophisticated methods proposed by other authors.

Chapter 6 presented a proposed extension of scattering networks to three-dimensional tomographic images. Since scattering networks output very high-dimensional feature vectors, scattering representations were combined with the fast Johnson-Lindenstrauss transform to enable linear classsification in a reduced-dimension space. This combined method was tested for AD detection based on structural MR images and tensor-based morphometry maps. It was observed that scattering representations improved over voxel-wise representation in terms of classification accuracy. However, the achieved accuracies were still lower than those reported in the literature for some other methods, which suggests that combining scattering networks with elements of methods proposed by other authors could improve the results further.

Chapter 7 discussed the potential future directions for the work presented in this dissertation, summarising the extensions proposed in individual chapters. It also highlighted the point that an interesting topic for future work would be to investigate how image acquisition and reconstruction with compressed sensing affects the performance of classification algorithms applied subsequently to images reconstructed in this way.

8.2. Conclusion

This dissertation contributes to the field of medical image computing, and in particular to the literature on compressed sensing MRI reconstruction and detection of AD based on tomographic brain images. All of the proposed methods were designed

for brain imaging and Alzheimer's disease imaging in particular, but applications beyond this domain may also be possible.

The proposed methods rely on algorithms and mathematical theory from fields ranging from wavelet-based signal processing, through sparse methods, to machine learning. In particular, chapter 7 combines wavelet-based scattering feature representations with efficient randomised projections (fast Johnoson-Lindenstrauss transform) and linear classification in a novel way, giving an algorithm that can learn efficiently from a very large number of features. This algorithm and its variations may be applicable to other problems in medical imaging and image processing in general.

Appendix: Lagrangian multipliers and Lagrange dual function

The following definition is a summary of the first part of Chapter 5 from [**21**].

Definition. Lagrange multipliers and Lagrange dual function

Consider the following optimisation problem:

$$\min f_0(\mathbf{x}) \quad \text{subject to} \begin{cases} f_i(\mathbf{x}) \leq 0, & i = 1, \ldots, m \\ h_i(\mathbf{x}) = 0, & i = 1, \ldots, p \end{cases} \tag{8.2.1}$$

where $\mathbf{x} \in \mathbb{R}^n$. The domain $\mathcal{D} = \bigcap_{i=1}^{m} \operatorname{dom} f_i \cap \bigcap_{i=1}^{p} \operatorname{dom} h_i$ is assumed to be non-empty and the optimal value is denoted $f^\star$.

The Lagrangian associated with problem 8.2.1 is the function $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$,

$$L(\mathbf{x}, \lambda, \nu) = f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^{p} \nu_i h_i(\mathbf{x})$$

with domain $\operatorname{dom} L = \mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^p$. The variables λ_i and ν_i are the Lagrange multipliers associated with their respective inequality (λ_i) or equality (ν_i) constraints. The vectors λ and ν are referred to as dual variables [**21**].

The Lagrange dual function $g : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ is the minimum of the Lagrangian over $\mathbf{x}$, *i.e.*

$$g(\lambda, \nu) = \inf_{\mathbf{x} \in \mathcal{D}} \left(f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^{p} \nu_i h_i(\mathbf{x}) \right)$$

and it gives a lower bound on the optimal value $f^\star$[**21**].

Bibliography

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(11):2274–2282, 2012.

[2] Dimitris Achlioptas. Database-friendly random projections. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, PODS '01, pages 274–281, New York, NY, USA, 2001. ACM.

[3] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311 –4322, 2006.

[4] N. Ailon and B. Chazelle. The fast Johnson-Lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on Computing*, 39(1):302–322, 2009.

[5] Nir Ailon and Edo Liberty. An almost optimal unrestricted fast Johnson-Lindenstrauss transform. *ACM Trans. Algorithms*, 9(3):21:1–21:12, June 2013.

[6] Daniel Aloise, Amit Deshpande, Pierre Hansen, and Preyas Popat. NP-hardness of Euclidean sum-of-squares clustering. *Machine Learning*, 75(2):245–248, 2009.

[7] Joakim Andén and Stéphane Mallat. Multiscale Scattering for Audio Classification. In *12th International Society for Music Information Retrieval Conference (ISMIR 2011)*, pages 657–662, 2011.

[8] Joakim Andén and Stéphane Mallat. Deep scattering spectrum. *Signal Processing, IEEE Transactions on*, 62(16):4114–4128, 2013.

[9] John Ashburner and Karl J. Friston. Voxel-based morphometry – the methods. *NeuroImage*, 11(6):805–821, 6 2000.

[10] Alzheimer's Association. 2014 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 10(2):e47–e92, 3 2014.

[11] S.P. Awate and E.V.R. DiBella. Spatiotemporal dictionary learning for undersampled dynamic MRI reconstruction via joint frame-based and dictionary-based sparsity. In *Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on*, pages 318–321, May 2012.

[12] S.D. Babacan, Xi Peng, Xian-Pei Wang, M.N. Do, and Zhi-Pei Liang. Reference-guided sparsifying transform design for compressive sensing MRI. In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, pages 5718–5721, Aug 2011.

[13] Akram Bakkour, John C. Morris, and Bradford C. Dickerson. The cortical signature of prodromal AD: Regional thinning predicts mild AD dementia. *Neurology*, 72(12):1048–1055, 2009.

[14] Roland Bammer. Basic principles of diffusion-weighted imaging. *European Journal of Radiology*, 45(3):169–184, 3 2003.

[15] Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28:253–263, 2008. 10.1007/s00365-007-9003-x.

[16] Peter J. Basser, Sinisa Pajevic, Carlo Pierpaoli, Jeffrey Duda, and Akram Aldroubi. In vivo fiber tractography using DT-MRI data. *Magnetic Resonance in Medicine*, 44(4):625–632, 2000.

[17] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*, volume 14, pages 585–591, 2001.

[18] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[19] A. Bilgin, Y. Kim, F. Liu, and M. S. Nadar. Dictionary design for compressed sensing MRI. In *Proceedings of the International Society for Magnetic Resonance in Medicine (ISMRM 2010)*, page 4887, 2010.

[20] Veronica Bolon-Canedo, Noelia Sanchez-Marono, and Amparo Alonso-Betanzos. A review of feature selection methods on synthetic data. *Knowledge and Information Systems*, 34(3):483–519, 2013.

[21] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[22] Richard G. Boyes, Daniel Rueckert, Paul Aljabar, Jennifer Whitwell, Jonathan M. Schott, Derek L.G. Hill, and Nicholas C. Fox. Cerebral atrophy measurements using Jacobian integration: Comparison with the boundary shift integral. *NeuroImage*, 32(1):159 – 169, 2006.

[23] Michael J Brammer. Multidimensional wavelet analysis of functional magnetic resonance images. *Human Brain Mapping*, 6(5-6):378–382, 1998.

[24] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[25] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

[26] J. Bruna and S. Mallat. Classification with scattering operators. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1561–1566, June 2011.

[27] J. Bruna and S. Mallat. Invariant scattering convolution networks. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):1872–1886, Aug 2013.

[28] Ed Bullmore, Jalal Fadili, Voichita Maxim, Levent Şendur, Brandon Whitcher, John Suckling, Michael Brammer, , and Michael Breakspear. Wavelets and functional magnetic resonance imaging of the human brain. *NeuroImage*, 23, Supplement 1:S234 – S249, 2004. Mathematics in Brain Imaging.

[29] B. Burdge, K. Kreutz-Delgado, and J. Murray. A unified FOCUSS framework for learning sparse dictionaries and non-squared error. In *Signals, Systems and Computers (ASILOMAR), 2010 Conference Record of the Forty Fourth Asilomar Conference on*, pages 2037 –2041, nov. 2010.

[30] C. S. Burrus. *Wavelets and Wavelet Transforms*. OpenStax CNX, February 2013.

[31] J. Caballero, A.N. Price, D. Rueckert, and J.V. Hajnal. Dictionary learning and time sparsity for dynamic MR data reconstruction. *Medical Imaging, IEEE Transactions on*, 33(4):979–994, Apr. 2014.

[32] Jose Caballero, Wenjia Bai, Anthony N. Price, Daniel Rueckert, and Joseph V. Hajnal. *Application-Driven MRI: Joint Reconstruction and Segmentation from Undersampled MRI Data*, pages 106–113. Springer International Publishing, Cham, 2014.

