UNIVERSITÀ DI PISA

Facoltà di Scienze Matematiche Fisiche e Naturali
Corso di Laurea Magistrale in Fisica Medica

Tesi di Laurea Magistrale

ANALYSIS OF BRAIN MAGNETIC RESONANCE IMAGES: VOXEL-BASED MORPHOMETRY AND PATTERN CLASSIFICATION APPROACHES

Relatrice:	**Candidata:**
Dott.ssa Alessandra Retico	**Alessia Giuliano**

Anno Accademico 2011/2012

Contents

List of Figures

List of Tables

List of Acronyms

ASD Autism Spectrum Disorders

AUC Area Under the receiver operating characteristic Curve

BA Broadmann Area

CNR Contrast to Noise Ratio

CSF Cerebro-Spinal Fluid

DARTEL Diffeomorphic Anatomical Registration using Exponentiated Lie algebra

DCT Discrete Cosine Transform

ESR Electron Spin Resonance

FID Free Induction Decay

FOV Field Of View

FWE Family Wise Error rate

GLM General Linear Model

GM Grey Matter

GRF Gaussian Random Field

ICBM Independent Community Bankers of Minnesota

ICM Iteration Conditional Modes

KKT Karush-Kuhn-Tucker

MAP Maximum A Posteriori

MNI Montreal Neurological Institute

MOG Mixture Of Gaussians model

MR Magnetic Resonance

MRI Magnetic Resonance Imaging

NMR Nuclear Magnetic Resonance

NVIQ Non Verbal Intelligence Quotient

RF Radio Frequency

RFE Recursive Feature Elimination

RFT Random Field Theory

ROC Receiver Operating Characteristic

SNR Signal to Noise Ratio

SPM Statistical Parametric Map

SVM Support Vector Machine

TBM Tensor-Based Morphometry

TPM Tissue Probability Map

VBM Voxel-Based Morphometry

WM White Matter

Introduction

This thesis aims to examine two different elaboration techniques of brain magnetic resonance imaging (MRI) data: the voxel-based morphometry (VBM) and the support vector machine (SVM) approaches. While the VBM is a standard and well-established mass-univariate method, the SVM multivariate analysis has been rarely implemented to investigate brain MRI data.

Standard univariate analysis of neuroimaging data has revealed several neuroanatomical differences between healthy individuals and patients suffering a large range of neurological and psychiatric disorders. Nevertheless, these findings have had limited clinical translation because of their significance only at group level. For this reason, recent attention has turned towards alternative types of analysis, including SVM machine learning, which allows classification of an individual's previously unseen data into a predefined group using a classification algorithm, developed on a training data set. An improvement of our knowledge on the pattern classification approach is necessary to be achieved, both to assess its exploratory capability and to point out advantages and disadvantages with respect to the more largely used VBM approach.

Despite these methods are potentially suitable to investigate a large variety of neurological and neuropsychiatric disorders such as Altzheimer's disease, schizofrenia, presymptomatic Huntington's and Parkinson's diseases, in the present study they have been employed with the purpose of detecting neuroanatomical and gender-related abnormalities in children with autism spectrum disorders (ASD). In fact, the differences in the neuroanatomy of young children with ASD are an intriguing and still poor investigated issue.

After a description of the physical principles of nuclear magnetic resonance and an overview of magnetic resonance imaging (see Chapters 1 and 2), we specified the two algorithms that represent the object of the current study: the voxel-based morphometry (see Chapter 3) and the support vector machines classification methods (see Chapter 4). Hence, we described the theoretical principles they are based on, pointing out schemes and procedures employed to implement these analysis approaches.

Then, we examined the application of VBM and SVM methods to an opportunely chosen sample of MRI data, composed by a total of 152 structural MRI scans. Specifically, as described in Chapter 5, our data set was characterized by 76 ASD children and 76 matched controls in the 2–7 year age range. The images were preprocessed applying the SPM8 algorithm, based on the diffeomorphic anatomical registration through exponentiated lie algebra (DARTEL) procedure.

The resulting grey matter (GM) segments were analyzed by applying the conventional voxel-wise two-sample t-test VBM analysis and employing the stringent family-wise error (FWE) rate correction according to random gaussian fields theory (see Chapter 6).

The same preprocessed GM segments were then analyzed using the SVM pattern classification approach, that presents the advantage of intrinsically taking into account interregional correlations (see Chapter 7). In fact, supervised machine learning methods are sensitive to spatially distributed and subtle effects in the brain that would be otherwise unobserved using standard univariate methods which focus on coarse differences at group level. Moreover, SVM would allow investigations about the predictive value of structural MRI scans, thus yielding results with a potentially high level of clinical translation. The pattern classification capability can be quantified in terms of the area under the receiver operating characteristic curve (AUC). The leave-pair-out cross-validation protocol has been adopted to evaluate the classification performance.

The recursive feature elimination (RFE) procedure has been implemented both to reduce the large number of features in the classification problem and to enhance the classification capability. The SVM-RFE allows also to localize the most discriminant voxels and to visualize them in a discrimination map. However, the pattern classification method was not employed in this study to predict the class membership of undiagnosed subjects, but as a figure of merit allowing to determine the optimal amount of voxels encoding the most relevant structural between-group differences that have to be retained in the discrimination maps.

With the aim of strengthening the SVM-based methods applied to brain data and to guarantee reliability and reproducibility of the results, we set up the following tests (see Chapter 8):

1. We evaluated the consistency among all discrimination maps, each obtained from one of the SVM leave-pair-out cross-validation steps, within the chosen range of number of retained features employed.

2. We assessed the dependency on the population of the training set within the cross-validation procedure. In this way we became able to check for the stability of our results with respect to the number of subjects employed during the learning phase. Furthermore, we can evaluate the classification performances for different cross-validation schemes.

Finally, as discussed in Chapter 9, we have reported results showing that SVM-RFE multivariate approach has the potential to contribute not only to distinguish ASD from matched control subjects, but also to disentangle the gender specificity of brain alterations, consistently with respect to the mass-univariate approach.

Chapter 1

Physical principles of Nuclear Magnetic Resonance

The main purpose of this chapter is to provide an overview of the basic principles of nuclear magnetic resonance (NMR). The intrinsic angular momentum of a hydrogen nucleus in a magnetic field precesses at the Larmor frequency, which depends linearly on the magnitude of the field. The physical fundamentals where independently discovered by F. Bloch and W. Purcell [1, 2] in the 1950s. At the beginnings, the related techniques were applied to study the physical and chemical properties of molecules. Only in the 1970s the potentialities of nuclear magnetic resonance as imaging technique were realized, as the first papers by Lauterbur and Mansfield testify. They had the idea that the introduction of a spatial variation of the magnetic field across the object, would induce a correlated spatial variation also in the Larmor frequencies. They pointed out that the different frequency components of a signal could be separated to give spatial information about the object. Magnetic resonance imaging (MRI) is possible since we can observe the way the protons contained in the human body respond to external magnetic fields. In the first stage of an MRI experiment, the proton spin orientation changes because of its interaction with a combination of applied magnetic fields. Subsequently, its changes in orientation can be measured with a coil detector. Although each proton field is tiny, a significant signal can be measured resulting from the sum of all fields of all affected protons of the body.

Magnetic resonance is a phenomenon found in magnetic systems that possess both magnetic and angular moments. As examples we have electron spins or nuclei of atoms. Furthermore, it is important to remark that, for static fields in the Tesla range, electrons have Larmor frequency in the microwave range. Here lies the reason why in medical field, resonance experiments are made on nuclei (NMR) and not on electrons (Electron Spin Resonance, ESR). In fact a microwave pulse would cause an excessive warming of biological tissues. For this reason, in the next sections we will refer to nuclei, especially to protons.

Extensive use of classical picture will be made, where the proton is viewed as a tiny spinning charge provided with a circulating electric current. Although much of NMR theory can be understood classically, we will show how to derive some results of this chapter in the quantum mechanical framework. Anyway, a complete and exhaustive treatment of NMR principles can be found within the classical texts of Abragam [3, 4], Carrington [5], Slichter [6], and Haacke [7], the last especially for MRI.

1.1 System of a single spin

In this section we focus on the response of a single spin, such as a proton, to an external field, ignoring the interactions of each proton with its surroundings. However, these interactions

Table 1.1: List of selected nuclear species with their spins, their associated magnetic moments in units of nuclear magneton μ_n ($\mu_n \equiv e\hbar/2m_p = 5.05 \times 10^{-27} A \cdot m^2$), gyromagnetic ratios, and their relative molar abundances (1M = 1molar = 1mole/liter). For comparison, the hydrogen (^{1}H) molarity of water is 110 M, and brain gray matter, for example, has a water content of 80% leading to an abundance of 88 M. The quoted body abundances will vary from tissue to tissue. Certain common elements are omitted, such as ^{12}C and ^{16}O, because their nuclear spins (and hence their magnetic moments) are zero. A negative sign for the moment and the gyromagnetic ratio refers to to the fact that the magnetic moment is anti-parallel to the angular momentum vector [7].

Nucleus	Spin (units of $\hbar$)	Magnetic moment (units of μ_n)	$\bar{\gamma}$ (MHz/T)	Abundance in human body
hydrogen ^{1}H	1/2	2.793	42.58	88M
sodium ^{23}Na	3/2	2.216	11.27	80mM
phosphorus ^{31}P	1/2	1.131	17.25	75mM
oxygen ^{17}O	5/2	−1.893	−5.77	16mM
fluorine ^{19}F	1/2	2.627	40.08	4μM

are very important in MRI, therefore they will be analyzed in the next section.

1.1.1 Interaction with a static magnetic field

We now consider an electrically charged particle that is rotating around its axis of inertia with constant angular velocity $\vec{\omega}$. This physical system possesses a magnetic moment $\vec{\mu}$ and an angular momentum $\vec{L}$ that are linked by the following relation:

$$\vec{\mu} = \gamma \vec{L}. \tag{1.1}$$

With the symbol γ we indicate the gyromagnetic (or magnetogyric) ratio of the particle, that for a particle characterized by charge q, mass m and Landé factor g is expressed by:

$$\gamma = g\frac{q}{2m}. \tag{1.2}$$

This proportionality constant depends on the particle or nucleus. For the proton, it is found to be:

$$\gamma = 2.675 \times 10^8 \ \frac{\text{rad}}{\text{s}}\text{T} \tag{1.3}$$

or, what may be referred as 'gamma-bar':

$$\bar{\gamma} \equiv \frac{\gamma}{2\pi} = 42.58. \ \text{MHz/T}. \tag{1.4}$$

Some values of $\bar{\gamma}$ for other nuclei are shown in table 1.1. Observing the values reported there, the reasons why imaging elements other than hydrogen is difficult in the human body are more clear. Smaller values of $\bar{\gamma}$ with respect to hydrogen are not the only reason, the problem is also their low concentration. We now introduce a magnetic field $\vec{B}$ that forms an angle θ with respect to the magnetic moment direction (figure 1.1). Therefore the system is subjected to a mechanic moment $\vec{\Gamma} = \vec{\mu} \wedge \vec{B}$ and the equation of temporal evolution for $\vec{L}$ is:

$$\frac{d\vec{L}}{dt} = \vec{\mu} \wedge \vec{B}. \tag{1.5}$$

Remembering the proportionality between $\vec{\mu}$ and $\vec{L}$ shown in eq. (1.1), we obtain:

$$\frac{d\vec{\mu}}{dt} = \vec{\mu} \wedge (\gamma \vec{B}). \tag{1.6}$$

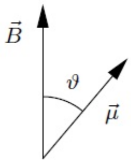


Figure 1.1: A charged particle with magnetic moment $\vec{\mu}$ that is precessing around a magnetic field $\vec{B}$.

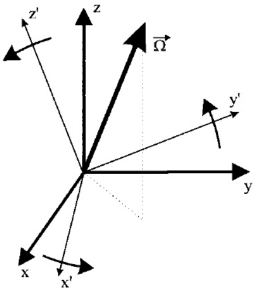


Figure 1.2: The primed reference frame rotating according to a general angular velocity $\vec{\Omega}$ relative to the unprimed reference frame. The rotation is counterclockwise around $\vec{\Omega}$ [7].

The last equation is valid independently of the field form. For example, if $\vec{B}$ does not change in time, from the eq. (1.6) we know that the angle between the magnetic moment and the field will remain constant every time, therefore $\vec{\mu}$ will design a cone around the rotation axis in a precession motion. The solution of the eq. (1.6) becomes easier to visualize if we introduce a new reference system $(\hat{x}', \hat{y}', \hat{z}')$ rotating with angular velocity $\vec{\Omega}$ with respect to the laboratory system $(\hat{x}, \hat{y}, \hat{z})$ (figure 1.2). We suppose that the two frames have coincident origins. The temporal evolution of the three versors of the rotating frame can be written as follows:

$$\frac{d\hat{x}'}{dt} = \vec{\Omega} \wedge \hat{x} \qquad \frac{d\hat{y}'}{dt} = \vec{\Omega} \wedge \hat{y} \qquad \frac{d\hat{z}'}{dt} = \vec{\Omega} \wedge \hat{z}. \tag{1.7}$$

Now we want to work out the temporal evolution equation for $\vec{\mu}$ in the rotating frame. If $\vec{\mu} = \mu_{x'}\hat{x}' + \mu_{y'}\hat{y}' + \mu_{z'}\hat{z}'$, we have[1]:

$$\begin{aligned} \frac{d\vec{\mu}}{dt} =& \frac{d\mu_{x'}}{dt}\hat{x}' + \mu_{x'}\frac{d\hat{x}'}{dt} + \frac{d\mu_{y'}}{dt}\hat{y}' + \mu_{y'}\frac{d\hat{y}'}{dt} + \frac{d\mu_{z'}}{dt}\hat{z}' + \mu_{z'}\frac{d\hat{z}'}{dt} = \\ & \frac{d\mu_{x'}}{dt}\hat{x}' + \frac{d\mu_{y'}}{dt}\hat{y}' + \frac{d\mu_{z'}}{dt}\hat{z}' + \vec{\Omega} \wedge \vec{\mu} = \\ & \frac{\partial\vec{\mu}}{\partial t} + \vec{\Omega} \wedge \vec{\mu}. \end{aligned} \tag{1.8}$$

Substituting in the last equation the eq. (1.6), we obtain:

$$\frac{\partial\vec{\mu}}{\partial t} + \vec{\Omega} \wedge \vec{\mu} = \vec{\mu} \wedge (\gamma\vec{B}) \tag{1.9}$$

that implies:

$$\frac{\partial\vec{\mu}}{\partial t} = \vec{\mu} \wedge (\gamma\vec{B} + \vec{\Omega}) = \vec{\mu} \wedge \gamma(\vec{B} + \frac{\vec{\Omega}}{\gamma}). \tag{1.10}$$

---

[1] We have chosen to distinguish the temporal derivative written in the laboratory frame from which in the rotating frame, using respectively the symbols: $\frac{d}{dt}$ and $\frac{\partial}{\partial t}$.

Evidently, the temporal evolution equation for $\vec{\mu}$, expressed in the coordinates of the rotating reference system, is the same of the one written in the laboratory system, (1.6), but with an effective magnetic field $\vec{B}_e$ defined as follows:

$$\vec{B}_e = \vec{B} + \frac{\vec{\Omega}}{\gamma}. \tag{1.11}$$

Therefore, if the system is subjected to a static magnetic field $\vec{B} = B_0\hat{z}$, a simple manner to solve the motion of $\vec{\mu}$ is to introduce a rotating system at $\vec{\Omega} = -\vec{\omega}_0$ ($\vec{\omega}_0 = \gamma B_0 \hat{z}$, the Larmor frequency) with respect to the laboratory system. In fact in the rotating frame we obtain:

$$\vec{B}_e = (B_0 + \frac{\omega_0}{\gamma})\hat{z} = 0 \implies \frac{\partial \vec{\mu}}{\partial t} = 0. \tag{1.12}$$

Then in the rotating frame $\vec{\mu}$ remains constant during time while in the laboratory it is rotating with respect to z axis at ω_0, moreover μ_z is constant, instead the expectation values of μ_x and μ_y are oscillating at the Larmor frequency.

We can obtain the same results with a quantum mechanical treatment. For this reason, we now consider a particle with spin $\vec{S}$ in the presence of a static magnetic field $\vec{B} = B_0\hat{z}$. The possession of both spin and charge confers on the particle a magnetic moment $\vec{\mu}$ which is proportional to the magnitude of the spin:

$$\vec{\mu} = \gamma \vec{S}. \tag{1.13}$$

Quantum theory requires that the available spin states are quantized, then the projections of the spin vector on a given direction can only take up one of a set of discrete values which are $+S, S-1, S-2, \ldots, -S$. The classical interaction Hamiltonian, which is also called Zeeman interaction, is:

$$\mathcal{H} = -\vec{\mu} \cdot \vec{B} = -\gamma \vec{S} \cdot \vec{B} = -\gamma B_0 \hbar m. \tag{1.14}$$

With m we indicated the angular momentum projection of the particle along z axis, measured in units of $\hbar$. For example, for a particle of spin $\frac{1}{2}$ there will be only two possible projections of $\vec{S}$ on z axis, then we can show the unique two non degenerate eigenvalues of energy (figure 1.3):

$$E_\downarrow = +\frac{\gamma B_0 \hbar}{2} \qquad E_\uparrow = -\frac{\gamma B_0 \hbar}{2}. \tag{1.15}$$

and the difference between the two energy levels is:

$$\Delta E = \gamma B_0 \hbar = \omega_0 \hbar \tag{1.16}$$

that is linearly proportional to the intensity of the static magnetic field and to the gyromagnetic ratio. Therefore, from a quantum mechanical point of view, the presence of a static magnetic field on a spin implies the splitting of its energy levels. Instead, classically, the field application leads to a precessional motion of the magnetic moment around the field direction. Now we want to work out the expected value of the dipole moment operator along the three axis directions. For this purpose, we begin to write a generic wave function as linear combination of the energy eigenstates of the interaction Hamiltonian:

$$|\psi(t)\rangle = C_\downarrow e^{-\frac{iE_\downarrow t}{\hbar}} |\downarrow\rangle + C_\uparrow e^{-\frac{iE_\uparrow t}{\hbar}} |\uparrow\rangle, \tag{1.17}$$

with $|\downarrow\rangle$ and $|\uparrow\rangle$ we indicated the eigenstates of the interaction Hamiltonian relative to the eigenvalues $E_\downarrow$ and $E_\uparrow$, and with $C_\downarrow$ and $C_\uparrow$ two generic complex constants. We aim to evaluate the following:

$$\langle \psi \,|\, \hat{\mu} \,|\, \psi \rangle = \gamma \left\langle \psi \,\middle|\, \hat{S} \,\middle|\, \psi \right\rangle = \frac{\gamma\hbar}{2} \langle \psi \,|\, \vec{\sigma} \,|\, \psi \rangle, \tag{1.18}$$

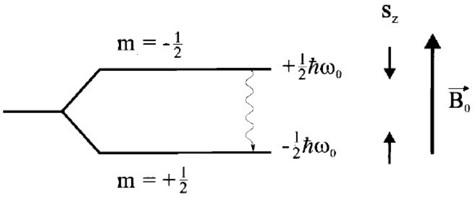


Figure 1.3: The Zeeman energy levels for a spin one-half system and a positive gyromagnetic ratio. The spin is parallel to the external field $B_0\hat{z}$ in the lower energy state. The wavy vertical line represents a transition from the higher to the lower state by photon emission [7].

with $\vec{\sigma}$ we have indicated the vector of the three Pauli matrices for spin $\frac{1}{2}$ particles. Now we can adopt polar coordinates to facilitate the calculation:

$$C_\downarrow = \cos\Theta e^{i\alpha_\downarrow} \qquad C_\uparrow = \sin\Theta e^{i\alpha_\uparrow} \tag{1.19}$$

and employing the following change of variables:

$$\Theta = \frac{\theta}{2} \qquad \phi_0 = \alpha_\uparrow - \alpha_\downarrow \tag{1.20}$$

we obtain the expectation values of $\hat{\mu}$ in the three cartesian directions:

$$\begin{aligned} \langle\hat{\mu}_x\rangle_t &= \frac{\gamma\hbar}{2}\sin\theta\cos(\phi_0 - \omega_0 t) \\ \langle\hat{\mu}_y\rangle_t &= \frac{\gamma\hbar}{2}\sin\theta\sin(\phi_0 - \omega_0 t) \\ \langle\hat{\mu}_z\rangle_t &= \frac{\gamma\hbar}{2}\cos\theta \end{aligned} \tag{1.21}$$

that are equivalent to the motion obtained using the classical approach, in fact the expectation value of the z component of the magnetic moment seems to be static, instead the transverse components have a rotatory motion with the Larmor frequency. In other words, the expectation value of the magnetic moment obeys the classical motion equation:

$$\frac{d}{dt}\left\langle\vec{\hat{\mu}}\right\rangle = \left\langle\vec{\hat{\mu}}\right\rangle \wedge \gamma\vec{B}. \tag{1.22}$$

1.1.2 Interaction with the radiofrequency field

The basic idea of a magnetic resonance experiment is to excite a macroscopic sample using some pulses of frequencies equal to the Larmor one and to observe how the system returns to equilibrium conditions. Since B_0 fields used in clinical and research medical MR applications are about few Tesla and the nuclei of potential interest in this field are those reported with the higher abundance in table 1.2, the Larmor frequencies fall in the radiofrequency range[2]. In general we will talk about radiofrequency (RF) pulse, and the instrument that generates this exciting pulse is called radiofrequency (RF) coil.

Rotating magnetic field

We now consider a magnetic moment $\vec{\mu}$ in the presence of both a static field $\vec{B_0} = B_0\hat{z}$ and a rotating field $\vec{B_1}(t) = B_1[\hat{x}cos(\omega t) - \hat{y}sen(\omega t)]$, that is left-circularly polarized, lying in the

---

[2] Larmor frequencies for ^{1}H are about 60–130 MHz for B_0 in the range 1.5–3 T (clinical research) and about 300 MHz when B_0=7 T (only research).

xy plane and generated by a RF coil. Then, indicating with $\vec{B}(t)$ the total magnetic field, we can write the motion equations for $\vec{\mu}$ both in the laboratory frame and in the reference system that is rotating at $\vec{\Omega}$ with respect to the first one:

$$\frac{d\vec{\mu}}{dt} = \vec{\mu} \wedge (\gamma \vec{B}(t)), \tag{1.23}$$

$$\frac{\partial\vec{\mu}}{\partial t} = \vec{\mu} \wedge [\gamma(\vec{B}(t) + \frac{\vec{\Omega}}{\gamma})]. \tag{1.24}$$

The effective magnetic field now becomes:

$$\vec{B}_{\text{eff}} = B_0\hat{z} + B_1[\hat{x}\cos(\omega t) - \hat{y}\sin(\omega t)] + \frac{\vec{\Omega}}{\gamma}. \tag{1.25}$$

If we adopt the resonance condition $\vec{\Omega} = -\vec{\omega}_0$, we obtain:

$$\vec{B}_{\text{eff}} = \vec{B}_1(t) \qquad \Longrightarrow \qquad \frac{\partial\vec{\mu}}{\partial t} = \vec{\mu} \wedge (\gamma \vec{B}_1(t)). \tag{1.26}$$

Therefore, in resonance condition, in the rotating frame the magnetic moment vector experiences only the rotating magnetic field and precedes around it at frequency $\omega_1 = \gamma B_1$. Instead, in the laboratory frame, the magnetic moment vector is preceding around the total magnetic field $\vec{B}(t)$. For example, we can choose a rotating reference system having the x'-axis coincident every instant with $\vec{B}_1$, therefore the spin will precede around the x'-axis at ω_1 (figure 1.4). If the RF pulse is experienced by the magnetic moment vector for a time interval τ, it will rotate with respect to the z axis of an angle $\Delta\theta$:

$$\Delta\theta = \gamma B_1 \tau \tag{1.27}$$

that is called *flip angle*. We can vary $\Delta\theta$ in arbitrary way changing the duration of the RF pulse, thus originating the so-called $\Delta\theta$-pulses (e. g. $\pi/2$-pulse, π-pulse). Instead, in the laboratory frame this motion will be seen as a spiral trajectory, resulting from the combination of the motion seen in the rotating frame and a rotation around the z axis at the Larmor frequency.

We can obtain the same results with a quantum approach. It is necessary to solve the Schrödinger equation for a particle of spin $\frac{1}{2}$ having magnetic moment $\vec{\mu}$:

$$i\hbar\frac{\partial}{\partial t}\psi = \mathcal{H}\psi \tag{1.28}$$

with ψ we have indicated the wave function of the particle and with $\mathcal{H}$ the interaction Hamiltonian, that has already been used:

$$\mathcal{H}(t) = -\vec{\mu}\cdot\vec{B}(t) = -\frac{\gamma\hbar}{2}[\sigma_z B_0 + B_1(\sigma_x cos(\omega t) - \sigma_y sin(\omega t))], \tag{1.29}$$

where we have reintroduced the Pauli matrices $\vec{\sigma}$. With the usual identifications for the frequencies, $\omega_0 \equiv \gamma B_0$ and $\omega_1 \equiv \gamma B_1$, we obtain:

$$\begin{aligned} \mathcal{H}(t) = &-\frac{\gamma\hbar}{2}\begin{pmatrix} B_0 & B_1[\cos(\omega t) + i\sin(\omega t)] \\ B_1[\cos(\omega t) - i\sin(\omega t)] & -B_0 \end{pmatrix} = \\ &= -\frac{\hbar}{2}\begin{pmatrix} \omega_0 & \omega_1 e^{i\omega t} \\ \omega_1 e^{-i\omega t} & -\omega_0 \end{pmatrix}. \end{aligned} \tag{1.30}$$

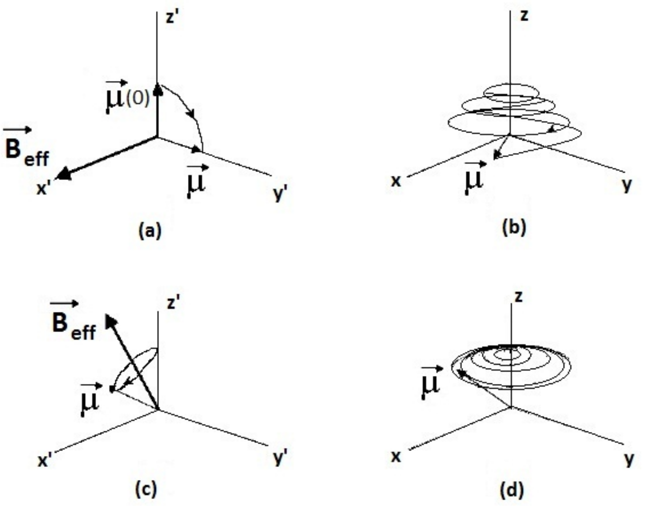


Figure 1.4: An on-resonance $\pi/2$ spin flip as viewed in the rotating frame (a) and in the laboratory frame (b). An off-resonance trajectory as viewed in the rotating frame (c) and in the laboratory frame (d). In MR applications, the frequency ω_1 would be much smaller in relation to the RF frequency, but the spiraling would then be much too dense to illustrate [7].

Therefore, the Schrödinger equation becomes:

$$i\hbar\frac{\partial}{\partial t}\psi = -\frac{\hbar}{2}\begin{pmatrix} \omega_0 & \omega_1 e^{i\omega t} \\ \omega_1 e^{-i\omega t} & -\omega_0 \end{pmatrix}\psi. \tag{1.31}$$

A simple set of differential equations can be found by a change of variables. For this reason we consider the following form for the wave function[3]:

$$\psi = \psi_1'(t)\psi_+ e^{\frac{i\omega_0 t}{2}} + \psi_2'(t)\psi_- e^{-\frac{i\omega_0 t}{2}} = \begin{pmatrix} \psi_1' e^{\frac{i\omega_0 t}{2}} \\ \psi_2' e^{-\frac{i\omega_0 t}{2}} \end{pmatrix} \tag{1.32}$$

In resonance condition ($\omega = \omega_0$) we obtain the temporal evolution equations for the two components of the transformed wave function:

$$i\hbar\frac{\partial}{\partial t}\psi' = -\frac{\hbar}{2}\mathcal{H}\psi', \tag{1.33}$$

$$i\hbar\begin{pmatrix} \frac{d\psi_1'}{dt}e^{\frac{i\omega_0 t}{2}} + \psi_1'\frac{i\omega_0}{2}e^{\frac{i\omega_0 t}{2}} \\ \frac{d\psi_2'}{dt}e^{-\frac{i\omega_0 t}{2}} - \psi_2'\frac{i\omega_0}{2}e^{-\frac{i\omega_0 t}{2}} \end{pmatrix} = -\frac{\hbar}{2}\begin{pmatrix} \omega_0\psi_1' e^{\frac{i\omega_0 t}{2}} + \omega_1\psi_2' e^{i\omega_0 t}e^{-\frac{i\omega_0 t}{2}} \\ \omega_1\psi_1' e^{\frac{i\omega_0 t}{2}}e^{-i\omega_0 t} - \omega_0\psi_2' e^{-i\frac{i\omega_0 t}{2}}, \end{pmatrix}, \tag{1.34}$$

$$\begin{aligned} \frac{d\psi_1'}{dt} &= \frac{i}{2}\omega_1\psi_2', \\ \frac{d\psi_2'}{dt} &= \frac{i}{2}\omega_1\psi_1'. \end{aligned} \tag{1.35}$$

The last have the following solutions:

$$\begin{aligned} \psi_1'(t) &= c_1\cos\Big(\frac{\omega_1 t}{2}\Big) + c_2\sin\Big(\frac{\omega_1 t}{2}\Big), \\ \psi_2'(t) &= -ic_2\cos\Big(\frac{\omega_1 t}{2}\Big) + ic_1\sin\Big(\frac{\omega_1 t}{2}\Big). \end{aligned} \tag{1.36}$$

---

[3] This change of variables corresponds to a wave function transformation from the laboratory to the Larmor rotating frame.

Having calculated the components of ψ', it is possible to obtain the expectation values of the magnetic moment operator along the three directions of the rotating frame (x', y', z). The explicit calculation is omitted here, but it is reported in detail in [6] and [7]. The final expressions of the desired expectation values are:

$$\langle\hat{\mu}_{x'}\rangle_t = \langle\hat{\mu}_{x'}\rangle_0\,, \tag{1.37}$$

$$\langle\hat{\mu}_{y'}\rangle_t = \langle\hat{\mu}_{y'}\rangle_0 \cos(\omega_1 t) + \langle\hat{\mu}_z\rangle_0 \sin(\omega_1 t), \tag{1.38}$$

$$\langle\hat{\mu}_z\rangle_t = -\langle\hat{\mu}_{y'}\rangle_0 \sin(\omega_1 t) + \langle\hat{\mu}_z\rangle_0 \cos(\omega_1 t). \tag{1.39}$$

Equations (1.38) and (1.39) represent a vector, of fixed magnitude, which is precessing clockwise the x' axis, in the rotating frame, with a precession frequency ω_1 as predicted by classical theory. Arbitrary initial conditions are obviously accomodated.

Oscillating magnetic field

From an experimental point of view it is easier to use an oscillating field perpendicular to the static field, instead of a rotating one. In fact it is sufficient to use a coil currying an oscillating current with appropriate frequency, having the main axis orientated perpendicularly to the static field direction. If the solenoid is orientated along the x-axis, the magnetic moment vector will be subjected to a transverse magnetic field:

$$\vec{B}_1(t) = \hat{x} B_1 \cos(\omega t). \tag{1.40}$$

An oscillating field can be seen as the superimposition of two fields: a rotating one $(\vec{B}_R(t))$ and a counter-rotating one $(\vec{B}_L(t))$:

$$\begin{aligned} \vec{B}_R(t) &= \hat{x}\frac{B_1}{2}\cos(\omega t) - \hat{y}\frac{B_1}{2}\sin(\omega t), \\ \vec{B}_L(t) &= \hat{x}\frac{B_1}{2}\cos(\omega t) + \hat{y}\frac{B_1}{2}\sin(\omega t). \end{aligned} \tag{1.41}$$

These two fields are identical after the substitution of ω with $-\omega$. Similarly, a rotating field can be obtained with the superimposition of two oscillating fields in quadrature, mutually perpendicular and with equal intensities. After these preliminary remarks, we can consider a magnetic moment in the presence of a static magnetic field and an oscillating field. We suppose that $\vec{B}_1(t)$ oscillates at ω_0 and we consider the rotating frame that has $\vec{B}_R$ along the x'-axis every time. In these conditions we have:

$$\begin{aligned} \frac{\partial\vec{\mu}}{\partial t} =& \vec{\mu}\wedge\gamma[\vec{B}_0 - \frac{\vec{\omega}_0}{\gamma} + \hat{x}'\frac{B_1}{2} + \hat{x}'\frac{B_1}{2}\cos(2\omega_0 t) + \hat{y}'\frac{B_1}{2}\sin(2\omega_0 t)] \\ =& \vec{\mu}\wedge\gamma\frac{B_1}{2}[\hat{x}'(1+\cos(2\omega_0 t)) + \hat{y}'\sin(2\omega_0 t)]. \end{aligned} \tag{1.42}$$

The precessional motion observed in the rotating frame would occur around an effective magnetic field obtained summing the static field along the x'-axis and the other, of the same intensity, that is rotating in the $x'y'$ plane at the frequency $2\omega_0$. But in the normal experimental conditions $B_1 \ll B_0$, therefore the precession observed in the rotating frame happens around a medium effective magnetic field equal to the rotating component $\vec{B}_R$ of $\vec{B}_1$, because the rapid rotation of the counter-rotating component results mediated to zero in a single precession (*rotating wave approximation*). Then, in the presence of an oscillating field of intensity $\vec{B}_1/2$, everything happens as the system was only subjected to its rotating component. Instead, when this approximation is no longer valid, the presence of the counter-rotating component cannot be neglected, therefore there is a shift of the observed resonance frequency (*Bloch-Siegart effect*).

1.2 System of many spins

In the previous sections we have always treated single spins, but it is important to extend our discussion up to physical systems containing a spin distribution. In fact, the interactions of the proton spin with its neighboring atoms lead to significant modifications to this behavior. The local fields change the spin precession frequency, and the proton can exchange energy with its surroundings. In this section we will model these effects, after having introduced the average magnetic dipole moment density ('magnetization').