[33] Erick Jorge Canales-Rodriguez, Joaquim Radua, Edith Pomarol-Clotet, Salvador Sarro, Yasser Aleman-Gomez, Yasser Iturria-Medina, and Raymond Salvador. Statistical analysis of brain tissue images in the wavelet domain: Wavelet-based morphometry. *NeuroImage*, 72(0):214 – 226, 2013.

[34] E.J. Candes. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346:589 – 592, 2008.

[35] E.J. Candes, Y.C. Eldar, D. Needell, and P. Randall. Compressed sensing with coherent and redundant dictionaries. *Applied and Computational Harmonic Analysis*, 31(1):59–73, 2011.

[36] E.J. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489 – 509, 2006.

[37] E.J. Candes and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203 – 4215, 2005.

[38] E.J. Candes and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *Information Theory, IEEE Transactions on*, 52(12):5406–5425, Dec 2006.

[39] E.J. Candes and M.B. Wakin. An introduction to compressive sampling. *Signal Processing Magazine, IEEE*, 25(2):21 –30, 2008.

[40] Emmanuel Candes, Laurent Demanet, David Donoho, and Lexing Ying. Fast discrete curvelet transforms. *Multiscale Modeling & Simulation*, 5(3):861–899, 2006.

[41] Emmanuel Candes and Justin Romberg. Sparsity and incoherence in compressive sampling. *Inverse Problems*, 23(3):969, 2007.

[42] Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006.

[43] Timothy I Cannings and Richard J Samworth. Random projection ensemble classification. *arXiv preprint arXiv:1504.04595*, 2015.

[44] Sandeep Chaplot, L.M. Patnaik, and N.R. Jagannathan. Classification of magnetic resonance brain images using wavelets as input to support vector machine and neural network. *Biomedical Signal Processing and Control*, 1(1):86 – 92, 2006.

[45] S.S. Chen, D.L. Donoho, and M.A. Saunders. Atomic decomposition by basis pursuit. *SIAM journal on scientific computing*, 20(1):33–61, 1999.

[46] Xu Chen, Xiuyuan Cheng, and Stephane Mallat. Unsupervised deep Haar scattering on graphs. In *Advances in Neural Information Processing Systems*, pages 1709–1717, 2014.

[47] Bo Cheng, Mingxia Liu, Daoqiang Zhang, B.C. Munsell, and Dinggang Shen. Domain transfer learning for MCI conversion prediction. *Biomedical Engineering, IEEE Transactions on*, 62(7):1805–1817, July 2015.

[48] Bo Cheng, Daoqiang Zhang, and Dinggang Shen. Domain transfer learning for MCI conversion prediction. In Nicholas Ayache, Hervé Delingette, Polina Golland, and Kensaku Mori, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012*, volume 7510 of *Lecture Notes in Computer Science*, pages 82–90. Springer Berlin Heidelberg, 2012.

[49] Youngsang Cho, Joon-Kyung Seong, Yong Jeong, and Sung Yong Shin. Individual subject classification for Alzheimer's disease based on incremental learning using a spatial frequency representation of cortical thickness data. *NeuroImage*, 59(3):2217–2230, 2 2012.

[50] Carlton Chu, Ai-Ling Hsu, Kun-Hsien Chou, Peter Bandettini, and ChingPo Lin. Does feature selection improve classification accuracy? impact of sample size and feature selection

on classification using anatomical magnetic resonance images. *NeuroImage*, 60(1):59 – 70, 2012.

[51] V. Chudacek, R. Talmon, J. Anden, S. Mallat, R.R. Coifman, P. Abry, and M. Doret. Low dimensional manifold embedding for scattering coefficients of intrapartum fetale heart rate variability. In *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, pages 6373–6376, Aug 2014.

[52] M. K. Chung, K. J. Worsley, T. Paus, C. Cherif, D. L. Collins, J. N. Giedd, J. L. Rapoport, and A. C. Evans. A unified statistical approach to deformation-based morphometry. *NeuroImage*, 14(3):595–606, 9 2001.

[53] Marie Chupin, Emilie Gérardin, Rémi Cuingnet, Claire Boutet, Louis Lemieux, Stéphane Lehéricy, Habib Benali, Line Garnero, Olivier Colliot, and The Alzheimer's Disease Neuroimaging Initiative. Fully automatic hippocampus segmentation and classification in Alzheimer's disease and mild cognitive impairment applied on data from ADNI. *Hippocampus*, 19(6):579–587, 06 2009.

[54] Mark S. Cohen, Robert M. Weisskoff, Richard R. Rzedzian, and Howard L. Kantor. Sensory stimulation by time-varying magnetic fields. *Magnetic Resonance in Medicine*, 14(2):409–414, 1990.

[55] R.R. Coifman, Y. Meyer, S. Quake, and M.V. Wickerhauser. Signal processing and compression with wavelet packets. Technical report, Numerical Algorithms Research Group, Department of Mathematics, Yale University, New Haven, Connecticut 06520, 1990.

[56] R.R. Coifman and M.V. Wickerhauser. Entropy-based algorithms for best basis selection. *IEEE Transactions on Information Theory*, 38(2):713 –718, 1992.

[57] Olivier Colliot, Gael Chetelat, Marie Chupin, Beatrice Desgranges, Benoit Magnin, Habib Benali, Bruno Dubois, Line Garnero, Francis Eustache, and Stephane Lehericy. Discrimination between Alzheimer disease, mild cognitive impairment, and normal aging by using automated segmentation of the hippocampus. *Radiology*, 248(1):194–201, 2015/09/02 2008.

[58] George A. Constantinides. *Discrete Mathematics and Computational Complexity (Lecture Notes)*. Imperial College London, 2006.

[59] Pierrick Coupe, Simon F. Eskildsen, Jose V. Manjon, Vladimir S. Fonov, Jens C. Pruessner, Michele Allard, and D. Louis Collins. Scoring by nonlocal image patch estimator for early detection of Alzheimer's disease. *NeuroImage: Clinical*, 1(1):141–152, 2012.

[60] Remi Cuingnet, Emilie Gerardin, Jerome Tessieras, Guillaume Auzias, Stephane Lehericy, Marie-Odile Habert, Marie Chupin, Habib Benali, and Olivier Colliot. Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods

using the ADNI database. *NeuroImage*, 56(2):766 – 781, 2011. <ce:title>Multivariate Decoding and Brain Reading</ce:title>.

[61] Xiao Da, Jon B Toledo, Jarcy Zee, David A Wolk, Sharon X Xie, Yangming Ou, Amanda Shacklett, Paraskevi Parmpi, Leslie Shaw, John Q Trojanowski, Christos Davatzikos, and for the Alzheimer's Neuroimaging Initiative. Integration and relative value of biomarkers for prediction of MCI to AD progression: Spatial patterns of brain atrophy, cognitive scores, APOE genotype and CSF biomarkers. *NeuroImage : Clinical*, 4:164–173, 2014.

[62] Anders M. Dale, Bruce Fischl, and Martin I. Sereno. Cortical surface-based analysis: I. segmentation and surface reconstruction. *NeuroImage*, 9(2):179–194, 2 1999.

[63] S. Dasgupta and Y. Freund. Random projection trees for vector quantization. *Information Theory, IEEE Transactions on*, 55(7):3229–3242, July 2009.

[64] Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

[65] Manoranjan Dash and Huan Liu. Consistency-based search in feature selection. *Artificial Intelligence*, 151(1–2):155 – 176, 2003.

[66] Ingrid Daubechies. Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, 41(7):909–996, 1988.

[67] Christos Davatzikos, Priyanka Bhatt, Leslie M. Shaw, Kayhan N. Batmanghelich, and John Q. Trojanowski. Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. *Neurobiology of Aging*, 32(12):2322.e19 – 2322.e27, 2011.

[68] Christos Davatzikos, Ahmet Genc, Dongrong Xu, and Susan M. Resnick. Voxel-based morphometry using the RAVENS maps: Methods and validation using simulated longitudinal atrophy. *NeuroImage*, 14(6):1361–1369, 12 2001.

[69] Mark A. Davenport, Marco F. Duarte, Michael B. Wakin, Jason N. Laska, Dharmpal Takhar, Kevin F. Kelly, and Richard G. Baraniuk. The smashed filter for compressive classification and target recognition, 2007.

[70] D. Van de Sompel and M. Brady. Simultaneous reconstruction and segmentation algorithm for positron emission tomography and transmission tomography. In *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1035–1038, May 2008.