1.2.1 Non-interacting spins

We can begin from a simplified point of view, considering N non-interacting spins. The total magnetization is represented by the mean value of the magnetic moments associated to a group of N spins contained in a small volume V:

$$\vec{M} = \lim_{V \to 0} \frac{1}{V} \sum_{i=1}^{N} \vec{\mu}_i. \tag{1.43}$$

This volume has to be small enough that external fields are to a good approximation constant over V, but also large to can contain an elevate number of magnetic dipoles. Because the spins are non-interacting, the magnetization interacts with the external magnetic field $\vec{B}_{ext} = B_0 \hat{z}$ obeying the same laws of a single magnetic dipole. Therefore the motion equation for the magnetization becomes:

$$\frac{d\vec{M}}{dt} = \gamma \vec{M} \wedge \vec{B}_{ext}. \tag{1.44}$$

It is more advantageous to separate the magnetization vector in a component ($\vec{M}_{\parallel}$) parallel to the static magnetic field and another ($\vec{M}_{\perp}$) perpendicular to it:

$$\vec{M}_{\parallel} = M_z \hat{z}, \tag{1.45}$$

$$\vec{M}_{\perp} = M_x \hat{x} + M_y \hat{y}. \tag{1.46}$$

These lead to the decoupled equations:

$$\frac{d\vec{M}_{\parallel}}{dt} = 0 \qquad \text{(non-interacting protons)}, \tag{1.47}$$

$$\frac{d\vec{M}_{\perp}}{dt} = \gamma \vec{M}_{\perp} \wedge \vec{B}_{ext} \qquad \text{(non-interacting protons)}. \tag{1.48}$$

1.2.2 Spin-Lattice interaction

Equation (1.47) is obviously wrong if we consider interacting protons. In fact, every magnetic moment will try to align with the static field to minimize its potential energy, through the exchange of energy with the surroundings. The classical formula for the potential energy of a magnetic moment subjected to an external magnetic field is:

$$U = -\vec{\mu} \cdot \vec{B}. \tag{1.49}$$

This means that the spin tends to line up parallel to the field direction in order to reach its minimum energy state, if energy can be transferred away. Since protons are considered to be in thermal contact with the lattice of nearby atoms, the thermal motion existing in the lattice can account for any change in a given proton spin energy. The energy associated to the magnetization vector is:

$$U_M = -\vec{M} \cdot \vec{B}_{ext} = -M_{\|} B_0, \tag{1.50}$$

involving only the longitudinal component of the magnetization. The magnetization has an equilibrium value M_0, at room temperature, that obeys Curie's law, depending on the absolute temperature T and the external field B_0:

$$M_0 = C\frac{B_0}{T}. \tag{1.51}$$

We can explicitly derive the constant C for a particle such as a proton. We have already explained that a particle of spin $\frac{1}{2}$, possessing a nuclear moment and subjected to a steady magnetic field can assume two possible physical configurations: the nuclear moment and the field can be parallel or antiparallel, having chosen the quantization axis along the field direction (figure 1.3). If we face with a macroscopic sample of N protons immersed in a steady magnetic field, we expect to find $N_\uparrow$ spins in the state $|\uparrow\rangle$ and $N_\downarrow$ in the state $|\downarrow\rangle$. It is obviously true that:

$$N = N_\uparrow + N_\downarrow. \tag{1.52}$$

We assume that every spin is in thermal contact with the rest of the set of spins and with the background lattice all at temperature T. The number N is taken very large with respect to the lattice size. We want to find the thermal equilibrium value of M_z making the calculation of the z-component of the average total magnetic moment for N spins distributed over all possible magnetic spin states, neglecting their translational motion. We expect that the populations are governed by the Boltzmann law:

$$N_\uparrow = \frac{N}{Z_{sing}} e^{-E_\uparrow/kT}, \tag{1.53}$$

$$N_\downarrow = \frac{N}{Z_{sing}} e^{-E_\downarrow/kT}, \tag{1.54}$$

$$\frac{N_\uparrow}{N_\downarrow} = e^{(-E_\uparrow + E_\downarrow)/kT} = e^{\frac{\gamma\hbar B_0}{kT}} = e^{\frac{\Delta E}{kT}}. \tag{1.55}$$

The normalization divisor Z_{sing} is the partition function for the single particle of spin 1/2:

$$Z_{sing} = \sum_{m=-1/2}^{1/2} e^{-E(m)/kT} = e^{-E_\uparrow/kT} + e^{-E_\downarrow/kT} = e^{\frac{\gamma\hbar B_0}{2kT}} + e^{-\frac{\gamma\hbar B_0}{2kT}} = 2cosh\Big(\frac{\gamma\hbar B_0}{2kT}\Big), \tag{1.56}$$

since for human body temperature ($\approx$ 310 K) and for protons it is true that $\hbar\omega_0/kT \approx 6.6 \times 10^{-6} B_0$, therefore:

$$e^{\hbar\omega_0/kT} = 1 + \frac{\hbar\omega_0}{kT} + \mathcal{O}\Big(\Big(\frac{\hbar\omega_0}{kT}\Big)^2\Big) + ... \approx 1 + 6.6 \times 10^{-6} B_0. \tag{1.57}$$

In this way we can estimate $Z_{sing} \approx 2$, that implies:

$$N_\uparrow \approx \frac{N}{2}\Big(1 + \frac{\gamma\hbar B_0}{2kT}\Big), \tag{1.58}$$

$$N_\downarrow \approx \frac{N}{2}\Big(1 - \frac{\gamma\hbar B_0}{2kT}\Big). \tag{1.59}$$

Now we can estimate the equilibrium value of the magnetization:

$$M_0 = \frac{1}{V}(N_\uparrow\mu_\uparrow + N_\downarrow\mu_\downarrow) = \frac{\gamma\hbar(N_\uparrow - N_\downarrow)}{2V} \approx \frac{\rho_0\gamma^2\hbar^2 B_0}{4kT} = \chi B_0. \tag{1.60}$$

Where $\rho_0 = N/V$ represents the density of spins per unit of volume and χ the magnetic susceptibility. Therefore, we have obtained that N nuclei with spin 1/2 have $\chi = \rho_0 \gamma^2 \hbar^2 / 4kT$ and $C = \chi T$. In general, the susceptibility of N nuclei with spin I is given by the following relation:

$$\chi = \rho_0 \frac{\gamma_I^2 \hbar^2 I(I+1)}{3kT}. \tag{1.61}$$

We remark that for a proton at human body temperature and indicating with ΔN the difference between the populations $N_\uparrow$ and $N_\downarrow$, we can estimate:

$$\frac{\Delta N}{N} \approx \frac{\gamma \hbar B_0}{2kT} \approx 6.6 \times 10^{-6} B_0, \tag{1.62}$$

therefore for a static field equal to 1 T, only 7 protons in one million contribute to the signal. However the proton density in each voxel is so large (about $10^{17}/\text{mm}^3$) that results a measurable signal, in fact 10^{11} of these contribute to the formation of the signal.

Now we suppose that the magnetization vector is disturbed from its equilibrium value $M_0 \hat{z}$, for example because of the application of a RF pulse. Then, after a characteristic time interval, the magnetization will return to its equilibrium value, as a result of the continue presence of the static field. It will verify a growth of the interaction rate with the lattice, that implies a rate of change of the longitudinal magnetization $(dM_z(t)/dt)$ proportionally to the difference $M_0 - M_z(t)$, with a proportionality constant that can be empirically determined. After these considerations, eq. (1.47) can be replaced by the following:

$$\frac{dM_z(t)}{dt} = \frac{1}{T_1}(M_0 - M_z(t)). \tag{1.63}$$

T_1 is the experimental 'spin-lattice relaxation time' and represents the time scale of the growth rate. This is due to the energy exchanges between the nuclei and the lattice and it describes the system of spins ability to reorganize themselves to achieve the equilibrium state, that is determined by the static magnetic field and by the temperature. Every molecule has some characteristic frequencies at which it exchanges energy with the surroundings, that depend on its microscopic structure and on its possible rotational and translational modes. One of the most important types of interaction is the dipolar one. When two magnetic dipoles are at a certain distance, they interact through a dipole field that depends on their relative angle and distance. If these frequencies would be comparable to the Larmor frequency, the system will be able to easily generate a magnetization through energy exchanges, therefore we expect a short T_1. If we regard a biological medium, water characteristic frequencies are greater then the Larmor frequency, therefore T_1 will be long, instead proteins will have short T_1. Molecules such as lipids, in particular the cholesterol, are relatively large, hence their T_1 will be short. For this reason in a MRI image obtained with T_1 contrast, lipids will appear light instead mediums with a considerable water concentration, such as the cerebrospinal fluid, will appear dark. Typical values of this parameter for different tissues are reported in table 1.2. After the application of a RF pulse, the longitudinal magnetization shows an exponential form in its evolution from the initial value $M_z(0)$ to the equilibrium value M_0 (figure 1.5 (a)):

$$M_z(t) = M_z(0) e^{-t/T_1} + M_0(1 - e^{-t/T_1}) \qquad (\vec{B}_{\text{ext}} \parallel \hat{z}). \tag{1.64}$$

1.2.3 Spin-Spin interaction

The spin-spin interaction is an important mechanism for the transverse magnetization decay. The total transverse magnetization is the vector sum of all the individual transverse

Table 1.2: Representative values of T_1 and T_2, for hydrogen components of different human body tissues at $B_0 = 1.5$ T and 37 ° C (human body temperature) [7].

Tissue	T_1 (ms)	T_2 (ms)
gray matter	950	100
white matter	600	80
muscle	900	50
cerebrospinal fluid	4500	2200
fat	250	60
blood	1200	100 – 200

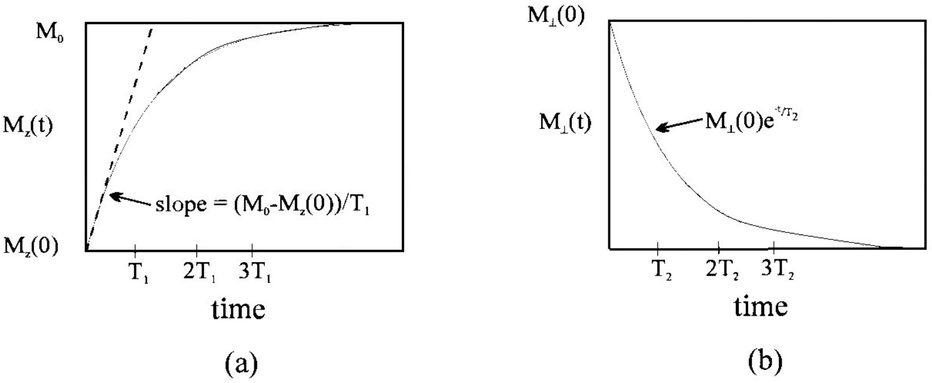


Figure 1.5: (a) The regrowth of the longitudinal component of magnetization from the initial value $M_z(0)$ to the equilibrium value M_0. (b) The decay of the magnitude of the transverse magnetization from an initial value $M_\perp(0)$ [7].

components of the magnetic moments. Spins interact not only with the static field $\vec{B}_{\text{ext}}$, but also experience the presence of local fields that are combinations of the applied fields and of the fields generated by their neighbors. The rate of reduction in the transverse magnetization can be characterized by an experimental parameter: the 'spin-spin relaxation time' T_2. This represents the characteristic time interval that the spins need to accumulate a phase shift capable to reset the transverse magnetization (see eq. 2.6). Therefore, the eq. (1.48) has to be modified in the following way:

$$\frac{d\vec{M}_\perp(t)}{dt} = \gamma \vec{M}_\perp(t) \wedge \vec{B}_{ext} - \frac{1}{T_2}\vec{M}_\perp(t). \tag{1.65}$$

In the rotating reference frame the last differential equation assumes the standard decay-rate form:

$$\frac{\partial \vec{M}_\perp(t)}{\partial t} = -\frac{1}{T_2}\vec{M}_\perp(t) \qquad \text{(rotating frame)} \tag{1.66}$$

with the solution:

$$\vec{M}_\perp(t) = \vec{M}_\perp(0)e^{-t/T_2} \qquad \text{(rotating frame)}, \tag{1.67}$$

that describes the exponential decay of the magnitude of $\vec{M}_\perp$ in either the laboratory or the rotating frame (figure 1.5 (b)).

Furthermore, there is another dephasing effect that contributes to decrease the time interval necessary to reset the transverse magnetization. In fact $\vec{B}_0$ is not homogeneous over all the sample because of local fields, generated by interactions between the spins. Therefore there are local differences between the precessional frequencies of the magnetic moments that lead the spins to dephase, conducting to a reduction of the net magnetization vector. To better explain this concept we consider two spins that at $t = 0$ are in phase in

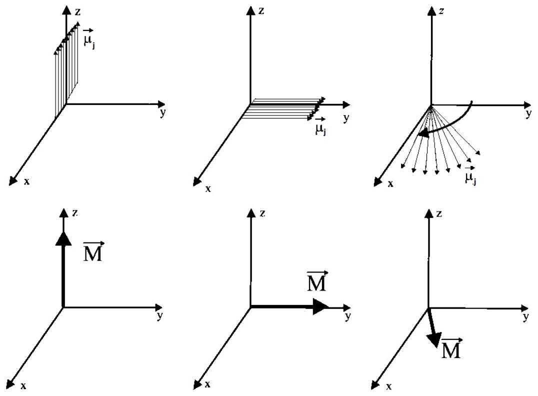


Figure 1.6: The upper sequence shows a 90° tip of a set of spins into the transverse plane such that they all lie along the y-axis of the laboratory frame at some instant time, as shown in the middle figure. Precession of the individual spins in the xy plane immediately follows (the recovery of the longitudinal magnetization is ignored since the focus is on transverse magnetization dephasing effects). The lower sequence shows the same process, but expressed in terms of the net transverse magnetization, which decreases in magnitude during the precession because of the fanning out of the spins [7].

the xy plane because of the application of an RF field. But, if they experience a different static magnetic field, a difference ΔB_0 will produce a phase shift $\Delta\omega$ in their precessional motion. Therefore, after a time t, they will have been accumulated a phase displacement equal to:

$$\Delta\phi = \Delta\omega t. \tag{1.68}$$

This reduction in an initial value of $\vec{M}_\perp$ is usually characterized by the decay time T_2'. We can define the relaxation rates as:

$$R_2 \equiv \frac{1}{T_2}, \qquad R_2' \equiv \frac{1}{T_2'}. \tag{1.69}$$

Then, the total relaxation rate is the sum of the internal and the external rates:

$$R_2^* = R_2 + R_2'. \tag{1.70}$$

In terms of an overall relaxation time $T_2^* \equiv 1/R_2^*$:

$$\frac{1}{T_2^*} = \frac{1}{T_2} + \frac{1}{T_2'}. \tag{1.71}$$

The loss of transverse magnetization due to T_2' is recoverable, instead the intrinsic T_2 losses are not recoverable and are related to local, random, time-dependent field variation. In the dephasing effect no energy is lost, therefore to eq. (1.65) corresponds a higher relaxation rate than eq. (1.63). Defining $R_1 = \frac{1}{T_1}$, we have:

$$R_2 > R_1 \qquad \text{or} \qquad T_2 < T_1. \tag{1.72}$$

Typical values of T_2 are shown in table 1.2. If the molecular environment makes possible many changes in spins position and orientation, these fluctuations will be mediated, therefore the phase loss will be less efficient and T_2 will result longer. For this reason blood and cerebrospinal fluid (CSF) have longer T_2 with respect to lipidic tissues.

1.2.4 Bloch equation

The differential equations (1.63) and (1.65) for the magnetization in the presence of a static magnetic field $\vec{B}_{ext} = B_0\hat{z}$, that include the relaxation terms, can be combined into the following empirical vector equation, that is more commonly called *Bloch equation*:

$$\frac{d\vec{M}}{dt} = \gamma\vec{M} \wedge \vec{B}_{ext} + \frac{1}{T_1}(M_0 - M_z)\hat{z} - \frac{1}{T_2}\vec{M}_\perp. \tag{1.73}$$

The last represents the fundamental equation of magnetic resonance and it was phenomenologically introduced for the first time by F. Bloch in 1946 [8]. We can derive the three component equations:

$$\frac{dM_z}{dt} = \frac{M_0 - M_z}{T_1}, \tag{1.74}$$

$$\frac{dM_x}{dt} = \omega_0 M_y - \frac{M_x}{T_2}, \tag{1.75}$$

$$\frac{dM_y}{dt} = -\omega_0 M_x - \frac{M_y}{T_2}. \tag{1.76}$$

The first equation is the same as the eq. (1.63) and obviously has the analogous solution. The last two can be solved with a simple change of variables, as explained in [7]. The solutions are the following:

$$M_x(t) = e^{-t/T_2}(M_x(0)\cos(\omega_0 t) + M_y(0)\sin(\omega_0 t)), \tag{1.77}$$

$$M_y(t) = e^{-t/T_2}(M_y(0)\cos(\omega_0 t) - M_x(0)\sin(\omega_0 t)), \tag{1.78}$$

$$M_z(t) = M_z(0)e^{-t/T_1} + M_0(1 - e^{-t/T_1}). \tag{1.79}$$

The eqs. (1.77) and (1.78) have sinusoidal terms modified by a decay factor owing to the transverse relaxation effect. Therefore the longitudinal component relaxes from its initial value to the equilibrium value M_0 and the transverse component precedes clockwise and it decreases in magnitude. An example of the resulting ‘corkscrew’ trajectory for an initial magnetization lying in the transverse plane i illustrated in figure 1.7. The equilibrium or steady-state solutions can be found applying the asymptotic limit $t \to \infty$. In that limit all the exponentials vanish and we obtain:

$$M_x(\infty) = M_y(\infty) = 0 \qquad M_z(\infty) = M_0. \tag{1.80}$$

Now we consider the more common situation in which there is a combination of a static and a RF field. In fact the RF pulse is necessary if we want to tip $\vec{M}$ from its equilibrium direction and then to observe how the system relaxes, with the aim of measuring the characteristic relaxation times T_1 and T_2. Therefore, for a left-circularly polarized RF field $\vec{B}_1$ which is at rest in the rotating frame and parallel to $\hat{x}'$, the total external field $\vec{B}.upext$ becomes:

$$\vec{B}_{\text{ext}} = B_0\hat{z} + B_1\hat{x}' \tag{1.81}$$

And the effective magnetic field is the following:

$$\vec{B}_{\text{eff}} = \left(B_0 - \frac{\omega}{\gamma}\right)\hat{z} + B_1\hat{x}'. \tag{1.82}$$

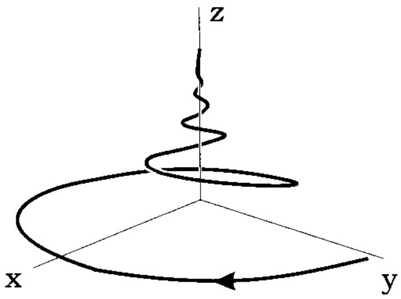


Figure 1.7: The trajectory of the tip of the magnetization vector showing the combined regrowth of the longitudinal magnetization and decay of the transverse components. The initial value was along the y-axis and the reference frame is the laboratory [7].

We can find the Bloch component equations for $\vec{B}_{\text{eff}}$ in the rotating frame:

$$\frac{\partial M_z}{\partial t} = -\omega_1 M_{y'} + \frac{M_0 - M_z}{T_1}, \tag{1.83}$$

$$\frac{\partial M_{x'}}{\partial t} = \Delta\omega M_{y'} - \frac{M_{x'}}{T_2}, \tag{1.84}$$

$$\frac{\partial M_{y'}}{\partial t} = -\Delta\omega M_{x'} - \omega_1 M_z - \frac{M_{y'}}{T_2}, \tag{1.85}$$

with $\Delta\omega \equiv \omega_0 - \omega$ we indicate the 'off-resonance' contributions. We repeat that ω_0 is the Larmor frequency, ω_1 is the spin frequency due to the RF field and ω is the RF laboratory frequence of oscillation. Now we want to solve these equations in dependence on the RF pulse duration.

In most MR measurements the RF pulses have a very small duration (τ_{rf}) with respect to the relaxation times. Therefore typical values for ω_1 are much greater then the decay rates $\frac{1}{T_1}$ and $\frac{1}{T_2}$. In these conditions we can solve the eqs. (1.83)-(1.84) firstly neglecting the relaxation terms, hence we have to employ the simple solutions of the eq. (1.24). Secondly, the RF pulse is considered to be turned off ($\omega_1 = 0$), therefore we face with the eqs. (1.74)-(1.76). In this way, in the rotating frame and in resonance condition, we have a simple exponential decay, described by the eqs. (1.77)-(1.79).

Even if in medicine it is unusual, sometimes a long duration RF pulse is used. The sample is said to be 'saturated' and the long-term behavior of the magnetization can be described by steady-state solutions, obtaining:

$$M_{x'} = \frac{M_0 \gamma B_1 \Delta\omega T_2^2}{1 + (\Delta\omega T_2)^2 + \gamma^2 B_1^2 T_1 T_2} \tag{1.86}$$

$$M_{y'} = \frac{M_0 \gamma B_1 T_2}{1 + (\Delta\omega T_2)^2 + \gamma^2 B_1^2 T_1 T_2} \tag{1.87}$$

$$M_z = \frac{M_0(1 + (\Delta\omega T_2)^2)}{1 + (\Delta\omega T_2)^2 + \gamma^2 B_1^2 T_1 T_2}. \tag{1.88}$$

We can underline that the term $\gamma^2 B_1^2 T_1 T_2$, said to be 'saturation term', can be neglected if $\omega_1 \ll \frac{1}{T_1}$, equivalently if the RF field amplitude is great with respect to the spins absorbtion and dissipation energy time scale, the system will be subjected to saturation phenomena and T_1 will disappear from the final solution. This is a proof of the fact that the longitudinal relaxation time contains all the information about the energetic exchanges. Moreover it is true that $M_{x'} = M_{y'}\Delta\omega T_2$, hence the transverse components are one in phase and the other in quadrature with the excitation, therefore the first is responsible of the absorbtion and the second of the excitation. In addition we can calculate the complex magnetic susceptibility $\chi = \chi' + i\chi''$:

$$\chi' = \frac{M_0\gamma T_2}{2}\frac{\Delta\omega T_2}{1+(\Delta\omega T_2)^2}, \tag{1.89}$$

$$\chi'' = \frac{M_0\gamma T_2}{2}\frac{1}{1+(\Delta\omega T_2)^2}, \tag{1.90}$$

and the absorbed power is:

$$< P > \propto \omega_0\chi'' \propto \frac{\omega_0^2}{1+(\Delta\omega T_2)^2}, \tag{1.91}$$

that is a Lorentian function in the variable ω centered in ω_0 and width proportional to T_2. We can also notice that the absorbed power grows proportional to ω_0, therefore is linear in the static field. The Fourier transform of the last function is the time signal that is revealed in a magnetic resonance measure, said to be Free Induction Decay (FID):

$$\text{signal} \propto e^{-t/T_2}\sin(\omega_0 t). \tag{1.92}$$

1.3 Magnetic Resonance signal detection

We suppose of having a sample in the presence of a static magnetic field along the z axis of the laboratory frame and that has been applied an RF pulse to rotate the magnetization in the xy plane. We want to reveal the transverse magnetization using a coil orientated in this plane. Generally this receive coil is the same of the one that transmit the RF pulse. An electromotive force ϵ would be created in any coil through which the spin's magnetic flux sweeps, as a consequence of Faraday's law. The time-dependent form of this current carries the information that will be transformed into an image of the sample. According to Faraday's law, the electromotive force induced in a coil by a change in its magnetic flux is:

$$\epsilon = -\frac{d\Phi}{dt}, \tag{1.93}$$

where the flux through the coil is:

$$\Phi = \int_{\text{coil area}} \vec{B}\cdot d\vec{S}. \tag{1.94}$$

As we shall see later, applying the principle of reciprocity, the last equation can be converted into a form which is more useful for MRI, where the roles of the magnetization source and the detection coil are reversed. The magnetization of the sample is associated with an effective current density:

$$\vec{J_M}(\vec{r},t) = \vec{\nabla}\wedge\vec{M}(\vec{r},t). \tag{1.95}$$

We can express the flux in terms of the vector potential at a position $\vec{r}$:

$$\vec{A}(\vec{r}) = \frac{\mu_0}{4\pi}\int d^3r'\frac{\vec{J_M}(\vec{r'})}{|\vec{r}-\vec{r'}|}. \tag{1.96}$$

The effects due to the time delay between the source and the measurement of the field are ignored. We can calculate the magnetic field:

$$\vec{B} = \vec{\nabla}\wedge\vec{A}. \tag{1.97}$$

Using the Stokes theorem we can obtain the following expression for the flux:

$$\Phi = \int_{\text{coil area}} (\vec{\nabla} \wedge \vec{A}) \cdot d\vec{S} = \oint d\vec{l} \cdot \vec{A}. \tag{1.98}$$

The use of eqs. (1.95), (1.96) and (1.98), an integration by parts (where a surface term can be neglected for finite sources), and the vector identity $\vec{A} \cdot (\vec{B} \wedge \vec{C}) = -(\vec{A} \wedge \vec{C}) \cdot \vec{B}$ give:

$$\begin{aligned} \Phi &= \oint d\vec{l} \cdot \left[\frac{\mu_0}{4\pi} \int d^3r' \frac{\vec{\nabla}' \wedge \vec{M}(\vec{r}')}{|\vec{r} - \vec{r}'|} \right] \\ &= \frac{\mu_0}{4\pi} \int d^3r' \oint d\vec{l} \cdot \left[\left(-\vec{\nabla}' \frac{1}{|\vec{r} - \vec{r}'|} \right) \wedge \vec{M}(\vec{r}') \right] \\ &= \frac{\mu_0}{4\pi} \int d^3r' \vec{M}(\vec{r}') \cdot \left[\vec{\nabla}' \wedge \left(\oint \frac{d\vec{l}}{|\vec{r} - \vec{r}'|} \right) \right]. \end{aligned} \tag{1.99}$$

For current loops the vector potential assumes the form that follows:

$$\vec{A}(\vec{r}') = \frac{\mu_0}{4\pi} \oint \frac{I d\vec{l}}{|\vec{r} - \vec{r}'|}. \tag{1.100}$$

The last show that the curl integral in the eq. (1.99) is exactly $\vec{\mathcal{B}}^{\text{receive}}$, the magnetic field per unit current that would be produced by the coil at the point $\vec{r}'$:

$$\vec{\mathcal{B}}^{\text{receive}}(\vec{r}') = \frac{\vec{B}(\vec{r}')}{I} = \vec{\nabla}' \wedge \left(\frac{\mu_0}{4\pi} \oint \frac{d\vec{l}}{|\vec{r} - \vec{r}'|} \right). \tag{1.101}$$

Finally, the flux can be written as:

$$\Phi(t) = \int_{\text{sample}} d^3r \vec{\mathcal{B}}^{\text{receive}}(\vec{r}) \cdot \vec{M}(\vec{r}, t). \tag{1.102}$$

It is important to notice that the flux in eq. (1.102) depends upon $\vec{\mathcal{B}}^{\text{receive}}$, the magnetic field per unit of current produced by the detection coil at all points where the magnetization is nonzero and this is an example of the principle of reciprocity. In fact, we began with a surface integration over the detection coil area and we are arrived to a volume integration over the region of nonzero magnetization, in other words we have to calculate the flux that has been emanated from the detection coil, per unit of current, through the rotating magnetization. The electromotive force induced in the coil is:

$$\epsilon = -\frac{d}{dt}\Phi(t) = -\frac{d}{dt} \int_{\text{sample}} d^3r \vec{M}(\vec{r}, t) \cdot \vec{\mathcal{B}}^{\text{receive}}(\vec{r}). \tag{1.103}$$

The dependence of the electromotive force on the excitation or transmit field is implicit in the dependence of eq. (1.103) on the magnetization. The fundamental signal in an MR measure derives from the detection of the electromotive force induced in the receive coil ($signal \propto \epsilon$). Then, inserting the known solutions for $\vec{M}(\vec{r}, t)$ (see eqs. (1.77)-(1.79)) into the integrand and evaluating the time derivative, it is possible to show that the longitudinal magnetization can be neglected, even when there is a nonzero z-component for the receive-coil field. This is true because for static fields at the Tesla level and considering protons, the Larmor frequency ω_0 is at least four order of magnitude larger than typical values of $1/T_1$ and $1/T_2$, therefore the proportionality factors, that appear in front of the time derivative of the z-term, can be neglected. Now we choose to adopt the more convenient complex formalism for the transverse magnetization:

$$M_+(t) \equiv M_x(t) + iM_y(t). \tag{1.104}$$

Therefore the solution for a static field in this representation is:

$$M_+(t) = e^{-i\omega_0 t - t/T_2} M_+(0), \tag{1.105}$$

$$M_+(t) = |M_+(t)|e^{i\phi(t)} = M_\perp(t)e^{i\phi(t)} = e^{-t/T_2} M_+(0)e^{i\phi(t)}, \tag{1.106}$$

since that $|M_+(t)| = M_\perp(t)$ and $M_+(0) = M_\perp(0)$. Therefore:

$$M_\perp(t) = M_\perp(0)e^{-t/T_2}, \tag{1.107}$$

with $\phi(t) = -\omega_0 t + \phi(0)$.
In this way we can obtain the following form for the detected MR signal:

$$signal \propto \omega_0 \int d^3r e^{-t/T_2(\vec{r})} M_\perp(\vec{r},0)[\mathcal{B}_x^{\text{receive}}(\vec{r})\sin(\omega_0 t - \phi_0(\vec{r})) + \mathcal{B}_y^{\text{receive}}(\vec{r})\cos(\omega_0 t - \phi_0(\vec{r}))]. \tag{1.108}$$

It is possible to simplify the last expression writing the receive field laboratory components in terms of the magnitude $\mathcal{B}_\perp$ and the angle $\theta_\mathcal{B}$:

$$\mathcal{B}_x^{\text{receive}} \equiv \mathcal{B}_\perp \cos\theta_\mathcal{B} \qquad \mathcal{B}_y^{\text{receive}} \equiv \mathcal{B}_\perp \sin\theta_\mathcal{B}. \tag{1.109}$$

Therefore we obtain:

$$signal \propto \omega_0 \int d^3r e^{-t/T_2(\vec{r})} M_\perp(\vec{r},0)\mathcal{B}_\perp(\vec{r})\sin(\omega_0 t + \theta_\mathcal{B}(\vec{r}) - \phi_0(\vec{r})). \tag{1.110}$$

In a more general situation we have to modify the last equation with the replacement $T_2 \rightarrow T_2^*$, that must be made in the presence of external field inhomogeneities. In studies of small homogeneous samples, where the spatial dependence can be neglected, we can consider all quantities inside the eq. (1.110) to be independent of $\vec{r}$. Hence, if the sample volume is V_s, we obtain:

$$\text{signal} \propto \omega_0 V_s e^{-t/T_2} M_\perp \mathcal{B}_\perp sin(\omega_0 t + \theta_\mathcal{B} - \phi_0) \qquad \text{(space-independent limit)}. \tag{1.111}$$

Since the signal is characterized by rapid oscillations at the Larmor frequency, it is necessary an electronic step of demodulation and filtering, with the aim of removing from the signal expression the presence of ω_0. Demodulation corresponds to the multiplication of the signal by a sinusoid or cosinusoid with a frequency at or near ω_0. The resulting demodulated signal is low pass filtered to eliminate the high frequency component. These signal processing operations are made in both a real and an imaginary channels. In other words, the signal, as measured in the laboratory, oscillates rapidly near the high Larmor frequency ($\omega_0 = \gamma B_0$ and $B_0 \approx$ Tesla). The demodulated and filtered signal, essentially the signal measured in the rotating frame defined by the reference frequency $\Omega = \omega_0 + \delta\omega$, is free of the rapid Larmor oscillation and it oscillates at the offset frequency $\delta\omega$ ($\delta\omega t \equiv \theta_\mathcal{B}(\vec{r}) - \phi_0(\vec{r})$). Finally, the eq. (1.110) can be replaced by:

$$s(t) \propto \omega_0 \int d^3r e^{-t/T_2(\vec{r})} M_\perp(\vec{r},0)\mathcal{B}_\perp(\vec{r})e^{i((\Omega-\omega_0)t+\phi_0(\vec{r})-\theta_\mathcal{B}(\vec{r}))}. \tag{1.112}$$

Chapter 2

Magnetic Resonance Imaging

The aim of MR imaging is not only to establish the presence of different nuclei, but also to determine the spatial distribution of a given species within a sample. Therefore, we aim to correlate a series of signal measurements with the spatial locations of the various sources. When all protons are represented by just one chemical species such as water, then the above spectroscopic analysis simply gives the total signal from all spins without their spatial location in the static field, as long as that field is uniform. Here lies the reason why MRI becomes possible only through the introduction of the gradient fields. In fact they enable us to produce a signal with spatially varying frequency components. In this chapter we briefly explain how to employ these gradient fields to transform the measured signal into an image. Moreover, we describe the main MR image properties, emphasizing how they depend on the imaging parameters. Finally, we provide some examples of MRI sequences.

2.1 From the Magnetic Resonance Signal to the Image

2.1.1 Signal and Effective Spin Density

The imaging of a body corresponds to the determination of the spin distribution, rather than the magnetization. The discussions that will follow are valid for data sampling taking place over times small compared to T_2^*. Now we consider the complex signal given by the eq. (1.112) and we make some assumptions. The transmitting and receiving RF coils are considered to be sufficiently uniform, so that ϕ_0, $\theta_{\mathcal{B}}$ and $\mathcal{B}_\perp$ are all independent of position. It is now introduced the constant Λ, which includes the gain factors from the electronic detection system. Moreover we neglect the relaxation effects and we take $\Omega = \omega_0$, therefore we obtain:

$$s(t) = \omega_0 \Lambda \mathcal{B}_\perp \int d^3r M_\perp(\vec{r}, 0) e^{i(\Omega t + \phi(\vec{r}, t))}. \tag{2.1}$$

The angle ϕ is the accumulated phase with the counterclockwise positive sign convention:

$$\phi(\vec{r}, t) = -\int_0^t dt' \omega(\vec{r}, t'), \tag{2.2}$$

where $\omega = \omega_0$ only if there is a uniform static field. Now we suppose that a perfect $\pi/2$-pulse has been applied uniformly over the sample. Since we are neglecting the relaxation effects, the initial transverse magnetization is equal to its equilibrium value M_0. As we have already seen in eq. (1.60), the equilibrium magnetization can be expressed in terms of $\rho_0(\vec{r})$, the number of proton spins per unit of volume:

$$M_\perp(\vec{r}, 0) = M_0(\vec{r}) = \frac{1}{4} \rho_0(\vec{r}) \frac{\gamma^2 \hbar^2}{kT} B_0. \tag{2.3}$$

Using eqs. (2.1) and (2.3), it is possible to express the signal in terms of the effective spin density $\rho(\vec{r})$:

$$s(t) = \int d^3r \rho(\vec{r}) e^{i(\Omega t + \phi(\vec{r},t))}, \tag{2.4}$$

$$\rho(\vec{r}) \equiv \omega_0 \Lambda \mathcal{B}_\perp M_0(\vec{r}) = \frac{1}{4}\omega_0 \Lambda \mathcal{B}_\perp \rho_0(\vec{r}) \frac{\gamma^2 \hbar^2}{kT} B_0. \tag{2.5}$$

2.1.2 The Imaging Equation

Now we suppose the presence of three time-dependent gradient fields:

$$\vec{G}(t) = G_x(t)\hat{x} + G_y(t)\hat{y} + G_z(t)\hat{z}, \tag{2.6}$$

where $\hat{x}$, $\hat{y}$ and $\hat{z}$ do not necessary correspond to the transverse and longitudinal directions of the magnetization components. Defining the three implicitly time-dependent components of $\vec{k}$ as follows:

$$k_x(t) = \bar{\gamma} \int^t G_x(t')dt', \qquad k_y(t) = \bar{\gamma} \int^t G_y(t')dt', \qquad k_z(t) = \bar{\gamma} \int^t G_z(t')dt', \tag{2.7}$$

where the integrations run from the onset of the gradient to time t. The use of each of these gradients has the object to establish a connection between the position of spins along some direction and their precessional rates, also called frequency encoding along that direction. In fact the Larmor frequencies acquire a spatial dependence:

$$\omega(\vec{r}, t) = \omega_0 + \omega_G(\vec{r}, t), \tag{2.8}$$

where $\omega_G(\vec{r}, t) = \gamma \vec{r} \cdot \vec{G}(t)$. Therefore we can write:

$$\omega(\vec{r}, t) = \gamma(B_0 + \vec{r} \cdot \vec{G}(t)). \tag{2.9}$$

In the hypothesis $\Omega = \omega_0$ ($\delta\omega = 0$), the eq. (2.4) becomes:

$$s(t) = \int d^3r \rho(\vec{r}) e^{i\phi(\vec{r},t)} = \int d^3r \rho(\vec{r}) e^{-i \int^t dt' \omega_G(\vec{r},t')} = \int d^3r \rho(\vec{r}) e^{-i\gamma\vec{r} \cdot \int^t dt' \vec{G}(t')}. \tag{2.10}$$

Finally, substituting the eq. (2.7) in the eq. (2.10), we can write the *3D-imaging equation*:

$$s(\vec{k}) = \int d^3r \rho(\vec{r}) e^{-i2\pi \vec{k} \cdot \vec{r}}. \tag{2.11}$$

This expression shows that the signal $s(\vec{k})$ is the Fourier transform of the spin density of the sample. Therefore, the spin density of the sample can be found by taking the inverse Fourier transform of the signal:

$$\rho(\vec{r}) = \int d^3k s(\vec{k}) e^{i2\pi \vec{k} \cdot \vec{r}}. \tag{2.12}$$

In general we have a set of discrete data, $s_m(\vec{k})$, from which we would like to reconstruct the image $\hat{\rho}(\vec{r})$, that represents an accurate estimate of the physical density $\rho(\vec{r})$. Since our measurements consists of a discrete set, what is practically used is the discrete inverse Fourier transform.