[71] M. Desco, J.A. Hernandez, A. Santos, and M. Brammer. Multiresolution analysis in fMRI: Sensitivity and specificity in the detection of brain activation. *Human Brain Mapping*, 14(1):16–27, 2001.

[72] Markus Deserno. How to generate equidistributed points on the surface of a sphere. Technical report, Max-Planck-Institut fur Polymerforschung, Ackermannweg 10, 55128 Mainz, Germany, 2004.

[73] Rahul S. Desikan, Howard J. Cabral, Christopher P. Hess, William P. Dillon, Christine M. Glastonbury, Michael W. Weiner, Nicholas J. Schmansky, Douglas N. Greve, David H. Salat, Randy L. Buckner, Bruce Fischl, and The Alzheimer's Disease Neuroimage Initiative. Automated MRI measures identify individuals with mild cognitive impairment and Alzheimer's disease. *Brain*, 05 2009.

[74] Ronald A. DeVore. Deterministic constructions of compressed sensing matrices. *Journal of Complexity*, 23(4–6):918 – 925, 2007. Festschrift for the 60th Birthday of Henryk Woźniakowski.

[75] Inderjit S. Dhillon, Subramanyam Mallela, and Rahul Kumar. A divisive information theoretic feature clustering algorithm for text classification. *J. Mach. Learn. Res.*, 3:1265–1287, March 2003.

[76] David L. Donoho. Wedgelets: nearly minimax estimation of edges. *Ann. Statist.*, 27(3):859–897, 06 1999.

[77] David L. Donoho. For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829, 2006.

[78] David L. Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via l1 minimization. *Proceedings of the National Academy of Sciences of the United States of America*, 100(5):pp. 2197–2202, 2003.

[79] David L. Donoho and Iain M. Johnstone. Ideal denoising in an orthonormal basis chosen from a library of bases. *Comptes Rendus Acad. Sci., Ser. I*, 319:1317–1322, 1994.

[80] David L. Donoho and Iain M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):pp. 1200–1224, 1995.

[81] David L. Donoho and Jain M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

[82] D.L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289 –1306, 2006.

[83] M.F. Duarte, M.A. Davenport, M.B. Wakin, J.N. Laska, Dharmpal Takhar, K.F. Kelly, and R.G. Baraniuk. Multiscale random projections for compressive classification. In *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, volume 6, pages VI – 161–VI – 164, Sept 2007.

[84] J.M. Duarte-Carvajalino and G. Sapiro. Learning to sense sparse signals: Simultaneous sensing matrix and sparsifying dictionary optimization. *IEEE Transactions on Image Processing*, 18(7):1395 –1408, 2009.

[85] H. Eavani, R. Filipovych, C. Davatzikos, T.D. Satterthwaite, R.E. Gur, and R.C. Gur. Sparse dictionary learning of resting state fMRI networks. In *Pattern Recognition in NeuroImaging (PRNI), 2012 International Workshop on*, pages 73–76, July 2012.

[86] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *Image Processing, IEEE Transactions on*, 15(12):3736 –3745, dec. 2006.

[87] M. Elad, M.A.T. Figueiredo, and Yi Ma. On the role of sparse and redundant representations in image processing. *Proceedings of the IEEE*, 98(6):972 –982, june 2010.

[88] Kjersti Engan, Karl Skretting, and John Håkon Husøy. Family of iterative LS-based dictionary learning algorithms, ILS-DLA, for sparse signal representation. *Digital Signal Processing*, 17(1):32 – 49, 2007.

[89] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *Dept. IRO, Université de Montréal, Tech. Rep*, 4323, 2009.

[90] Simon F. Eskildsen, Pierrick Coupé, Daniel García-Lorenzo, Vladimir Fonov, Jens C. Pruessner, and D. Louis Collins. Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning. *NeuroImage*, 65:511 – 521, 2013.

[91] Michael Ewers, Cathal Walsh, John Q. Trojanowski, Leslie M. Shaw, Ronald C. Petersen, Clifford R. Jack Jr., Howard H. Feldman, Arun L. W. Bokde, Gene E. Alexander, Philip Scheltens, Bruno Vellas, Bruno Dubois, Michael Weiner, and Harald Hampel. Prediction of conversion from mild cognitive impairment to Alzheimer's disease dementia based upon biomarkers and neuropsychological test performance. *Neurobiology of Aging*, 33(7):1203–1214.e2, 7 2012.

[92] M.J. Fadili and E.T. Bullmore. A comparative evaluation of wavelet-based methods for hypothesis testing of brain activation maps. *NeuroImage*, 23(3):1112 – 1128, 2004.

[93] Yong Fan, Nematollah Batmanghelich, Chris M. Clark, and Christos Davatzikos. Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. *NeuroImage*, 39(4):1731 – 1743, 2008.

[94] Yong Fan, Susan M. Resnick, Xiaoying Wu, and Christos Davatzikos. Structural and functional biomarkers of prodromal Alzheimer's disease: A high-dimensional pattern classification study. *NeuroImage*, 41(2):277 – 285, 2008.

[95] Yong Fan, Dinggang Shen, R.C. Gur, R.E. Gur, and C. Davatzikos. COMPARE: Classification of morphological patterns using adaptive regional elements. *Medical Imaging, IEEE Transactions on*, 26(1):93–105, Jan 2007.

[96] Xiaoli Zhang Fern and Carla E Brodley. Random projection for high dimensional data clustering: A cluster ensemble approach. In *ICML*, volume 3, pages 186–193, 2003.

[97] Roman Filipovych and Christos Davatzikos. Semi-supervised pattern classification of medical images: Application to mild cognitive impairment (MCI). *NeuroImage*, 55(3):1109 – 1119, 2011.

[98] Bruce Fischl and Anders M Dale. Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proceedings of the National Academy of Sciences of the United States of America*, 97(20):11050–11055, 09 2000.

[99] Bruce Fischl, Martin I. Sereno, and Anders M. Dale. Cortical surface-based analysis: II: Inflation, flattening, and a surface-based coordinate system. *NeuroImage*, 9(2):195–207, 2 1999.

[100] Christopher M Florkowski. Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: Communicating the performance of diagnostic tests. *The Clinical Biochemist Reviews*, 29(Suppl 1):S83–S87, 08 2008.

[101] P Frankl and H Maehara. The Johnson-Lindenstrauss lemma and the sphericity of some graphs. *Journal of Combinatorial Theory, Series B*, 44(3):355–362, 6 1988.

[102] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1–22, 2010.

[103] Karl J Friston, Andrew P Holmes, et al. Statistical parametric maps in functional imaging: a general linear approach. *Human Brain Mapping*, 2:189 – 210, 1995.

[104] U. Gamper, P. Boesiger, and S. Kozerke. Compressed sensing in dynamic MRI. *Magnetic Resonance in Medicine*, 59(2):365–373, 2008.

[105] C. Gargour, M. Gabrea, V. Ramachandran, and J.-M. Lina. A short introduction to wavelets and their applications. *Circuits and Systems Magazine, IEEE*, 9(2):57–68, Second 2009.

[106] Emilie Gerardin, Gael Chetelat, Marie Chupin, Remi Cuingnet, Beatrice Desgranges, Ho-Sung Kim, Marc Niethammer, Bruno Dubois, Stephane Lehericy, Line Garnero, Francis Eustache, and Olivier Colliot. Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging. *NeuroImage*, 47(4):1476–1486, 10 2009.

[107] P. Ghosh, D. H. Laidlaw, K. W. Fleischer, A. H. Barr, and R. E. Jacobs. Pure phase-encoded MRI and classification of solids. *IEEE Transactions on Medical Imaging*, 14(3):616–620, Sep 1995.

[108] Catriona D. Good, Ingrid S. Johnsrude, John Ashburner, Richard N. A. Henson, Karl J. Friston, and Richard S. J. Frackowiak. A voxel-based morphometric study of ageing in 465 normal adult human brains. *NeuroImage*, 14(1):21–36, 7 2001.

[109] Katherine Gray. *Machine learning for image-based classification of Alzheimer's disease*. PhD thesis, Imperial College London, November 2012. Available at http://www.doc.ic.ac.uk/ krg03/krgray_03102012_phdthesis.pdf.

[110] Katherine Gray, Paul Aljabar, R. A. Heckemann, A. Hammers, and Daniel Rueckert. Random Forest-Based Manifold Learning for Classification of Imaging Data in Dementia. In *MLMI Workshop, MICCAI 2011*, volume 7009 of *Lecture Notes in Computer Science*, pages 159–166, 2011. Winner of the best paper award for the workshop.