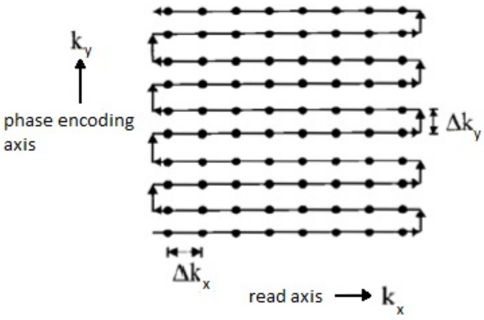


Figure 2.1: The k-space coverage for a 2D example. Each dot represents a sampled point. Lines of connected dots along the read direction are referred to data collected during the same read period [7].

2.1.3 The 2D Representation

In this section we want to describe how to obtain a 2D representation, for example of a slice of a body. It is necessary to sample only k_x and k_y. The 2D k-space coverage is realized using two gradient fields, G_x and G_y, the latter responsible of the phase encoding, the former used as read gradient. Each set of points along a given k_y is phase encoded as $\gamma G_y \tau_y y$. This y-dependent phase contribution remains unchanged during data sampling along the k_x-axis. A standard coverage of 2D k-space is based on a sequence that employs a single RF pulse to sample the xy plane with parallel lines as shown in figure 2.1. The gradients along $\hat{x}$ and $\hat{y}$ move the position in k-space from the origin to the bottom left corner in figure 2.1. This 2D k-space sampling configuration is made possible by alternately turning on the G_x gradient and then the G_y gradient. The read gradient G_x is reversed after each application of G_y and the phase encoding steps ΔG_y are applied when the read gradient is turned off. While the read gradient is turned on data are sampled, instead data are not collected during the short vertical steps. In this configuration, all information along y at a given x is projected on to x. For example, for the line $k_y = 0$, using the eq. (2.11) we can write:

$$s(k_x, 0) = \int d^3r \rho(\vec{r}) e^{-i2\pi k_x x}. \tag{2.13}$$

Now, calculating the inverse Fourier transform of this line of data, we can obtain the 1D projection reconstructed image:

$$\begin{aligned}\hat{\rho}_{1D}(x) &= \int dk_x s(k_x, 0) e^{i2\pi k_x x} = \int d^3r' \rho(\vec{r}') \int dk_x e^{-i2\pi k_x (x-x')} \\ &= \int\int\int dx' dy' dz' \rho(x', y', z') \delta(x - x') = \int\int dy dz \rho(x, y, z).\end{aligned} \tag{2.14}$$

The generalization of the encoding to an arbitrary value of k_y follows:

$$s(k_x, k_y) = \int\int\int dx dy dz \rho(x, y, z) e^{-i2\pi(k_x x + k_y y)}. \tag{2.15}$$

Finally, after a calculation similar to eq. (2.14), the 2D image projected along $\hat{z}$ is:

$$\hat{\rho}(x, y) = \int dk_x s(k_x, k_y) e^{i2\pi(k_x x + k_y y)} = \int dz \rho(x, y, z). \tag{2.16}$$

2.1.4 3D coverage of k-space

The above discussion about the coverage of 2D k-space can be generalized to the 3D case by considering a series of parallel planes. The set can be obtained in 3D k-space by the application of two phase encoding gradients, one along k_y and the other along k_z. The imaging of three dimensions by the use of two phase encoding directions perpendicular to the read direction is called '3D imaging', to contrast it with 'multi-slice 2D imaging'. The difference between the two methods interests the RF excitation. In the latter approach, a series of RF pulses defines the set of slices filling out the third dimension. Instead, in the former approach, a RF pulse may be used to excite a thicker slice which is then phase encoded in the slice select gradient direction.

Now we want to describe how to employ the gradients to obtain a discretized coverage in 3D imaging. The x-gradient defines the read direction of the set of lines. During the continuous application of G_z along this direction, it is possible to carry out a set of measurements at finite time steps Δt. The associated step in the k_x direction is:

$$\Delta k_x = \bar{\gamma} G_x \Delta t. \tag{2.17}$$

The orthogonal gradients, G_x and G_y, are turned off during the read sampling, in order to detain each line parallel to the x-axis. before the read data are taken, the (k_y, k_z) position of each line is determined by applying the orthogonal gradients for times τ_y and τ_z. After a given line has been sampled, an adjacent parallel line is considered by turning the orthogonal gradients back on with the same amplitudes ΔG_y and ΔG_z and for the same times. The corresponding shifts in k-space are:

$$\Delta k_y = \bar{\gamma} \Delta G_y \tau_y, \tag{2.18}$$

$$\Delta k_z = \bar{\gamma} \Delta G_z \tau_z. \tag{2.19}$$

It is rare to sample all k-space points after a single RF excitation. Only a small number of lines of k-space can be collected after each RF excitation before the signal is lost due to T_2 or T_2^* decay. Generally, after a single RF pulse, it is possible to sample only one line of data. After a new RF excitation, the phase encoding gradients are increased with step sizes ΔG_y and ΔG_z, followed by the acquisition of another line of k-space data. This process is repeated every T_R (time between two successive RF pulses) until all of the necessary k-space data are acquired. The total time for a 3D imaging method is given by:

$$T_{\text{acq}} = N_y N_z T_R. \tag{2.20}$$

The plane or set of lines in 2D or 3D imaging could be replaced, in principle, by an arbitrary trajectory through the k-space region of interest as long as the same points are sampled. A rich variety of RF pulses and gradient combinations can be used to achieve the same coverage of the k-space. Finally, in 3D imaging, the total imaging time is given by:

$$T_T = N_{\text{acq}} N_y N_z T_R. \tag{2.21}$$

2.1.5 Slice Selection

In general, magnetic resonance images are generated by exciting a single thin slice of the body by using a combination of gradient fields and spatially selective RF pulses. Firstly we have to choose the direction perpendicular to the plane of the desired slice, called 'slice select axis', and the gradient along this direction is defined as the 'slice select gradient'. Choosing the z-axis as slice select axis leads to a *transverse* slice of the body. Instead, if we select the y-axis, the slice obtained is called *coronal*. Finally, if the x-axis is chosen, the

Table 2.1: The relationship between the slice select gradient used and the slice plane obtained applying it. The nomenclature has an anatomical basis [7].

Applied slice select gradient	Name	Slice plane orientation
G_x	sagittal	parallel to yz plane
G_y	coronal	parallel to xz plane
G_z	transverse	parallel to xy plane

slice is said to be *sagittal*. This nomenclature is shown in figure 2.2 and summarized in table 2.1. The presence of a slice select gradient causes the precessional frequency to be linearly dependent on the position along the corresponding slice select axis. In MR measurements it is usually chosen the z-axis as slice select direction. Therefore the frequency at position z is:

$$f(z) = f_0 + \gamma z G_z, \tag{2.22}$$

where $f_0 = \gamma B_0$ is the Larmor frequency at $z = 0$. The object is to excite uniformly a slice, defined by the coordinate z, such that all the spins contained here have identical phase and flip angle after slice selection. Since the frequency spread of a realistic RF pulse is bandlimited, a region of finite thickness along the z-axis would have its spins tipped, while spins outside this region would ideally remain aligned with B_0. Therefore, the exciting of an infinitesimal slice, using an RF pulse, represents only an ideal case. Now we consider the more realistic case in which we want to uniformly excite a slice of finite thickness extending from $z_0 - \Delta z/2$ to $z_0 + \Delta z/2$. The RF pulse should have a frequency profile, in the rotating frame, which is unitary inside the range of frequencies $(\gamma G_z z_0 - \gamma G_z \Delta z/2, \gamma G_z z_0 + \gamma G_z \Delta z/2)$ and equal to zero outside. The bandwidth BW_{RF} of the RF pulse is given by:

$$BW_{RF} \equiv \Delta f = (\gamma G_z z_0 + \gamma G_z \Delta z/2) - (\gamma G_z z_0 - \gamma G_z \Delta z/2) = \gamma G_z \Delta z. \tag{2.23}$$

Hence, thanks to the presence of the gradient G_z, exists a range of frequencies which can be excited to create a transverse magnetization in a slice with thickness Δz, orthogonal to the z-axis. It is useful to introduce a new notation for the slice thickness:

$$\Delta z \equiv TH. \tag{2.24}$$

As just seen, we can express the slice thickness as a function of the bandwidth of the RF pulse an of the applied gradient:

$$TH = \frac{BW_{RF}}{\gamma G_z}. \tag{2.25}$$

Since we desire a uniform flip angle across the selected slice, the RF excitation profile as a function of frequency has to be a rect function ($rect(f/\Delta f)$) of bandwidth Δf. This necessity imposes that the temporal envelope of the RF pulse ($B_1(t)$), which is the inverse Fourier transform of the frequency profile, would be a sinc function:

$$B_1(t) \propto sinc(\pi \Delta f t). \tag{2.26}$$

The sinc envelope corresponds to the amplitude modulation of the RF oscillations in the laboratory frame. The center frequency of the excited bandwidth is $\gamma G_z z_0$ in the Larmor rotating frame and $f_0 + \gamma G_z z_0$ in the laboratory frame. We have to remark that, since a sinc function is defined from $-\infty$ to $+\infty$, it is necessary to cut off the temporal envelope, therefore in the frequency domain we will not obtain a prefect rect function.

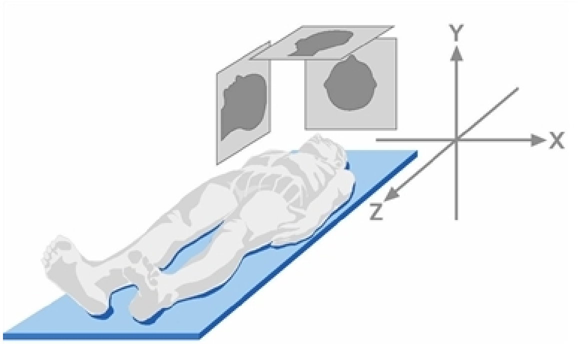


Figure 2.2: The orientation of the orthogonal slice planes correlated with the standard axis of a whole-body MRI magnet system. The subject is usually placed head first into the magnet and carried in by a sliding gantry on the patient table. The feet point along the $+\hat{z}$, left shoulder along $+\hat{x}$, and nose along $+\hat{y}$ when the person is supine.

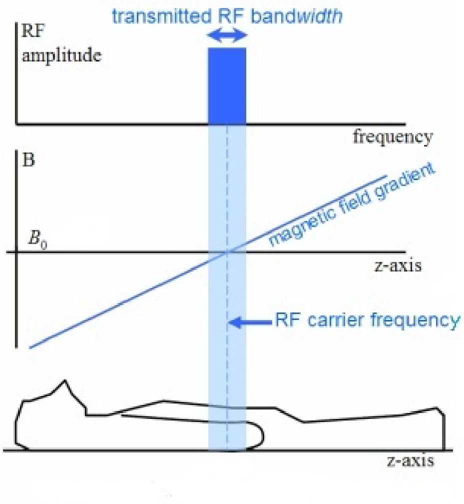


Figure 2.3: From the top: the RF excitation profile in the frequency domain and the application of the slice select gradient along z-axis.

2.2 Image properties

2.2.1 Signal and Noise

The ability to understand if some object is within a voxel or not depends critically on whether the signal for the object in that voxel can be distinguished from noise. The signal-to-noise ratio (SNR) is an essential parameter that characterizes the degree to which noise affects a measurement. In MRI, if the SNR is not high enough, it becomes impossible to discriminate a tissue from another one or from the background.

As just seen in the precedent sections, the data are assumed to be sampled at a set of points in 3D k-space. These points are separated by Δk_x, Δk_y and Δk_z, respectively, in the three orthogonal k-space directions. It is useful to define three sets of integers that run over the number points sampled in each direction of the image space:

$$p' \in (-N_x, N_x - 1) \qquad q' \in (-N_y, N_y - 1) \qquad r' \in (-N_z, N_z - 1). \tag{2.27}$$

The image is reconstructed by calculating the discrete inverse Fourier transform of the set of data represented by $s_m(p'\Delta k_x, q'\Delta k_y, r'\Delta k_z)$, where $s(\vec{k})$ has already been defined in eq. (2.11).

The image is represented by $\hat{\rho}_m(p\Delta x, q\Delta y, r\Delta z)$, where the integers p, q and r typically cover the same range as p', q' and r':

$$\hat{\rho}_m(p\Delta x, q\Delta y, r\Delta z) = \frac{1}{N_x N_y N_z} \sum_{p',q',r'} s(p'\Delta k_x, q'\Delta k_y, r'\Delta k_z) e^{i2\pi\left(\frac{pp'}{N_x} + \frac{qq'}{N_y} + \frac{rr'}{N_z}\right)}. \tag{2.28}$$

The last equation represents the effective spin density and it is also known as the 'voxel signal' $S(\vec{r})$, since it is the signal which will be represented in the volume element $\Delta x \Delta y \Delta z$ at position $(p\Delta x, q\Delta y, r\Delta z)$ within the reconstructed image. It is assumed that voxels are small volumes characterized by uniform spin densities. Therefore, we will neglect the possibility that a voxel contains more than one tissue type. As we have already seen in eq. (1.111), the peak signal for homogeneous object is $\omega_0 M_0 \mathcal{B}_\perp V_{\text{sample}}$, which, for proton imaging and using the eq. (1.60), becomes:

$$S \equiv \hat{\rho}_m(p\Delta x, q\Delta y, r\Delta z) \propto \frac{\gamma^3 \hbar^2}{4kT} B_0^2 \mathcal{B}_\perp(p\Delta x, q\Delta y, r\Delta z) V_{\text{voxel}}. \tag{2.29}$$

Therefore, the voxel signal is directly related to $\rho\mathcal{B}_\perp$ and to B_0^2, in addition it shows an inverse dependence on the sample temperature.

One of the principal objects in an MRI experiment is to achieve enough voxel signal relative to noise to observe tissues of interest. The measured signal is affected by random thermal fluctuations, which are called 'white' fluctuations because they are characterized by equal expected noise power components at all frequencies within the readout bandwidth (BW_{read}):

$$\sigma_{\text{thermal}} = \sqrt{4kTRBW_{\text{read}}}, \tag{2.30}$$

R is the effective resistance of the coil loaded by the body, and BW_{read} is the bandwidth of the noise-voltage detecting system, which can be expressed as follows:

$$BW_{\text{read}} = \frac{1}{\Delta t} = \gamma G_x L_x, \tag{2.31}$$

where Δt is the sampling time step and L_x is the Field of View (FOV) along the read gradient field direction. Now we want to give an expression for the SNR, but not before having noticed that repeating an entire imaging experiment N_{acq} times and averaging the signal over these measurements improve the SNR. In fact this parameter betters as the square root of the number of acquisitions if the noise is uncorrelated from one experiment to the next. Therefore, we can write:

$$SNR/\text{voxel}(p\Delta x, q\Delta y, r\Delta z) = \sqrt{N_{\text{acq}}} \frac{\frac{\gamma^3\hbar^2}{4kT} B_0^2 \mathcal{B}_\perp(p\Delta x, q\Delta y, r\Delta z) V_{\text{voxel}}}{\sqrt{4kTRBW_{\text{read}}}}. \tag{2.32}$$

Finally, we aim to remark the SNR dependencies on imaging parameters, which are under experimenter control:

$$SNR/\text{voxel} \propto \sqrt{N_{acq}} \frac{V_{\text{voxel}}\sqrt{N_x N_y N_z}}{\sqrt{BW_{\text{read}}}} \propto \Delta x \Delta y \Delta z \sqrt{N_{acq} N_y N_z T_s}. \tag{2.33}$$

To estimate the last we have employed the identity $T_s = N_x \Delta t$, where T_s is the time necessary to sample one line of the k-space. Moreover we have used the relation:

$$\sigma_0^2(p\Delta x) = \frac{\sigma_m^2}{N}, \tag{2.34}$$

between the measured variance σ_m^2 of any point in k-space and the variance σ_0^2 in the image domain. We have taken N as the number of samples collected along k_x direction. The variance measured in any voxel in the image space is N times smaller than in the detected signal and is the same for all voxels. From the eq. (2.33) it is evident that the experimenter may augment or decrease the SNR adjusting the number of voxels, the number of acquisitions or the voxel dimensions. Nevertheless $N_x \Delta t$ cannot be arbitrarily changed because of relaxation phenomena.

Since often the acquisition is repeated to augment the SNR, it is convenient to introduce another parameter: the imaging efficiency Υ. This represents the SNR/voxel normalized to the square root of the total imaging time (see eq. (2.21)). Therefore, for a fixed T_R, the imaging efficiency is defined as:

$$\Upsilon \equiv \frac{(SNR/\text{voxel})_{3D}}{\sqrt{T_T}} \propto \frac{\Delta x \Delta y \Delta z \sqrt{N_{acq} N_y N_z T_s}}{\sqrt{N_{\text{acq}} N_y N_z}} \propto \Delta x \Delta y \Delta z \sqrt{T_s}. \tag{2.35}$$

In other words, an image with better spatial resolution, such as with Δy and Δz halved, is considered only half as SNR-efficient as an image acquired with voxel size Δy or Δz, respectively. The main conclusion is that high resolution can be achieved efficiently only in the read direction, and even that is limited by the need to keep T_s on the order of T_2^* or less.

2.2.2 Contrast, Contrast-to-Noise Ratio and Visibility

Even the highest SNR does not guarantee a useful image. An important object of imaging for diagnostic purposes is to be capable of distinguish between diseased and neighboring normal tissues. This problem falls under the category of the 'signal detection' problem and requires an understanding of the importance of Contrast-to-Noise Ratio (CNR).