[111] Katherine Gray, Paul Aljabar, R. A. Heckemann, A. Hammers, Daniel Rueckert, and The Alzheimer's Disease Neuroimaging Initiative. Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. *NeuroImage*, 65:167–175, January 2013.

[112] Katherine Gray, Robin Wolz, Shiva Keihaninejad, R. A. Heckemann, Paul Aljabar, A. Hammers, and Daniel Rueckert. Regional analysis of FDG-PET for use in the classification of Alzheimer's disease. In *ISBI 2011*, pages 1082–1085, 2011.

[113] Andreas Grunauer and Markus Vincze. Using dimension reduction to improve the classification of high-dimensional data. *arXiv preprint arXiv:1505.06907*, 2015.

[114] Ricardo Guerrero, Christian Ledig, and Daniel Rueckert. Manifold alignment and transfer learning for classification of Alzheimer's disease. In Guorong Wu, Daoqiang Zhang, and Luping Zhou, editors, *Machine Learning in Medical Imaging*, volume 8679 of *Lecture Notes in Computer Science*, pages 77–84. Springer International Publishing, 2014.

[115] Kanghui Guo and Demetrio Labate. Optimally sparse multidimensional representation using shearlets. *SIAM journal on mathematical analysis*, 39(1):298–318, 2007.

[116] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, March 2003.

[117] Kerstin Hackmack, Friedemann Paul, Martin Weygandt, Carsten Allefeld, and John-Dylan Haynes. Multi-scale classification of disease using structural MRI and wavelet transform. *NeuroImage*, 62(1):48 – 58, 2012.

[118] C. Haense, K. Herholz, W. J. Jagust, and W. D. Heiss. Performance of FDG PET for detection of Alzheimer's disease in two independent multicentre samples (NEST-DD and ADNI). *Dementia and Geriatric Cognitive Disorders*, 28(3):259–266, 2009.

[119] J.V. Hajnal, D.L.G. Hill, and D.J. Hawkes. *Medical Image Registration*. CRC Press, 2001.

[120] Mark A Hall. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999.

[121] Mark A Hall and Lloyd A Smith. Practical feature subset selection for machine learning. 1998.

[122] C. L. G. Ham, J. M. L. Engels, G. T. van de Wiel, and A. Machielsen. Peripheral nerve stimulation during MRI: Effects of high gradient amplitudes and switching rates. *Journal of Magnetic Resonance Imaging*, 7(5):933–937, 1997.

[123] Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The Elements of Statistical Learning*. Springer, 2009.

[124] J. Haupt, R. Castro, R. Nowak, G. Fudge, and A. Yeh. Compressive sampling for signal classification. In *Signals, Systems and Computers, 2006. ACSSC '06. Fortieth Asilomar Conference on*, pages 1430–1434, Oct 2006.

[125] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[126] R. A. Heckemann, Christian Ledig, Paul Aljabar, Katherine Gray, Daniel Rueckert, JV Hajnal, and A. Hammers. Label propagation using group agreement - DISPATCH. In *MICCAI 2012 Grand Challenge and Workshop on Multi-Atlas Labeling*, September 2012. Available in proceedings at https://masi.vuse.vanderbilt.edu/workshop2012/images/c/c8/MICCAI_2012_Workshop_v2.pdf (pages 75-78).

[127] Juha Heinonen. *Lectures on Lipschitz analysis*. Department of Mathematics, University of Michigan, Ann Arbor, MI 48109, USA, 2005.

[128] K. Herholz, E. Salmon, D. Perani, J-C. Baron, V. Holthoff, L. Frölich, P. Schönknecht, K. Ito, R. Mielke, E. Kalbe, G. Zündorf, X. Delbeuck, O. Pelati, D. Anchisi, F. Fazio, N. Kerrouche, B. Desgranges, F. Eustache, B. Beuthien-Baumann, C. Menzel, J. Schröder, T. Kato, Y. Arahata, M. Henze, and W-D. Heiss. Discrimination between Alzheimer dementia and controls by automated analysis of multicenter FDG PET. *NeuroImage*, 17(1):302 – 316, 2002.

[129] Chris Hinrichs, Vikas Singh, Lopamudra Mukherjee, Guofan Xu, Moo K. Chung, and Sterling C. Johnson. Spatially augmented LPboosting for AD classification with evaluations on the ADNI dataset. *NeuroImage*, 48(1):138 – 149, 2009.

[130] Chris Hinrichs, Vikas Singh, Guofan Xu, and Sterling C. Johnson. Predictive markers for AD in a multi-modality framework: An analysis of MCI progression in the ADNI population. *NeuroImage*, 55(2):574 – 589, 2011.

[131] Yoko Hirata, Hiroshi Matsuda, Kiyotaka Nemoto, Takashi Ohnishi, Kentaro Hirao, Fumio Yamashita, Takashi Asada, Satoshi Iwabuchi, and Hirotsugu Samejima. Voxel-based morphometry to discriminate early Alzheimer's disease from controls. *Neuroscience Letters*, 382(3):269 – 274, 2005.

[132] Kieren Grant Hollingsworth. Reducing acquisition time in clinical MRI by data undersampling and compressed sensing reconstruction. *Physics in Medicine and Biology*, 60(21):R297, 2015.

[133] Ing-Tsung Hsiao, A. Rangarajan, and G. Gindi. Joint-MAP reconstruction/segmentation for transmission tomography using mixture-models as priors. In *Nuclear Science Symposium, 1998. Conference Record. 1998 IEEE*, volume 3, pages 1689–1693 vol.3, 1998.

[134] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, STOC '98, pages 604–613, New York, NY, USA, 1998. ACM.

[135] Clifford R. Jack, Matt A. Bernstein, Nick C. Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J. Britson, Jennifer L. Whitwell, Chadwick Ward, Anders M. Dale, Joel P. Felmlee, Jeffrey L. Gunter, Derek L.G. Hill, Ron Killiany, Norbert Schuff, Sabrina Fox-Bosetti, Chen Lin, Colin Studholme, Charles S. DeCarli, Gunnar Krueger, Heidi A. Ward, Gregory J. Metzger, Katherine T. Scott, Richard Mallozzi, Daniel Blezek, Joshua Levy, Josef P. Debbins, Adam S. Fleisher, Marilyn Albert, Robert Green, George Bartzokis, Gary Glover, John Mugler, and Michael W. Weiner. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, 27(4):685–691, 2008.

[136] C.R. Jack, R.C. Petersen, Y.C. Xu, P.C. O'Brien, G.E. Smith, R.J. Ivnik, B.F. Boeve, S.C. Waring, E.G. Tangalos, and E. Kokmen. Prediction of AD with MRI-based hippocampal volume in mild cognitive impairment. *Neurology*, 52(7):1397, 1999.

[137] William J. Jagust, Dan Bandy, Kewei Chen, Norman L. Foster, Susan M. Landau, Chester A. Mathis, Julie C. Price, Eric M. Reiman, Daniel Skovronsky, and Robert A. Koeppe. The

Alzheimer's Disease Neuroimaging Initiative positron emission tomography core. *Alzheimer's & Dementia*, 6(3):221–229, 5 2010.

[138] Eva Janousova, Maria Vounou, Robin Wolz, Katherine Gray, Daniel Rueckert, Giovanni Montana, and The Alzheimer's Disease Neuroimaging Initiative. Biomarker discovery for sparse classification of brain images in Alzheimer's disease. *Annals of the BMVA*, 2012(2):1–11, September 2012. Available at http://www.bmva.org/annals/2012/2012-0002.pdf.

[139] Biao Jie, Daoqiang Zhang, Bo Cheng, and Dinggang Shen. Manifold regularized multi-task feature selection for multi-modality classification in Alzheimer's disease. In Kensaku Mori, Ichiro Sakuma, Yoshinobu Sato, Christian Barillot, and Nassir Navab, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*, volume 8149 of *Lecture Notes in Computer Science*, pages 275–283. Springer Berlin Heidelberg, 2013.

[140] William B Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary mathematics*, 26(189-206):1, 1984.

[141] P. Jost, P. Vandergheynst, S. Lesage, and R. Gribonval. MoTIF: An efficient algorithm for learning translation invariant dictionaries. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 5, page V, May 2006.

[142] Joe H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

[143] Hong Jung, Kyunghyun Sung, Krishna S. Nayak, Eung Yeop Kim, and Jong Chul Ye. k-t FOCUSS: A general compressed sensing framework for high resolution dynamic MRI. *Magnetic Resonance in Medicine*, 61(1):103–116, 2009.

[144] B. Kainz, M. Steinberger, W. Wein, M. Kuklisova-Murgasova, C. Malamateniou, K. Keraudren, T. Torsney-Weir, M. Rutherford, P. Aljabar, J.V. Hajnal, and D. Rueckert. Fast volume reconstruction from motion corrupted stacks of 2D slices. *Medical Imaging, IEEE Transactions on*, 34(9):1901–1913, Sept 2015.

[145] S Sathiya Keerthi, KB Duan, Shirish Krishnaj Shevade, and Aun Neow Poo. A fast dual algorithm for kernel logistic regression. *Machine learning*, 61(1-3):151–165, 2005.

[146] S. Khullar, A. Michael, N. Correa, T. Adali, S. Baum, and V. Calhoun. Wavelet-based denoising and independent component analysis for improving multi-group inference in fMRI data. In *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*, pages 456 –459, 30 2011-april 2 2011.

[147] Nick Kingsbury. Complex wavelets for shift invariant analysis and filtering of signals. *Applied and Computational Harmonic Analysis*, 10(3):234 – 253, 2001.

[148] Juha Koikkalainen, Jyrki Lötjönen, Lennart Thurfjell, Daniel Rueckert, Gunhild Waldemar, and Hilkka Soininen. Multi-template tensor-based morphometry: Application to analysis of Alzheimer's disease. *NeuroImage*, 56(3):1134–1144, 6 2011.

[149] Daphne Koller and Mehran Sahami. Toward optimal feature selection. Technical Report 1996-77, Stanford InfoLab, February 1996. Previous number = SIDL-WP-1996-0032.

[150] Igor Kononenko. Estimating attributes: analysis and extensions of RELIEF. In *Machine Learning: ECML-94*, pages 171–182. Springer, 1994.

[151] Felix Krahmer and Rachel Ward. New and improved Johnson-Lindenstrauss embeddings via the restricted isometry property. *CoRR*, abs/1009.0744, 2010.

[152] Kenneth Kreutz-Delgado, Joseph F. Murray, Bhaskar D. Rao, Kjersti Engan, Te-Won Lee, and Terrence J. Sejnowski. Dictionary learning algorithms for sparse representation. *Neural Computation*, 15(2):349–396, 2003.

[153] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[154] Demetrio Labate, Wang-Q Lim, Gitta Kutyniok, and Guido Weiss. Sparse multidimensional representation using shearlets, 2005.

[155] Jessica B.S. Langbaum, Kewei Chen, Wendy Lee, Cole Reschke, Dan Bandy, Adam S. Fleisher, Gene E. Alexander, Norman L. Foster, Michael W. Weiner, Robert A. Koeppe, William J. Jagust, and Eric M. Reiman. Categorical and correlational analyses of baseline fluorodeoxyglucose positron emission tomography images from the Alzheimer's Disease Neuroimaging Initiative (ADNI). *NeuroImage*, 45(4):1107 – 1116, 2009.

[156] Zhiqiang Lao, Dinggang Shen, Zhong Xue, Bilge Karacali, Susan M. Resnick, and Christos Davatzikos. Morphological classification of brains via high-dimensional shape transformations and machine learning methods. *NeuroImage*, 21(1):46 – 57, 2004.

[157] Patrice Latinne, Olivier Debeir, and Christine Decaestecker. Limiting the number of trees in random forests. In Josef Kittler and Fabio Roli, editors, *Multiple Classifier Systems*, volume 2096 of *Lecture Notes in Computer Science*, pages 178–187. Springer Berlin Heidelberg, 2001.

[158] Quoc Le, Marc'Aurelio Ranzato, Rajat Monga, Matthieu Devin, Kai Chen, Greg Corrado, Jeff Dean, and Andrew Ng. Building high-level features using large scale unsupervised learning. In *International Conference in Machine Learning*, 2012.

[159] Quoc Le, Tamas Sarlos, and Alexander Smola. Fastfood - computing Hilbert space expansions in loglinear time. In *Proceedings of the 30th International Conference on Machine Learning*, pages 244–252, 2013.

[160] Denis Le Bihan, Jean-François Mangin, Cyril Poupon, Chris A. Clark, Sabina Pappata, Nicolas Molko, and Hughes Chabriat. Diffusion tensor imaging: Concepts and applications. *Journal of Magnetic Resonance Imaging*, 13(4):534–546, 2001.

[161] E. Le Pennec and S. Mallat. Sparse geometric image representations with bandelets. *Image Processing, IEEE Transactions on*, 14(4):423–438, April 2005.

[162] Honglak Lee, Chaitanya Ekanadham, and Andrew Y. Ng. Sparse deep belief net model for visual area V2. In J.C. Platt, D. Koller, Y. Singer, and S.T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 873–880. Curran Associates, Inc., 2008.

[163] K. Lee, S. Tak, and J. C. Ye. A data-driven sparse GLM for fMRI analysis using sparse dictionary learning with MDL criterion. *Medical Imaging, IEEE Transactions on*, 30(5):1076 –1089, may 2011.

[164] Jason P. Lerch, Jens Pruessner, Alex P. Zijdenbos, D. Louis Collins, Stefan J. Teipel, Harald Hampel, and Alan C. Evans. Automated cortical thickness measurements from MRI can accurately separate Alzheimer's patients from normal elderly controls. *Neurobiology of Aging*, 29(1):23–30, 1 2008.

[165] Shutao Li and Leyuan Fang. An efficient learned dictionary and its application to non-local denoising. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 1945 –1948, 2010.

[166] Dong Liang, Bo Liu, JiunJie Wang, and Leslie Ying. Accelerating SENSE using compressed sensing. *Magnetic Resonance in Medicine*, 62(6):1574–1584, 2009.

[167] S.G. Lingala and M. Jacob. Blind compressed sensing with sparse dictionaries for accelerated dynamic MRI. In *Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on*, pages 5–8, April 2013.

[168] S.G. Lingala and M. Jacob. Blind compressive sensing dynamic MRI. *Medical Imaging, IEEE Transactions on*, 32(6):1132–1145, June 2013.

[169] Bo Liu, Yi Ming Zou, and L. Ying. Sparsesense: Application of compressed sensing in parallel MRI. In *Information Technology and Applications in Biomedicine, 2008. ITAB 2008. International Conference on*, pages 127 –130, May 2008.

[170] Feng Liu, Heung-Il Suk, Chong-Yaw Wee, Huafu Chen, and Dinggang Shen. High-order graph matching based feature selection for Alzheimer's disease identification. In Kensaku Mori, Ichiro Sakuma, Yoshinobu Sato, Christian Barillot, and Nassir Navab, editors, *Medical*

Image Computing and Computer-Assisted Intervention – MICCAI 2013, volume 8150 of *Lecture Notes in Computer Science*, pages 311–318. Springer Berlin Heidelberg, 2013.

[171] Feng Liu, Chong-Yaw Wee, Huafu Chen, and Dinggang Shen. Inter-modality relationship constrained multi-modality multi-task feature selection for Alzheimer's disease and mild cognitive impairment identification. *NeuroImage*, 84(0):466 – 475, 2014.

[172] Li Liu and P.W. Fieguth. Texture classification from random features. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(3):574–586, March 2012.

[173] Manhua Liu, Daoqiang Zhang, and Dinggang Shen. Ensemble sparse classification of Alzheimer's disease. *NeuroImage*, 60(2):1106 – 1116, 2012.

[174] Manhua Liu, Daoqiang Zhang, Dinggang Shen, and The Alzheimer's Disease Neuroimaging Initiative. Hierarchical fusion of features and classifier decisions for Alzheimer's disease diagnosis. *Human brain mapping*, 35(4):1305–1319, 04 2014.

[175] Qiegen Liu, Shanshan Wang, Kun Yang, Jianhua Luo, Yuemin Zhu, and Dong Liang. Highly undersampled magnetic resonance image reconstruction using two-level Bregman method with dictionary updating. *Medical Imaging, IEEE Transactions on*, 32(7):1290–1301, July 2013.