The common approach to the contrast concept is to examine the absolute difference in the signal between two tissues of interest. If we label these two tissues with A and B, respectively characterized by voxel signals S_A and S_B, their signal difference is defined as the 'contrast':

$$C_{AB} \equiv S_A - S_B. \tag{2.36}$$

Although the contrast may be large enough to distinguish them, if noise is substantial, the signal difference would not be detectable. Therefore, we can introduce a more appropriate quantity, that is the contrast-to-noise ratio (CNR):

$$\text{CNR}_{AB} \equiv \frac{C_{AB}}{\sigma_0} = \frac{S_A - S_B}{\sigma_0} = \text{SNR}_A - \text{SNR}_B, \tag{2.37}$$

where the noise in each voxel of the image, σ_0, is assumed to be the same.
If multiple independent signal acquisitions N_{acq} are effectuated, as already seen in the in the precedent sections, an average of these signal measurements leads to a decrease of the effective noise standard deviation:

$$\sigma_{\text{eff}} = \frac{\sigma_0}{\sqrt{N_{\text{acq}}}}. \tag{2.38}$$

In an image where the tissue A occupies n_{voxel} voxels, each of which has signal S_A with independent additive noise with standard deviation σ_0, we can introduce a new measure called 'object visibility':

$$\nu_{AB} \equiv \frac{C_{AB}}{\sigma_{\text{eff}}} = \frac{C_{AB}}{\sigma_0}\sqrt{n_{\text{voxel}}} = \text{CNR}_{AB}\sqrt{n_{\text{voxel}}}. \tag{2.39}$$

2.2.3 Contrast Mechanisms in Magnetic Resonance Imaging

MRI has the advantage to elaborate the tissue signal in many ways, leading to numerous contrast mechanism. This huge flexibility originates from the presence of lots of imaging and tissue parameters. The most basic contrast generating mechanisms are based on spin density, and T_1 and T_2 differences between tissues. Additional examples are magnetic susceptibility differences, magnetization transfer contrast, tissue saturation methods, contrast enhancing agents and diffusion. For the purpose of understanding the basic concepts regarding the three main methods, we write the expression of the contrast between tissues A and B for a 90° flip angle gradient echo experiment:

$$\begin{aligned} C_{AB} &= S_A(T_E) - S_B(T_E) = \\ &= \rho_{0,A}(1 - e^{-T_R/T_{1,A}})e^{-T_E/T^*_{2,A}} - \rho_{0,B}(1 - e^{-T_R/T_{1,B}})e^{-T_E/T^*_{2,B}}. \end{aligned} \tag{2.40}$$

Where the signal is assumed to be determined by the tissue signal solution from the Bloch equation at the echo time T_E. C_{AB} can be maximized with respect to either T_R or T_E.
We will not describe the methods that have just been mentioned, but we will give a brief description of one of the main contrast mechanism: T_1-weighted contrast. This choice is justified by the fact that we have employed T_1-weighted images as input data in the analyses which will be detailed described in the next chapters.

T_1-Weighting

Normal soft tissues T_1 values are quite different from one another. Therefore, T_1-weighted contrast offers a very powerful method to reveal different tissues. For T_1 weighting, only the effects of T_2 can be minimized, because the effects of spin density differences cannot be neglected. Using the gradient echo example as before, we can notice that the choice of a very short T_E, with respect to $T^*_{2,A}$ and $T^*_{2,B}$, reduces any T^*_2 contrast, in fact we have:

$$T_E \ll T^*_{2,A} \qquad \Longrightarrow \qquad e^{-T_E/T^*_{2,A}} \approx 1, \tag{2.41}$$

$$T_E \ll T^*_{2,B} \qquad \Longrightarrow \qquad e^{-T_E/T^*_{2,B}} \approx 1. \tag{2.42}$$

Therefore, the eq. (2.40) becomes:

$$\begin{aligned} C_{AB} \approx& \rho_{0,A}(1 - e^{-T_R/T_{1,A}}) - \rho_{0,B}(1 - e^{-T_R/T_{1,B}}) = \\ &= (\rho_{0,A} - \rho_{0,B}) - (\rho_{0,A}e^{-T_R/T_{1,A}} - \rho_{0,B}e^{-T_R/T_{1,B}}). \end{aligned} \tag{2.43}$$

To optimize the T_1-weighted contrast, the eq. (2.43) has to be maximized with respect to T_R. Differentiating C_{AB} with respect to T_R and setting the derivative equal to zero, we obtain the following:

$$\frac{\partial C_{AB}}{\partial T_R} = 0 \implies \frac{\rho_{0,A} e^{-T_R/T_{1,A}}}{T_{1,A}} = \frac{\rho_{0,B} e^{-T_R/T_{1,B}}}{T_{1,B}}, \tag{2.44}$$

$$\implies T_{R_{\mathrm{opt}}} = \frac{\ln \frac{\rho_{0,B}}{T_{1,B}} - \ln \frac{\rho_{0,A}}{T_{1,A}}}{\frac{1}{T_{1,B}} - \frac{1}{T_{1,A}}}. \tag{2.45}$$

Some *a priori* knowledge of tissue properties are evidently very useful. In fact, when several tissues are present, it may be difficult to choose a single T_R which optimizes at the same time all contrasts.

2.3 Magnetic Resonance Imaging sequences

An MRI sequence is represented by the combination of RF pulses and gradients. Particular choices lead to different image contrasts and different k-space sampling. Today, really a large number of sequences are employed, but the classical ones are: Spin Echo and Inversion Recovery. The latter makes possible a T_1 measurement, instead the former method is used to effectuate a T_2 measurement. It will follows a panoramic explanation of these two main sequences.

2.3.1 Spin Echo

The spin echo method was for the first time introduced by Hahn in 1950 [12], but it is still very used. The spin echo sequence is based on the application of two types of RF pulses (see the sequence diagram in figure 2.4): a $\pi/2$-pulse followed by a π one, said also 'refocusing' pulse. In the first step, we can assume that the magnetization of a sample is instantly tipped by the $\pi/2$-pulse into the transverse plane. Then, at $t = 0$, the spins at different positions begin to dephase, relative to each other, as they experience different field strengths, each of which, in general, is not exactly equal to B_0. The dephasing effect takes place because of the T_2^*, which represents a combination of external field induced (T_2') and thermodynamic (T_2) effects. After a time interval $T_E/2$, the sample is excited by another RF pulse, but of the π type. Therefore, the spins which had previously accumulated extra positive phase now have, at the instant after the π-pulse application, the negative of that phase, and vice versa. Hence, the phase differences between spins gradually decrease and will vanish at $t = T_E$, where T_E is called echo time. The echo signal $s(T_E)$ will be characterized by a decrease in magnitude because the loss of phase due to T_2 is not recoverable:

$$s(T_E) = s_0 e^{-T_E/T_2}. \tag{2.46}$$

After a time $T_E/2$ from the echo signal, we can apply another π-pulse and obtain a new echo signal at $t = 2T_E$. Up to now we have reconstructed two points of the T_2 decay curve, but we can continue to fill it with other measured points and finally fit them to estimate T_2 (figure 2.6).

Regarding the applied gradients, we can say that both the RF pulses are slice selective. As it is illustrated in figure 2.4, the read gradient is turned on during the echo and the phase encoding one is activated between the $\pi/2$ and the π pulses.

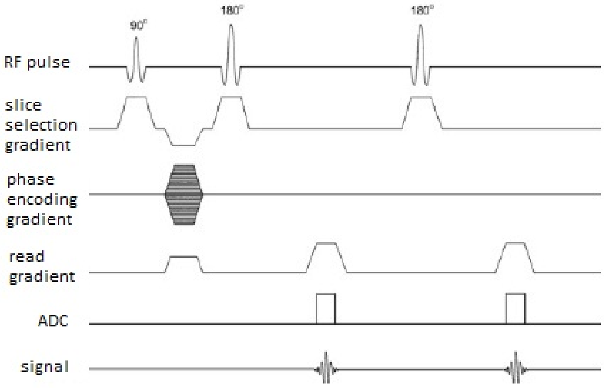


Figure 2.4: Schematic diagram showing the spin echo sequence.

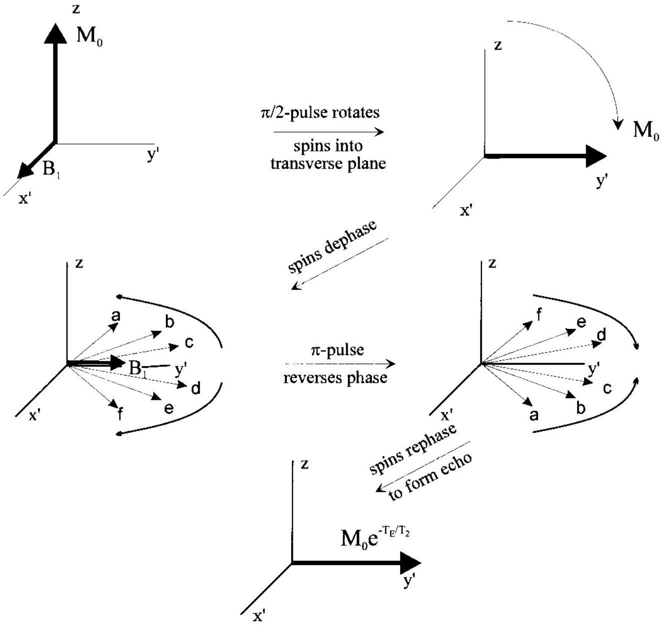


Figure 2.5: A simulation of an ensemble of spins in the rotating reference frame during a spin echo experiment [7].

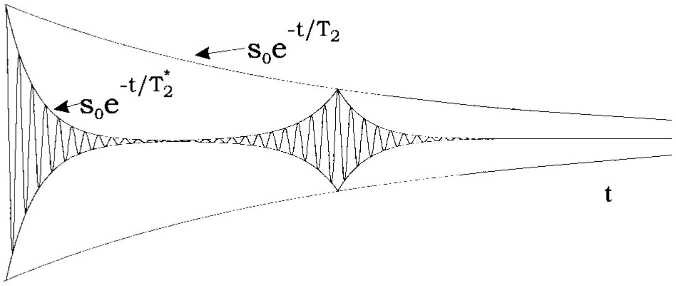


Figure 2.6: Decay envelopes in the laboratory frame, obtained according to the spin echo method [7].

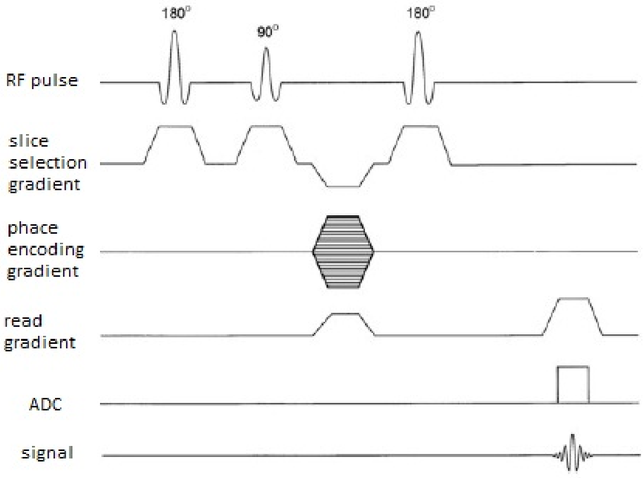


Figure 2.7: Schematic diagram showing the inversion recovery sequence.

2.3.2 Inversion Recovery

Spin echo experiments are useful for determining the T_2 properties of a sample. As single experiments, however, they are not sensitive to T_1. But, as repeated measurements, their signals must depend on T_1, and the spin-lattice parameters could, in line of principle, be determined by performing a number of repeated sequences. There is another experiment, called inversion recovery (figure 2.7), that is sensitive to T_1. This sequence becomes with the application of a π-pulse that inverts the longitudinal magnetization sign, therefore, we have:

$$M_z(0^+) = -M_0. \tag{2.47}$$

Hence, the magnetization becomes to relax towards its equilibrium value M_0 because of spin-lattice interaction, according to the following law:

$$M_z(t) = -M_0 e^{-t/T_1} + M_0(1 - e^{-t/T_1}) = M_0(1 - 2e^{-t/T_1}), \qquad 0 < t < T_I. \tag{2.48}$$

After, the longitudinal magnetization is tipped into the transverse plane to provide the initial signal, at $t = T_I$, employing a $\pi/2$-pulse. Hence, the magnitude of the transverse magnetization evolves as:

$$M_\perp(t) = |M_0(1 - 2e^{-T_I/T_1})| e^{-(t-T_I)/T_2^*}, \qquad t > T_I. \tag{2.49}$$

The $\pi/2$-pulse has converted the longitudinal magnetization into transverse magnetization whose precession can be detected as MR signal. This signal represents the beginning of a spin echo sequence. Repeating the experiment with different T_I it is possible to measure the T_1 relaxation time.

Chapter 3

Voxel-Based Morphometry

The purpose of the current chapter is to give the fundamentals of voxel-based morphometry (VBM), since that it represents the first technique that we have employed to analyse our T1-weighted images. The software used to do this is the SPM8 package (Statistical Parametric Mapping, Wellcome Department of Imaging Neuroscience, London, UK, `http://www.fil.ion.ucl.ac.uk/spm`), which allows for an easy implementation of the VBM protocol [13].

3.1 An introduction to morphometric methods

Morphometrics is about studying the variability of the form (size and shape) of organisms or objects [15]. A pioneer in this field was D'Arcy Thompson who first proposed the concept of spatial normalization, according to which diverse and dissimilar brains "can be referred as a whole to identical functions of very different coordinate systems" [16]. The brain of an individual subject can be modeled by a canonical brain, which is deformed, or warped, to a different shape. Within this model there is the assumption of a one-to-one mapping between the anatomy of one brain and that of another, and that is possible to transfer this configuration of homologous points between them. Morphometry, or computational anatomy, involves analysing features that characterize the shapes of the brains. A large number of approaches for characterizing differences in the shape and neuro-anatomical configuration of different brains have recently emerged due to improved resolution of anatomical human brain MRI scans (see the MRI scan of a healthy subject in figure 3.1) and the development of new advanced image processing techniques.

The traditional approach to morphometry involves manually measuring lengths, angles, areas, volumes and so on. However, the measures of interest are usually constituted by volumes. A number of interactive tools are available for manually making volume measurements from MRI data, and it is probably the most widely accepted form of morphometry among the medical community. Nevertheless, this traditional approach has some disadvantages, e.g. the excessive time-consuming of manual outlining of brain regions. Moreover, this procedure is subjective and quite a little reproducible. Once the measurements have been made, they usually would be analyzed in order to detect statistically significant differences between populations of subjects. Such analyses are often performed independently, region-by-region [15].

Studies of brain shape have been carried out by many researchers on a number of different populations, including patients with Alzheimer's disease, schizophrenia, autism, dyslexia and Turner's syndrome. Often, the morphometric measurements of these studies have been obtained from brain regions that can be clearly defined. These measures typically regard volumes of unambiguous structures such as the hippocampi. However, there are many morphometric features that may be more difficult to quantify by inspection. Therefore, it is necessary to involve new morphometric approaches that are not biased to one particular

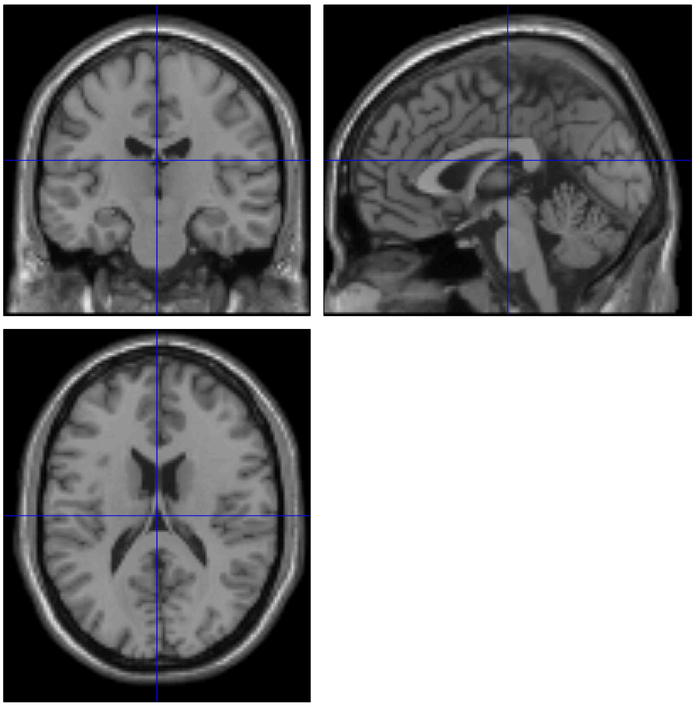


Figure 3.1: T1-weighted high resolution image of a healthy subject (voxel size 2x2x2 mm^3, image dimension 91x109x91).

structure or tissue and that give an even-handed and comprehensive assessment of anatomical differences throughout the brain [18]. In this context we can introduce another approach for comparing the volume of a tissue type among populations of subjects: voxel-based morphometry (VBM), the morphometric approach which will be explained in this chapter.

VBM is based on a voxel-wise comparison of the local concentration of grey matter between two groups of subjects. The procedure is relatively direct, and involves spatially normalizing high-resolution MR images, and smoothing these grey-matter segments. Voxel-wise parametric statistical tests are performed, which compare the smoothed grey-matter images from the groups. Corrections for multiple comparisons are employed using the theory of Gaussian random fields. VBM has the purpose of localizing significant structural differences among populations resulting in a statistical parametric map of regional differences. In the case of VBM, statistical parametric maps are derived from univariate data, where there is a single variable for each voxel, but there are other morphometric techniques that involve multivariate data, where there is a number of variables at each voxel. An example of such methods is tensor-based morphometry (TBM), in which the aim is to identify differences in the local shape of brain structure from the gradients of deformation fields, rather than differences in the local concentration of GM.

3.2 The Voxel-Based Morphometry Method

The steps that characterize the VBM method can be summarized as follows [13, 17, 18].

- The algorithm begins with the affine registration of all subjects' images to the same stereotaxic space. Therefore, the images become all referred to the same template. This is necessary because the segmentation procedure in SPM8 requires the images to be registered with the tissue probability maps [13]. The resulting images are then divided into different tissue classes according to the segmentation methods that are described in detail in [20, 21], and that we will try to explain in section 3.3. Many tissue classification methods produce images containing the *a posteriori* probability for each voxel to belong to a particular tissue class according to the model. Moreover, the segmentation step includes parameters that account for image intensity nonuniformity, which make possible to perform a bias correction.

- To increase the accuracy of inter-subject alignment the Diffeomorphic Anatomical Registration using Exponentiated Lie algebra (DARTEL) algorithm [26] can be implemented (see also section 3.4). It models the shape of each brain using millions of parameters (three parameters for each voxel), instead the simpler inter-subject registration model included in SPM5 used about 1000 coefficients to parametrize the shape of the brain. DARTEL works by aligning grey matter among the images, while simultaneously matching white matter. This is possible because of an iterative alignment between data and an average template data. In the current step a study-specific template is generated and the deformation fields that warp each segmented image to the average template are also estimated.