[176] Sidong Liu, Yang Song, Weidong Cai, Sonia Pujol, Ron Kikinis, Xiaogang Wang, and Dagan Feng. Multifold Bayesian kernelization in Alzheimer's diagnosis. In Kensaku Mori, Ichiro Sakuma, Yoshinobu Sato, Christian Barillot, and Nassir Navab, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*, volume 8150 of *Lecture Notes in Computer Science*, pages 303–310. Springer Berlin Heidelberg, 2013.

[177] Yawu Liu, Teemu Paajanen, Yi Zhang, Eric Westman, Lars-Olof Wahlund, Andrew Simmons, Catherine Tunnard, Tomasz Sobow, Patrizia Mecocci, Magda Tsolaki, Bruno Vellas, Sebastian Muehlboeck, Alan Evans, Christian Spenger, Simon Lovestone, and Hilkka Soininen. Analysis of regional MRI volumes and thicknesses as predictors of conversion from mild cognitive impairment to Alzheimer's disease. *Neurobiology of Aging*, 31(8):1375 – 1385, 2010. Alzheimer's Disease Neuroimaging Initiative (ADNI) Studies.

[178] Nikos K. Logothetis. What we can do and what we cannot do with fMRI. *Nature*, 453(7197):869–878, 06 2008.

[179] Nikos K. Logothetis. What we can do and what we cannot do with fMRI - supplementary material. *Nature*, 453(7197):869–878, 06 2008.

[180] A. Lucchi, K. Smith, R. Achanta, G. Knott, and P. Fua. Supervoxel-based segmentation of mitochondria in EM image stacks with learned shape features. *Medical Imaging, IEEE Transactions on*, 31(2):474–486, Feb 2012.

[181] M. Lustig, D. Donoho, and J.M. Pauly. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007.

[182] M. Lustig, D.L. Donoho, J.M. Santos, and J.M. Pauly. Compressed sensing MRI. *Signal Processing Magazine, IEEE*, 25(2):72 –82, 2008.

[183] M. Lustig, J.M. Santos, D.L. Donoho, and J.M. Pauly. k-t SPARSE: High frame rate dynamic MRI exploiting spatio-temporal sparsity. In *Proceedings of the 13th Annual Meeting of ISMRM, Seattle*, page 2420, 2006.

[184] Michael Lustig, Juan M Santos, David L Donoho, and John M Pauly. kt SPARSE: High frame rate dynamic MRI exploiting spatio-temporal sparsity. In *Proceedings of the 13th Annual Meeting of ISMRM, Seattle*, volume 2420, 2006.

[185] J. Mairal, G. Sapiro, and M. Elad. Multiscale sparse image representation with learned dictionaries. In *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, volume 3, pages III –105 –III –108, 162007-oct.19 2007.

[186] Julien Mairal, Guillermo Sapiro, and Michael Elad. Learning multiscale sparse representations for image and video restoration. *Multiscale Modeling & Simulation*, 7(1):214–241, 2008.

[187] S. Mallat. *A Wavelet Tour of Signal Processing: The Sparse Way*. Elsevier/Academic Press, Amsterdam, 3rd edition, 2009.

[188] S.G. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 11(7):674–693, 1989.

[189] S.G. Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *Signal Processing, IEEE Transactions on*, 41(12):3397 –3415, dec 1993.

[190] Stephane Mallat. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10):1331–1398, 2012.

[191] Nicolai Meinshausen and Peter Buhlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

[192] Vincent Michel, Alexandre Gramfort, Gael Varoquaux, Evelyn Eger, Christine Keribin, and Bertrand Thirion. A supervised clustering approach for fMRI-based inference of brain states. *Pattern Recognition*, 45(6):2041 – 2049, 2012.

[193] Satoshi Minoshima, Kirk A Frey, Robert A Koeppe, Norman L Foster, and David E Kuhl. A diagnostic approach in Alzheimer's disease using three-dimensional stereotactic surface projections of fluorine-18-FDG PET. *J Nucl med*, 36:1238–1248, 1995.

[194] Chandan Misra, Yong Fan, and Christos Davatzikos. Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: Results from ADNI. *NeuroImage*, 44(4):1415 – 1422, 2009.

[195] Silvia Morbelli, Arnoldo Piccardo, Giampiero Villavecchia, Barbara Dessi, Andrea Brugnolo, Alessandra Piccini, Anna Caroli, Giovanni Frisoni, Guido Rodriguez, and Flavio Nobili. Mapping brain morphological and functional conversion patterns in amnestic MCI: a voxel-based MRI and FDG-PET study. *European Journal of Nuclear Medicine and Molecular Imaging*, 37(1):36–45, 2010.

[196] Karsten Mueller, Gabriele Lohmann, Stefan Zysset, and D. Yves von Cramon. Wavelet statistics of functional MRI data and the general linear model. *Journal of Magnetic Resonance Imaging*, 17(1):20–30, 2003.

[197] S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett. The Alzheimer's Disease Neuroimaging Initiative. *Neuroimaging Clinics of North America*, 15(4):869–877, 2005. Alzheimer's disease: 100 years of progress.

[198] Bruno A. Olshausen and David J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, June 1996.

[199] ThaisMayumi Oshiro, PedroSantoro Perez, and Jose Augusto Baranauskas. How many trees in a random forest? In Petra Perner, editor, *Machine Learning and Data Mining in Pattern Recognition*, volume 7376 of *Lecture Notes in Computer Science*, pages 154–168. Springer Berlin Heidelberg, 2012.

[200] Ricardo Otazo, Daniel Kim, Leon Axel, and Daniel K. Sodickson. Combination of compressed sensing and parallel imaging for highly accelerated first-pass cardiac perfusion MRI. *Magnetic Resonance in Medicine*, 64(3):767–776, 2010.

[201] S.G. Ozkaya and D. Van De Ville. Anatomically adapted wavelets for integrated statistical analysis of fMRI data. In *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*, pages 469 –472, 30 2011-april 2 2011.

[202] Y.C. Pati, R. Rezaiifar, and P.S. Krishnaprasad. Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. In *Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on*, pages 40 –44 vol.1, nov 1993.

[203] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[204] H. Peng, Fulmi Long, and C. Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1226–1238, Aug 2005.

[205] Richard J. Perrin, Anne M. Fagan, and David M. Holtzman. Multimodal techniques for diagnosis and prognosis of Alzheimer's disease. *Nature*, 461(7266):916–922, 10 2009.

[206] G. Peyre. Best basis compressed sensing. In *Proceedings of the 1st International Conference on Scale Space and Variational Methods in Computer Vision*, SSVM'07, pages 80–91, Berlin, Heidelberg, 2007. Springer-Verlag.

[207] G. Peyre. Best basis compressed sensing. *IEEE Transactions on Signal Processing*, 58(5):2613 –2622, 2010.

[208] Olivier Querbes, Florent Aubry, Jeremie Pariente, Jean-Albert Lotterie, Jean-Francois Demonet, Veronique Duret, Michele Puel, Isabelle Berry, Jean-Claude Fort, Pierre Celsis, and The Alzheimer's Disease Neuroimage Initiative. Early diagnosis of Alzheimer's disease using cortical thickness: impact of cognitive reserve. *Brain*, 132(8):2036–2047, 07 2009.

[209] Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1313–1320. Curran Associates, Inc., 2009.

[210] K. Ramchandran and M. Vetterli. Best wavelet packet bases in a rate-distortion sense. *IEEE Transactions on Image Processing*, 2(2):160 –175, 1993.

[211] S. Ravishankar and Y. Bresler. MR image reconstruction from highly undersampled k-space data by dictionary learning. *IEEE Transactions on Medical Imaging*, 30(5):1028 –1041, 2011.

[212] S Ravishankar and Y Bresler. Multiscale dictionary learning for MRI. In *Proc. ISMRM*, page 2830, 2011.

[213] S. Ravishankar and Y. Bresler. Learning sparsifying transforms. *Signal Processing, IEEE Transactions on*, 61(5):1072–1086, March 2013.

[214] S. Ravishankar and Y. Bresler. Sparsifying transform learning for compressed sensing MRI. In *Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on*, pages 17–20, April 2013.

[215] H. Reboredo, F. Renna, R. Calderbank, and M.R.D. Rodrigues. Compressive classification. In *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*, pages 674–678, July 2013.

[216] Eric M. Reiman and William J. Jagust. Brain imaging in the study of Alzheimer's disease. *NeuroImage*, 61(2):505–516, 6 2012.