- The final template image obtained in the previous step is now affine registered to MNI[1] space. Employing the last transformation and the deformation fields, the segmented brain tissues are then spatially normalized to MNI space. This procedure Jacobian scales the data, with the aim of preserving the total amount of signal from each region in the images (‘modulation’). In this way the final VBM statistics will be reflective of the absolute amount of grey matter (GM) in different regions rather than the relative concentration of GM (the proportion of grey matter to other tissue types within a region). Finally, the resulting images are smoothed through the convolution with a Gaussian kernel. In this way, every voxel belonging to the smoothed images contains the average amount of GM from around the voxel. This is often called ‘gray matter density’. This step is necessary because the smoothing has the effect of rendering the data more normally distributed, thus augmenting the validity of parametric statistical tests [29].

- The main step of a VBM analysis is based on the voxel-wise statistical tests (see also section 3.5). The results consist of a statistical parametric map (SPM), showing significant regional differences among the populations involved in the study. Given that a SPM contains the results of many statistical tests, it is necessary to correct for multiple dependent comparisons, employing the theory of Gaussian random fields (GRF). The statistical tests are performed in the context of the general linear model (GLM). The GLM is a really flexible framework, that for example allows to apply statistical tests with the aim of identifying regions of GM concentration that are related to particular covariates such as disease severity or age. Standard tests are t-test and F-test, performed to test the hypothesis, in more accurate terms to reject with a certain probability a null hypothesis.

3.3 Segmentation

Healthy brain tissue can generally be classified into three broad tissue types on the basis of an MR image. These are grey matter (GM), white matter (WM) and cerebrospinal fluid (CSF). We could perform manually this classification on a good quality T1 image, by simply selecting appropriate image intensity ranges which include most of the voxel intensities of a

---

[1] Standard brains from the Montreal Neurological Institute (MNI). The MNI defines a new standard brain by using a large series of MRI scans on normal controls. It wants to define a brain that is more representative of the population.

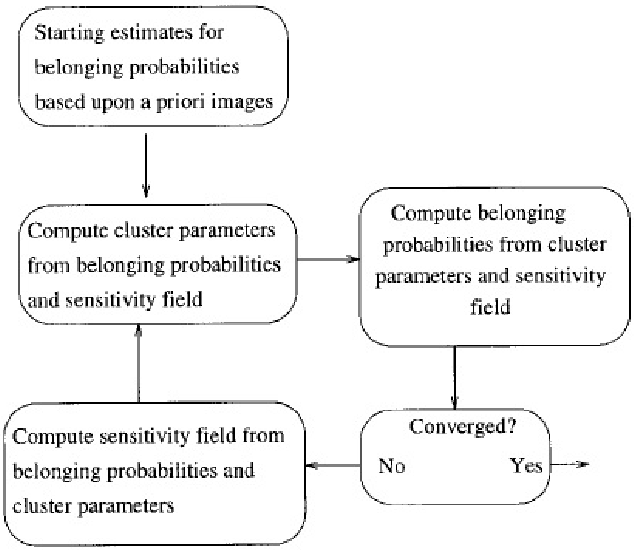


Figure 3.2: A flow diagram for the tissue classification [17].

particular tissue type. However, this manual selection of thresholds is extremely subjective. Hence, in morphometric studies automatic segmentation methods are performed.

The classical segmentation approach employs the tissue classification method. It requires the images to be registered with tissue classification maps [22], at least one for each brain tissue (GM, WM, CSF) and eventually others describing voxels outside the brain. After registration, these maps represent the prior probability of different tissue classes being found at each location in an image [24]. Bayes' rule can then be used to combine these priors with tissue type probabilities derived from voxel intensities to provide the posterior probability [21]. The disadvantage is that this procedure requires the knowledge of many parameters, and to estimate each of these it is necessary to know other parameters. Therefore, the tissue classification approach is circular, because the registration requires an initial tissue classification, and the tissue classification requires an initial registration (figure 3.2). This circularity is resolved by combining both components into a single generative model [21]. Here classification, registration and bias correction steps are unified, instead to be serially applied. This unified procedure is detailed described in the Unified segmentation paper [21], but we will sum up its main elements, because this procedure is employed by SPM8 software to perform tissues segmentation.

3.3.1 The Unified Segmentation framework

The Unified segmentation method [21] works within the mixture of Gaussians model (MOG), where the intensity distribution of each tissue class is represented by one or more gaussians. From this model it is possible to define an objective function that is minimized by optimum parameters. This objective function can be extended to model smooth intensity nonuniformity and to accommodate deformations of tissue probability maps.

The MOG model

The distribution of voxel intensities within a brain image can be modeled by a mixture of K Gaussians (clusters). In the most simple case each Gaussian distribution models one tissue class, but in more advanced parameterizations, more clusters represent one tissue type. This is a standard technique that is employed in many tissue classification algorithms [25]. For univariate data, the k-th Gaussian is characterized by three parameters: the mean cluster intensity μ_k, the standard deviation σ_k^2 and the mixing proportion γ_k, which represents the proportion of voxels that belong to that Gaussian. For the last parameter stands the following:

$$\sum_{k=1}^{K} \gamma_k = 1 \qquad \gamma_k \geq 0. \tag{3.1}$$

To segment our image we have to find the parameters that describe every cluster. Therefore, to fit this MOG model we have to maximize the probability of observing the I elements of data $\vec{y}$, given the parametrization of the Gaussians. The probability of observing a datum with intensity y_i given that it belongs to the k-th cluster is:

$$P(y_i|c_i = k, \mu_k, \sigma_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} exp\bigg(-\frac{(y_i - \mu_k)^2}{2\sigma_k^2}\bigg). \tag{3.2}$$

The prior probability that every voxel possesses of belonging to the k-th cluster, independently on its intensity is given by the mixing proportion, that represent the prior probability:

$$P(c_i = k|\gamma_k) = \gamma_k. \tag{3.3}$$

The joint probability of cluster k and intensity y_i is:

$$P(y_i, c_i = k|\mu_k, \sigma_k, \gamma_k) = P(y_i|c_i = k, \mu_k, \sigma_k)P(c_i = k|\gamma_k). \tag{3.4}$$

By integrating the last over the total number of clusters used in the model, we obtain:

$$P(y_i|\vec{\mu}, \vec{\sigma}, \vec{\gamma}) = \sum_{k=1}^{K} P(y_i, c_i = k|\mu_k, \sigma_k, \gamma_k). \tag{3.5}$$

Moreover, the probability of the entire dataset (image composed by I voxels) by assuming that all elements are independent is:

$$P(\vec{y}|\vec{\mu}, \vec{\sigma}, \vec{\gamma}) = \prod_{i=1}^{I} P(y_i|\vec{\mu}, \vec{\sigma}, \vec{\gamma}) = \prod_{i=1}^{I} \Bigg(\sum_{k=1}^{K} \frac{\gamma_k}{\sqrt{2\pi\sigma_k^2}} exp\bigg(-\frac{(y_i - \mu_k)^2}{2\sigma_k^2}\bigg)\Bigg). \tag{3.6}$$

The last expression is maximized when the following cost function (ϵ) is minimized:

$$\epsilon = -logP(\vec{y}|\vec{\mu}, \vec{\sigma}, \vec{\gamma}) = -\sum_{i=1}^{I} \Bigg(\sum_{k=1}^{K} \frac{\gamma_k}{\sqrt{2\pi\sigma_k^2}} exp\bigg(-\frac{(y_i - \mu_k)^2}{2\sigma_k^2}\bigg)\Bigg). \tag{3.7}$$

We have to remark the fact that the assumption that voxels are independent is hard to believe. Nevertheless the spatial priors will be modified to lead a higher degree of spatial dependency.

Intensity nonuniformity

MR images are usually corrupted by smooth, spatially varying artifacts that modulate the intensity of the image (bias). These nonuniformities, although not usually a problem for visual inspection, can impede automated processing of the images. Parametric bias correction models are often an integral part of tissue classification methods. Most algorithms assume that the bias is multiplicative. In this type of parametric model the observed signal (y_i) is considered to be an uncorrupted signal (μ_i), scaled by some bias (ρ_i) with added

Gaussian noise (n_i) that is independent on the bias. The noise source is assumed to be from the scanner itself:

$$y_i = \mu_i/\rho_i + n_i. \tag{3.8}$$

Another type of model expects that the noise is added before the signal is scaled. In this case, the noise is assumed to be due to variations of tissue properties.

$$y_i = (\mu_i + n_i)/\rho_i. \tag{3.9}$$

Furthermore a combination of the scanner and tissue noise models has been adopted in some studies. This would probably be a better model, especially for images corrupted by a large amount of bias. However, the single source model was mainly chosen for its simplicity.

The current method adopts a multiplicative noise model, which assumes that the errors originate from tissue variability rather than additive Gaussian noise from the scanner. Within the Unified segmentation framework, bias correction is included in the MOG model through extra parameters ($\vec{\beta}$) that account for smooth intensity variations. The field modelling the intensity variation at element i is denoted by $\rho_i(\vec{\beta})$. Therefore, intensity y_i that belongs to the kth cluster assumes the following expression:

$$\begin{aligned} P(y_i|c_i = k, \mu_k, \sigma_k, \vec{\beta}) &= \frac{1}{\sqrt{2\pi(\sigma_k/\rho_i(\vec{\beta}))^2}} exp\left(-\frac{(y_i - \mu_k/\rho_i(\vec{\beta}))^2}{2(\sigma_k/\rho_i(\vec{\beta}))^2}\right) = \\ &= \rho_i(\vec{\beta})\frac{1}{\sqrt{2\pi\sigma_k^2}} exp\left(-\frac{(\rho_i(\vec{\beta})y_i - \mu_k)^2}{2\sigma_k^2}\right), \end{aligned} \tag{3.10}$$

hence, the objective function becomes:

$$\epsilon = -\sum_{i=1}^{I} log\left(\rho_i(\vec{\beta})\sum_{k=1}^{K}\frac{\gamma_k}{\sqrt{2\pi\sigma_k^2}} exp\left(-\frac{(\rho_i(\vec{\beta})y_i - \mu_k)^2}{2\sigma_k^2}\right)\right). \tag{3.11}$$

The model adopted in this paper parameterises the bias field as the exponential of a linear combination of low frequency basis functions. Since the bias tends to be smooth, a small number of basis functions are employed.

Spatial priors

Prior probability maps are usually generated by registering a large number of subjects together, assigning voxels to different tissue types and averaging tissue classes over subjects. The default tissue probability maps are modified versions of the Independent Community Bankers of Minnesota (ICBM[2]) Tissue Probabilistic Atlases. The original data are represented by 452 T1-weighted scans, which were aligned with an atlas space, corrected for scan inhomogeneities, and classified into GM, WM and CSF. These data were affine registered to MNI space, and down-sampled to 2mm resolution [14]. The most recent version of the software (SPM8), the one that we will use, is characterized by an important improvement with respect to the paper we are referring [21]. It employs six tissue probability maps, which are displayed in figures 3.3-3.5. Therefore, the tissue classes that are used regard GM, WM, CSF, soft tissue and air/background. These maps give the prior probability of any voxel in a registered image belonging of each tissue class, irrespective of its intensity. We can modify

---

[2] The ICBM atlas is an average of T1-weighted MRIs of normal young adult brains. The space the atlas is in is not based on any single subject. Instead it is an average space constructed from the average position, orientation, scale, and shear from all the individual subjects. The atlas, therefore, is both an average of intensities and of spatial positioning (`http://www.loni.ucla.edu/`).

the model in eq. (3.11), instead of using stationary mixing proportions ($P(c_i = k|\gamma) = \gamma_k$), allowing the prior probabilities to vary over voxels, such that the prior probability of voxel i being drawn from the k-th Gaussian can be written as follows:

$$P(c_i = k|\vec{\gamma}) = \frac{\gamma_k b_{ik}}{\sum_{j=1}^{K} \gamma_j b_{ij}}, \tag{3.12}$$

where b_{ik} is the tissue probability for class k at voxel i. As we have already outlined, a tissue probability map can be represented by more than one cluster. The assumption of a single Gaussian distribution for each class does not hold because a voxel may be not purely of one tissue type and instead contain signal from different tissues (partial volume effects). Typical number of Gaussians could be two for grey matter, two for white matter, two for CSF, three for bone, four for other soft tissues and two for air/background.

Deformable spatial priors

According to the paper we are referring to [21], the model can be improved by allowing the tissue probability maps to be deformed through parameters $\vec{\alpha}$. This makes possible registration to a standard space to be included within the same generative model. The eq. (3.12) assumes the following form:

$$P(c_i = k|\vec{\gamma}, \vec{\alpha}) = \frac{\gamma_k b_{ik}(\vec{\alpha})}{\sum_{j=1}^{K} \gamma_j b_{ij}(\vec{\alpha})}. \tag{3.13}$$

After including the full priors, the objective function can be rewritten:

$$\epsilon = -\sum_{i=1}^{I} log\left(\frac{\rho_i(\vec{\beta})}{\sum_{k=1}^{K} \gamma_k b_{ik}(\vec{\alpha})} \sum_{k=1}^{K} \frac{\gamma_k b_{ik}(\vec{\alpha})}{\sqrt{2\pi\sigma_k^2}} exp\left(-\frac{(\rho_i(\vec{\beta}) y_i - \mu_k)^2}{2\sigma_k^2} \right) \right). \tag{3.14}$$

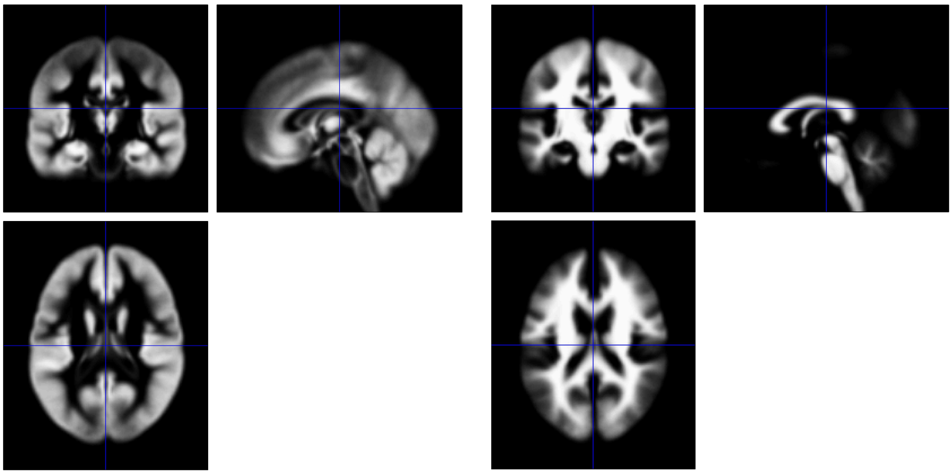
The current way of parameterize how the tissue probability maps could be deformed adopts a low-dimensional approach, which model the deformations by a linear combination of about a thousand of cosine transform bases [23].

Final overview

Since there is not a closed form solution for the minimization of the objective function shown in eq. (3.14), optimal values for different parameters are dependent on the values of others. The employed approach is called Iteration Conditional Modes (ICM) and it begins by starting estimates for the parameters and then iterating until a locally optimal solution is found. Each step of the iterative procedure involves alternating between estimating different groups of parameters, while holding the others fixed at their current best solution. Therefore, the adopted procedure is a local optimization, and the initial starting values are randomly assigned.

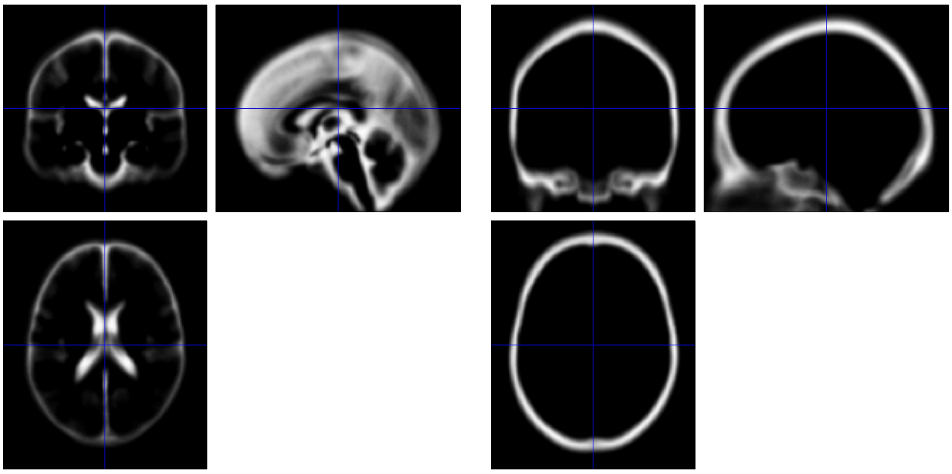
The current unified segmentation method was evaluated by J. Ashburner and K. J. Friston in [21], showing that the components of the generative model described above all contribute to segmentation performance, in at least one data context, improving the segmented output with respect to the previous segmentation approaches.

In conclusion, within the same probabilistic framework, the procedure described in [21] allows the unification of tissue classification, bias correction and nonlinear warping. This integration approach presents the disadvantage of being more complex and difficult to implement. In addition, implementation time is longer and more expertise is required. Nevertheless, this unified procedure is more accurate and makes better use of the information available in the data. An example of Unified segmentation output through the 'New Segment' tool of SPM8 package is shown in figure 3.6.



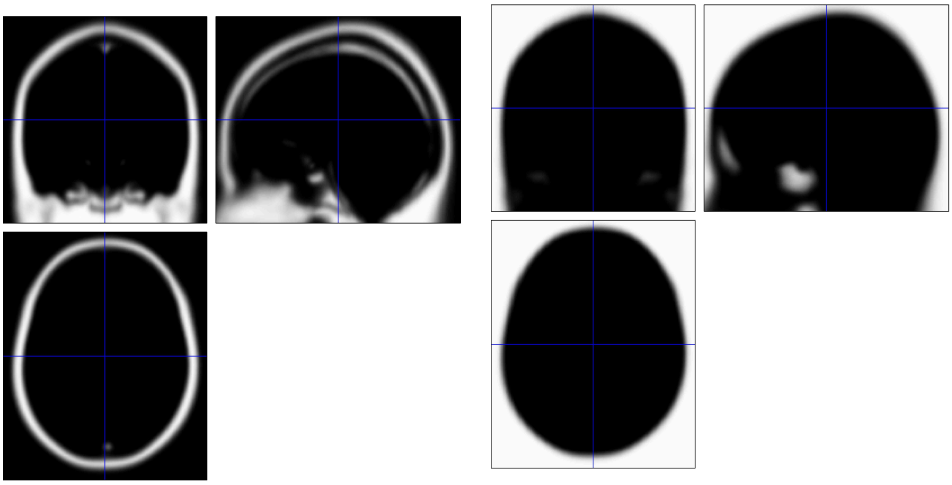
(a) Grey matter TPM. **(b)** White matter TPM.