[217] Michal P Romaniuk, Anil W Rao, Robin Wolz, Joseph V Hajnal, and Daniel Rueckert. Learning best wavelet packet bases for compressed sensing of classes of images: application to brain MR imaging. In *STMI 2012 – MICCAI Workshop on Sparsity Techniques in Medical Imaging*, 2012.

[218] Ruth Rosenholtz. Statistical methods in brain and cognitive science (spring 2004). In *MIT OpenCourseWare*. Massachusetts Institute of Technology, 2004.

[219] Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[220] R. Rubinstein, A.M. Bruckstein, and M. Elad. Dictionaries for sparse representation modeling. *Proceedings of the IEEE*, 98(6):1045 –1057, 2010.

[221] D. Rueckert, L.I. Sonoda, C. Hayes, D.L.G. Hill, M.O. Leach, and D.J. Hawkes. Nonrigid registration using free-form deformations: application to breast MR images. *Medical Imaging, IEEE Transactions on*, 18(8):712–721, Aug 1999.

[222] Urs E. Ruttimann, M. Unser, R.R. Rawlings, D. Rio, N.F. Ramsey, V.S. Mattay, D.W. Hommer, J.A. Frank, and D.R. Weinberger. Statistical analysis of functional MRI data in the wavelet domain. *Medical Imaging, IEEE Transactions on*, 17(2):142–154, April 1998.

[223] Yvan Saeys, Inaki Inza, and Pedro Larranaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.

[224] D. Salas-Gonzalez, J. M. Górriz, J. Ramírez, I. A. Illán, M. López, F. Segovia, R. Chaves, P. Padilla, C. G. Puntonet, and The Alzheimer's Disease Neuroimage Initiative. Feature selection using factor analysis for Alzheimer's diagnosis using F18-FDG PET images. *Medical Physics*, 37(11):6084–6095, 2010.

[225] Scikit-learn contributors. *sklearn.ensemble.RandomForestClassifier documentation*, 2016.

[226] I.W. Selesnick, R.G. Baraniuk, and N.C. Kingsbury. The dual-tree complex wavelet transform. *Signal Processing Magazine, IEEE*, 22(6):123 – 151, nov. 2005.

[227] S. Seth and J.C. Principe. Variable selection: A statistical dependence perspective. In *Machine Learning and Applications (ICMLA), 2010 Ninth International Conference on*, pages 931–936, 2010.

[228] Dinggang Shen and Christos Davatzikos. Very high-resolution morphometry using mass-preserving deformations and HAMMER elastic registration. *NeuroImage*, 18(1):28–41, 1 2003.

[229] Feng Shi, Li Wang, Guorong Wu, Yu Zhang, Manhua Liu, John H Gilmore, Weili Lin, and Dinggang Shen. Atlas construction via dictionary learning and group sparsity. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012*, pages 247–255. Springer, 2012.

[230] Laurent Sifre and Stéphane Mallat. Combined scattering for rotation invariant texture analysis. In *European Symposium on Artificial Neural Networks*, 2012.

[231] Laurent Sifre and Stéphane Mallat. Rotation, scaling and deformation invariant scattering for texture discrimination. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 1233–1240. IEEE, 2013.

[232] Laurent SIfre and Stéphane Mallat. Rigid-motion scattering for texture classification. *arXiv preprint arXiv:1403.1687*, 2014.

[233] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034, 2013.

[234] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[235] Nikhil Singh, Angela Y. Wang, Preethi Sankaranarayanan, P.Thomas Fletcher, and Sarang Joshi. Genetic, structural and functional imaging biomarkers for early detection of conversion from MCI to AD. In Nicholas Ayache, Herve Delingette, Polina Golland, and Kensaku Mori, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012*, volume 7510 of *Lecture Notes in Computer Science*, pages 132–140. Springer Berlin Heidelberg, 2012.

[236] Ying Song, Zhen Zhu, Yang Lu, Qiegen Liu, and Jun Zhao. Reconstruction of magnetic resonance imaging by three-dimensional dual-dictionary learning. *Magnetic Resonance in Medicine*, 71(3):1285–1298, 2014.

[237] A. Sotiras, C. Davatzikos, and N. Paragios. Deformable medical image registration: A survey. *Medical Imaging, IEEE Transactions on*, 32(7):1153–1190, July 2013.

[238] C. Studholme, D. L. G. Hill, and D. J. Hawkes. An overlap invariant entropy measure of 3D medical image alignment. *Pattern Recognition*, 32(1):71 – 86, 1999.

[239] Heung-Il Suk, Seong-Whan Lee, and Dinggang Shen. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. *NeuroImage*, 101:569–582, 11 2014.

[240] Heung-Il Suk and Dinggang Shen. Deep learning-based feature representation for AD/MCI classification. In Kensaku Mori, Ichiro Sakuma, Yoshinobu Sato, Christian Barillot, and Nassir Navab, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*, volume 8150 of *Lecture Notes in Computer Science*, pages 583–590. Springer Berlin Heidelberg, 2013.

[241] Heung-Il Suk and Dinggang Shen. Clustering-induced multi-task learning for AD/MCI classification. In Polina Golland, Nobuhiko Hata, Christian Barillot, Joachim Hornegger, and Robert Howe, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014*, volume 8675 of *Lecture Notes in Computer Science*, pages 393–400. Springer International Publishing, 2014.

[242] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[243] Richard Szeliski. Image alignment and stitching: A tutorial. *Foundations and Trends in Computer Graphics and Computer Vision*, 2(1), December 2006.

[244] Stefan J. Teipel, Christine Born, Michael Ewers, Arun L. W. Bokde, Maximilian F. Reiser, Hans-Jürgen Möller, and Harald Hampel. Multivariate deformation-based analysis of brain atrophy to predict Alzheimer's disease in mild cognitive impairment. *NeuroImage*, 38(1):13–24, 10 2007.

[245] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[246] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):pp. 267–288, 1996.

[247] Robert Tibshirani, Jacob Bien, Jerome Friedman, Trevor Hastie, Noah Simon, Jonathan Taylor, and Ryan J. Tibshirani. Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):245–266, 2012.

[248] Tong Tong, Robin Wolz, Pierrick Coupé, Joseph V. Hajnal, and Daniel Rueckert. Segmentation of MR images via discriminative dictionary learning and sparse coding: Application to hippocampus labeling. *NeuroImage*, 76:11 – 23, 2013.

[249] Tong Tong, Robin Wolz, Qinquan Gao, JosephV. Hajnal, and Daniel Rueckert. Multiple instance learning for classification of dementia in brain MRI. In Kensaku Mori, Ichiro Sakuma, Yoshinobu Sato, Christian Barillot, and Nassir Navab, editors, *Medical Image Computing and*

Computer-Assisted Intervention – MICCAI 2013, volume 8150 of *Lecture Notes in Computer Science*, pages 599–606. Springer Berlin Heidelberg, 2013.

[250] I. Tosic and P. Frossard. Dictionary learning. *Signal Processing Magazine, IEEE*, 28(2):27 –38, march 2011.

[251] J.A. Tropp and S.J. Wright. Computational methods for sparse solution of linear inverse problems. *Proceedings of the IEEE*, 98(6):948 –958, june 2010.

[252] Muhammad Usman, David Atkinson, Freddy Odille, Christoph Kolbitsch, Ghislain Vaillant, Tobias Schaeffter, Philip G. Batchelor, and Claudia Prieto. Motion corrected compressed sensing for free-breathing dynamic cardiac MRI. *Magnetic Resonance in Medicine*, 70(2):504–516, 2013.

[253] Laurens Van der Maaten, Eric Postma, and Jaap Van den Henrik. Dimensionality reduction: A comparative review. Technical report, TiCC, Tilburg University, Oct. 2009.

[254] Vladimir Vapnik. *The nature of statistical learning theory*. Springer Science & Business Media, 2000.

[255] Gael Varoquaux, Alexandre Gramfort, Fabian Pedregosa, Vincent Michel, and Bertrand Thirion. Multi-subject dictionary learning to segment an atlas of brain spontaneous activity. In *Proceedings of the 22nd international conference on Information processing in medical imaging*, IPMI'11, pages 562–573, Berlin, Heidelberg, 2011. Springer-Verlag.

[256] S.S. Vasanawala, M.J. Murphy, M.T. Alley, P. Lai, K. Keutzer, J.M. Pauly, and M. Lustig. Practical parallel imaging compressed sensing MRI: Summary of two years of experience in accelerating body MRI of pediatric patients. In *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*, pages 1039 –1043, 30 2011-april 2 2011.

[257] Prashanthi Vemuri, Jeffrey L. Gunter, Matthew L. Senjem, Jennifer L. Whitwell, Kejal Kantarci, David S. Knopman, Bradley F. Boeve, Ronald C. Petersen, and Clifford R. Jack Jr. Alzheimer's disease diagnosis in individual subjects using structural MR images: Validation studies. *NeuroImage*, 39(3):1186 – 1197, 2008.

[258] Dimitri Van De Ville, Mohamed L. Seghier, Francois Lazeyras, Thierry Blu, and Michael Unser. WSPM: Wavelet-based statistical parametric mapping. *NeuroImage*, 37(4):1205 – 1217, 2007.

[259] Maria Vounou, Eva Janousova, Robin Wolz, Jason L. Stein, Paul M. Thompson, Daniel Rueckert, and Giovanni Montana. Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease. *NeuroImage*, 60(1):700 – 716, 2012.

[260] Yanhua Wang and L. Ying. Compressed sensing dynamic cardiac cine MRI using learned spatiotemporal dictionary. *Biomedical Engineering, IEEE Transactions on*, 61(4):1109–1120, April 2014.

[261] Yanhua Wang, Yihang Zhou, and L. Ying. Undersampled dynamic magnetic resonance imaging using patch-based spatiotemporal dictionaries. In *Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on*, pages 294–297, April 2013.

[262] Michael W. Weiner, Dallas P. Veitch, Paul S. Aisen, Laurel A. Beckett, Nigel J. Cairns, Robert C. Green, Danielle Harvey, Clifford R. Jack, William Jagust, Enchi Liu, John C. Morris, Ronald C. Petersen, Andrew J. Saykin, Mark E. Schmidt, Leslie Shaw, Li Shen, Judith A. Siuciak, Holly Soares, Arthur W. Toga, and John Q. Trojanowski. The Alzheimer's Disease Neuroimaging Initiative: A review of papers published since its inception. *Alzheimer's & Dementia*, 9(5):e111–e194, 9 2013.

[263] Nick Weiss, Daniel Rueckert, and Anil Rao. Multiple sclerosis lesion segmentation using dictionary learning and sparse coding. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013*, pages 735–742. Springer, 2013.

[264] Eric Westman, J-Sebastian Muehlboeck, and Andrew Simmons. Combining MRI and CSF measures for classification of Alzheimer's disease and prediction of mild cognitive impairment conversion. *NeuroImage*, 62(1):229–238, 8 2012.

[265] Jason Weston, André Elisseeff, Bernhard Schölkopf, and Mike Tipping. Use of the zero norm with linear models and kernel methods. *J. Mach. Learn. Res.*, 3:1439–1461, March 2003.

[266] M.V. Wickerhauser. *Adapted Wavelet Analysis: From Theory to Software*. AK Peters, Ltd., Wellesley, Massachusetts, 1994.

[267] Robin Wolz, Rolf A. Heckemann, Paul Aljabar, Joseph V. Hajnal, Alexander Hammers, Jyrki Lötjönen, and Daniel Rueckert. Measurement of hippocampal atrophy using 4D graph-cut segmentation: Application to ADNI. *NeuroImage*, 52(1):109–118, 8 2010.

[268] Robin Wolz, Valtteri Julkunen, Juha Koikkalainen, Eini Niskanen, Dong Ping Zhang, Daniel Rueckert, Hilkka Soininen, Jyrki Lötjönen, et al. Multi-method analysis of MRI images in early diagnostics of Alzheimer's disease. *PloS one*, 6(10):e25446, 2011.

[269] J. Wright, Yi Ma, J. Mairal, G. Sapiro, T.S. Huang, and Shuicheng Yan. Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, 98(6):1031 –1044, 2010.

[270] Meng Wu, Andreas Maier, Qiao Yang, and Rebecca Fahrig. Improve path seeking accuracy for iterative reconstruction using the Karush-Kuhn-Tucker conditions. In *The 13th International Meeting on Fully ThreeDimensional Image Reconstruction in Radiology and Nuclear Medicine*, 2015.

[271] Igor Yakushev, Alexander Hammers, Andreas Fellgiebel, Irene Schmidtmann, Armin Scheurich, Hans-Georg Buchholz, Jürgen Peters, Peter Bartenstein, Klaus Lieb, and Mathias Schreckenberger. SPM-based count normalization provides excellent discrimination of mild Alzheimer's disease and amnestic mild cognitive impairment from healthy aging. *NeuroImage*, 44(1):43 – 50, 2009.

[272] Jonathan Young, Marc Modat, Manuel J. Cardoso, Alex Mendelson, Dave Cash, and Sebastien Ourselin. Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. *NeuroImage: Clinical*, 2(0):735 – 745, 2013.

[273] Hsiang-Fu Yu, Fang-Lan Huang, and Chih-Jen Lin. Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning*, 85(1):41–75, 2010.

[274] Lei Yu and Huan Liu. Efficient feature selection via analysis of relevance and redundancy. *The Journal of Machine Learning Research*, 5:1205–1224, 2004.

[275] Hyuk Jin Yun, Kichang Kwak, Jong-Min Lee, The Alzheimer's Disease Neuroimaging Initiative, et al. Multimodal discrimination of Alzheimer's disease based on regional cortical atrophy and hypometabolism. *PloS one*, 10(6):e0129250, 2015.

[276] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, volume 8689 of *Lecture Notes in Computer Science*, pages 818–833. Springer International Publishing, 2014.

[277] Daoqiang Zhang and Dinggang Shen. Semi-supervised multimodal classification of Alzheimer's disease. In *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*, pages 1628–1631, March 2011.

[278] Daoqiang Zhang and Dinggang Shen. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *NeuroImage*, 59(2):895 – 907, 2012.

[279] Daoqiang Zhang, Dinggang Shen, The Alzheimer's Disease Neuroimaging Initiative, et al. Predicting future clinical changes of MCI patients using longitudinal and multimodal biomarkers. *PloS one*, 7(3):e33182, 2012.

[280] Daoqiang Zhang, Yaping Wang, Luping Zhou, Hong Yuan, and Dinggang Shen. Multimodal classification of Alzheimer's disease and mild cognitive impairment. *NeuroImage*, 55(3):856 – 867, 2011.

[281] Kaihua Zhang, Lei Zhang, and Ming-Hsuan Yang. Real-time compressive tracking. In *Proceedings of ECCV*, 2012.

[282] Zheng Zhao and Huan Liu. Searching for interacting features. In *IJCAI*, volume 7, pages 1156–1161, 2007.

[283] Ji Zhu and Trevor Hastie. Support vector machines, kernel logistic regression and boosting. In *Multiple Classifier Systems*, pages 16–26. Springer, 2002.

[284] Ji Zhu and Trevor Hastie. Classification of gene microarrays by penalized logistic regression. *Biostatistics*, 5(3):427–443, 2004.

[285] Ji Zhu, Saharon Rosset, Trevor Hastie, and Rob Tibshirani. 1-norm support vector machines. *Advances in neural information processing systems*, 16(1):49–56, 2004.

[286] Xiaofeng Zhu, Heung-Il Suk, and Dinggang Shen. Multi-modality canonical feature selection for Alzheimer's disease diagnosis. In Polina Golland, Nobuhiko Hata, Christian Barillot, Joachim Hornegger, and Robert Howe, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014*, volume 8674 of *Lecture Notes in Computer Science*, pages 162–169. Springer International Publishing, 2014.

[287] Xiaofeng Zhu, Heung-Il Suk, and Dinggang Shen. A novel multi-relation regularization method for regression and classification in AD diagnosis. In Polina Golland, Nobuhiko Hata, Christian Barillot, Joachim Hornegger, and Robert Howe, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014*, volume 8675 of *Lecture Notes in Computer Science*, pages 401–408. Springer International Publishing, 2014.

[288] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

Permission for figure 1.1.4:

From: Katherine Gray <katherine.gray03@imperial.ac.uk>
To: michal.romaniuk06@imperial.ac.uk
Date: Fri 16/10/2015, 22:33

Hi Michal,

No problem, you're welcome to use this

Kat

From: Romaniuk, Michal
To: Katherine Gray <katherine.gray03@imperial.ac.uk>
Date: Wed 14/10/2015, 17:02

HI Kat,

Can I ask for permission to use Figure 1.5 from Chapter 1 of your thesis in my thesis? It's the one that explains some concepts about PET scanners.

Michal