Figure 3.3: Grey matter and white matter Tissue Probability Maps (TPMs) employed in SPM8 software.



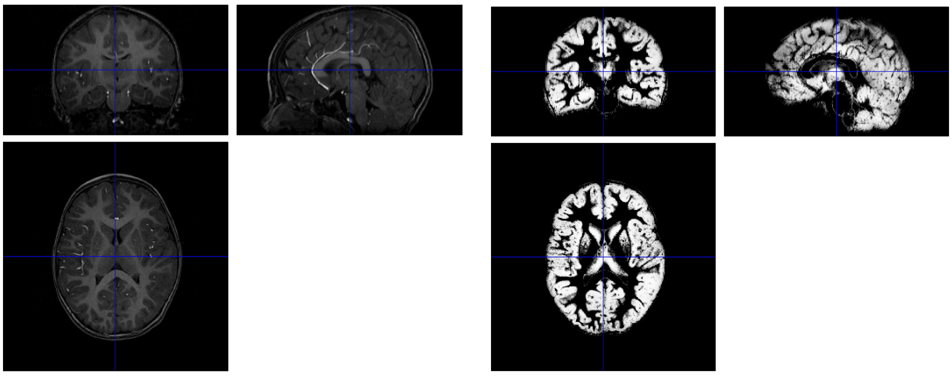
(a) CSF TPM. **(b)** Bone TPM.

Figure 3.4: CSF and bone Tissue Probability Maps (TPMs) employed in SPM8 software.

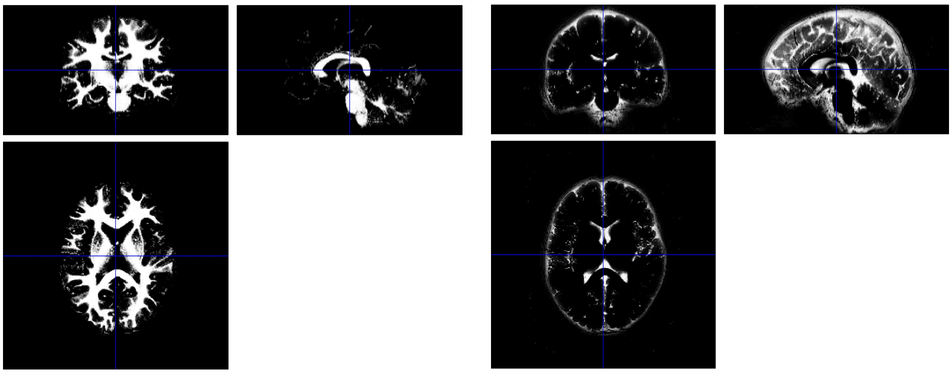


(a) Soft tissue TPM. **(b)** Air/background TPM.

Figure 3.5: Soft tissue and air/background Tissue Probability Maps (TPMs) employed in SPM8 software.



(a) Original T1-weighted image. **(b)** GM segmented image.



(c) WM segmented image. **(d)** CSF segmented image.

Figure 3.6: A T1-weighted image (3.6a) that has been segmented in three tissue types: grey matter (3.6b), white matter (3.6c) and cerebrospinal fluid (3.6d).

3.4 The DARTEL algorithm

Since VBM compares the whole brain volumes between two groups of subjects, achieving an accurate inter-subject registration seems to be a necessary request. For this reason the Diffeomorphic Anatomical Registration using Exponentiated Lie algebra (DARTEL) algorithm represents a fundamental step to process the segmentation output. This procedure implements a diffeomorphic warping of each GM segmented image to a study-specific template generated within the same algorithm. The SPM8 package offers a specific tool to implement at the same time the study-specific template generation and the diffeomorphic image registration. This procedure is extensively illustrated in the paper of Ashburner [26], of which we will remark the main elements.

Image registration is based on the estimation of a smooth continuous mapping between the points in one image and those in another. Then, the parameters characterizing this mapping can be used to determine the relative shapes of the images. The aim is to obtain the single best set of values for these parameters. We can distinguish between two ways of modelling such mappings:

- The small-deformation model does not necessarily preserve topology and parameterizes a displacement field ($\boldsymbol{u}$), which is simply added to an identity transform ($\boldsymbol{x}$):

  $$\boldsymbol{\Phi}(\boldsymbol{x}) = \boldsymbol{x} + \boldsymbol{u}(\boldsymbol{x}) \tag{3.15}$$

  In this case, the inverse transformation is sometimes approximated by subtracting the displacement, which fails for larger deformations.

- The large-deformation model generates deformations (diffeomorphisms) that possess the important property of enforcing the preservation of topology. A differomorphism represents a one-to-one smooth and continuous mapping with derivatives that are invertible (i.e. nonzero Jacobian determinant). If the mapping is not diffeomorphic, then topology is not necessarily preserved.

The DARTEL algorithm regards the latter framework, which shows the positive characteristic of consistency under composition of deformations. In fact, if these are diffeomorphic, the result of the composition operation will also be diffeomorphic. However, as deformations are represented discretely by a finite number of parameters, there may be some small violations. Instead, perfect (i.e. infinitely dimensional) diffeomorphisms form a Lie group under the composition operation, because they satisfy the requirements of closure, associativity, inverse and identity.

If $\boldsymbol{u}^{(t)}$ is a velocity field at time t, then the diffeomorphism Φ evolves as follows:

$$\frac{d\boldsymbol{\Phi}}{dt} = \boldsymbol{u}^{(t)}\big(\boldsymbol{\Phi}^{(t)}\big). \tag{3.16}$$

According to the paper we are referring to [26], the DARTEL procedure considers a single (flow) velocity field, which remains constant over unit time. Registration involves simultaneously minimising a measure between the image and the warped template, while also minimising an energy measure of the deformations used to warp the template. This energy can be computed by integrating the energy of the velocity fields over unit time. The fixed velocity field employed has to describe the whole trajectory of an evolving diffeomorphism, therefore it may be forced to assume complicated and high energy trajectories in order to achieve good correspondence between images. In Group theory, the flow field may be considered as a member of the *Lie algebra*, which is *exponentiated* to produce a doformation, which is a member of a *Lie group*. Because the Jacobian of a deformation that adapts to an exponentiated flow field is always positive, we can be sure that the mapping is diffeomorphic and, implicitly, that the forward and inverse transformations can be generated from the same flow field:

$$\boldsymbol{\Phi} = Exp(\boldsymbol{u}). \tag{3.17}$$

3.4.1 The DARTEL optimization procedure

This diffeomorphic image registration procedure uses a model which contains a number of unknown parameters that describe how the image is deformed to match with another. The optimization procedure recurs to a discrete parametrization of the velocity field, $\boldsymbol{u}(\boldsymbol{x})$, that can be written as a combination of basis functions:

$$\boldsymbol{u}(\boldsymbol{x}) = \sum_i \nu_i \boldsymbol{\rho}_i(\boldsymbol{x}), \tag{3.18}$$

where $\boldsymbol{\nu}$ is a vector of coefficients and $\boldsymbol{\rho}_i(\boldsymbol{x})$ is the i-th basis function at position $\boldsymbol{x}$. The purpose of the optimization procedure is to individuate the single best set of values that parameterize the velocity field. We can write the posterior probability of the parameters given the image data (D):

$$p(\boldsymbol{\nu}|D) = \frac{p(D|\boldsymbol{\nu})p(\boldsymbol{\nu})}{p(D)}. \tag{3.19}$$

that is proportional to the likelihood, i.e. the probability of the image data given the parameters ($p(D|\boldsymbol{\nu})$) and to the priors of the parameters ($p(\boldsymbol{\nu})$). The probability of the data is a constant ($p(D)$). The eq. (3.19) is the objective function has to be maximized to obtain a *maximum a posteriori* estimate for the parameters. Normally, the objective function is represented by the logarithm of the *a posteriori* probability (in which case it is maximized) or the negative logarithm (in which case it is minimized). It may be seen as the sum of two terms: a prior term and a likelihood term.

$$-\log p(\boldsymbol{\nu}|D) = \log p(\boldsymbol{\nu}) - \log p(D|\boldsymbol{\nu}), \tag{3.20}$$

or

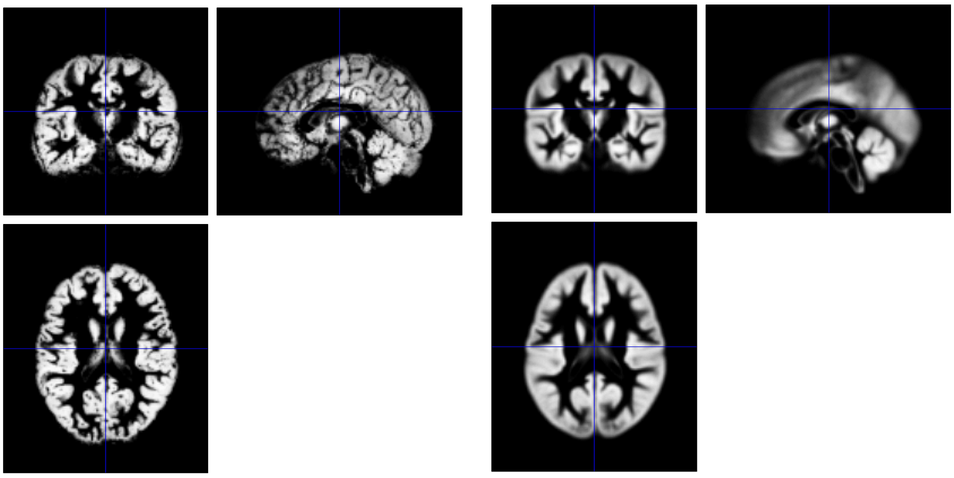
$$\epsilon(\boldsymbol{\nu}) = \epsilon_1(\boldsymbol{\nu}) + \epsilon_2(\boldsymbol{\nu}). \tag{3.21}$$

Nevertheless, in practice there are many technical difficulties that can impede a simple Bayesian interpretation of the problem. The procedure adopted to solve this problem consists in a local optimization, that is practically actuated through the Lavenberg-Marquardt (LM) algorithm [27]. The necessary matrix solutions are obtained in reasonable time using a multigrid method.

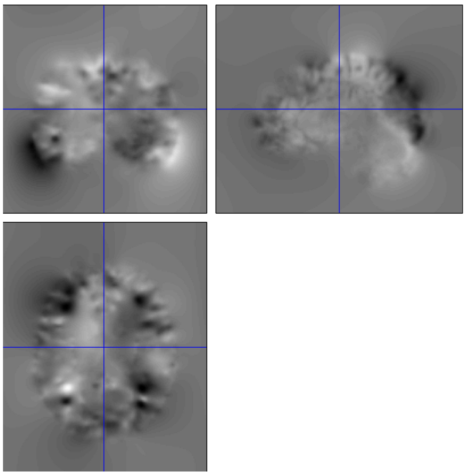
3.4.2 Group-wise registration

Until now, we have discussed about how to match a pair of images through diffeomorphic image registration. Nevertheless, our initial purpose was to warp together multiple subjects' images. For this reason, a study-specific template is usually generated and then the deformation fields that warp each image to this average template are estimated. The scheme that the DARTEL algorithm follows to construct the study-specific template involves an iterative procedure. Therefore, the DARTEL is used to map the scans to their average, and then the warped images are used to compute a new average template. This cycle is repeated until the spatial precision required has been achieved. An example of output images implementing DARTEL algorithm in SPM8 is shown in figure 3.7.

This procedure has been applied to intersubject registration of 471 whole brain images, and the resulting deformations were evaluated in terms of how well they encode the



(a) GM segmented and DARTEL-imported image. **(b)** Study-specific template.



(c) Flow field that warps the 3.7a into the 3.7b.

Figure 3.7: An example of flow field estimation through the DARTEL algorithm. The T1-weighted image of one subject has been GM segmented and then spatially normalized to a lower resolution space (i.e. the segmented image has been DARTEL-imported (3.7a)). The flow field that warps the latter to the study-specific template (3.7b), both computed through the DARTEL algorithm, is shown in 3.7c.

shape information necessary to separate male and female subjects and to predict the age of the subjects [26]. As explained in the latter paper, cross-validation was used to assess classification accuracy, showing a performance better than a small-deformation approach.

3.5 Voxel-wise statistical tests

Voxel-based morphometry is a technique that essentially compares segments of grey (or white) matter obtained from segmented MR images. VBM adopts the General Linear Model (GLM) to lay out parametric statistical tests to identify and make inferences about regionally specific differences. The output of the method is a statistic parametric map showing regions where GM density significantly differs among groups.
The statistical analysis involves two steps. Firstly, tests indicating evidence against a null hypothesis of no effect at each voxel are computed. An image representing these results is then produced. Secondly, this statistical image has to be assessed, well locating voxels where an effect has been detected, at the same time limiting the possibility of false positives

[28].
The current methods for evaluating the data are parametric. Therefore, the data are assumed to be distributed according to a specific form of probability density, and hypotheses expressed in terms of these distributions depending on unknown parameters are then tested. In this way, it is possible to compute the parameters and to form a statistic reflecting evidence against the null hypothesis. The majority of the statistical models represent special cases of the GLM. In this section we are determined to introduce the GLM with the purpose of explain the two-sample t-test framework. In fact, the last is one of the main statistical test that the VBM method employs in order to make voxel-wise comparisons of GM segments among two groups of subjects. In the following treatment of the GLM we will refer to [28].

3.5.1 An introduction to the General Linear Model

We now suppose to conduct an experiment in which we can measure a *response variable* Y_j, where $j = 1, ..., J$ indicates the observation. Y_j represents a random variable. Moreover, we will introduce for each observation a set of L ($L < J$) *explanatory variables* (each considered without error) denoted by x_{jl}, where $l = 1, ..., L$. The explanatory variables can represent covariates, function of covariates, or they may be *dummy variables* indicating the levels of an experimental factor.

A GLM models the response variable Y_j as a linear combination of L explanatory variables plus an error term ϵ_j:

$$Y_j = x_{j1}\beta_1 + ... + x_{jl}\beta_l + ... + x_{jL}\beta_L + \epsilon_j. \tag{3.22}$$

The β_l are the unknown parameters, corresponding to each of the L explanatory variables x_{jl}. The errors ϵ_j are supposed to be independent and identically distributed (iid) normal random variables with zero mean and variance σ^2, therefore we can write:

$$\epsilon_j \overset{iid}{\sim} \mathcal{N}(0, \sigma^2). \tag{3.23}$$

Linear models having other error distributions are called *Generalized Linear Models.*

Matrix formulation

It can be useful to introduce the matrix formulation of the GLM. We can write out the eq. (3.22) for each observation j, obtaining a set of J equation:

$$\begin{aligned} Y_1 &= x_{11}\beta_1 + ... + x_{1l}\beta_l + ... + x_{1L}\beta_L + \epsilon_1 \\ \vdots &= \vdots \\ Y_j &= x_{j1}\beta_1 + ... + x_{jl}\beta_l + ... + x_{jL}\beta_L + \epsilon_j \\ \vdots &= \vdots \\ Y_J &= x_{J1}\beta_1 + ... + x_{Jl}\beta_l + ... + x_{JL}\beta_L + \epsilon_J. \end{aligned} \tag{3.24}$$

We can also write the equivalent matrix form:

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_j \\ \vdots \\ Y_J \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1l} & \cdots & x_{1L} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{j1} & \cdots & x_{jl} & \cdots & x_{jL} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{J1} & \cdots & x_{Jl} & \cdots & x_{JL} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_l \\ \vdots \\ \beta_L \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_j \\ \vdots \\ \epsilon_J \end{pmatrix},$$

which can be written in the following way:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3.25}$$

where $\mathbf{Y}$ is the column vector of observations, $\boldsymbol{\epsilon}$ the column vector of error terms, and $\boldsymbol{\beta}$ the column vector of the unknown parameters. The $J \times L$ matrix is said to be the *design matrix*, which has one row per observation and one column per model parameter. Into the design matrix we can quantify our experimental knowledge about the expected signal.

Parameter estimation

Usually, the equations (3.24) implied by the general linear model (with $\epsilon = 0$) cannot be solved, because the number of parameters L is typically chosen to be less than the number of observations J. Hence, it becomes necessary to introduce a method of estimating parameters that best fit the data. Therefore, the method of *ordinary least squares* is employed.

We can indicate with $\tilde{\boldsymbol{\beta}}$ the set of estimated parameters, with $\tilde{\mathbf{Y}}$ the fitted values, and with $\mathbf{e}$ the residual errors:

$$\tilde{\boldsymbol{\beta}} = [\tilde{\beta}_1, ..., \tilde{\beta}_L]^T \tag{3.26}$$

$$\tilde{\mathbf{Y}} = [\tilde{Y}_1, ..., \tilde{Y}_J]^T = \mathbf{X}\tilde{\boldsymbol{\beta}} \tag{3.27}$$

$$\mathbf{e} = [e_1, ..., e_J]^T = \mathbf{Y} - \tilde{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}}. \tag{3.28}$$

Therefore, we can write the residual sum of squares $S = \sum_{j=1}^{J} e_j^2$, that measures the fit of the model with these parameter estimates. The least squares estimates are those parameter values which minimize the residual sum of squares:

$$S = \sum_{j=1}^{J} (Y_j - x_{j1}\tilde{\beta}_1 - ... - x_{jL}\tilde{\beta}_L)^2, \tag{3.29}$$

that is minimized in the following case:

$$\frac{\partial S}{\partial \tilde{\beta}_l} = 2\sum_{j=1}^{J} (-x_{jl})(Y_j - x_{j1}\tilde{\beta}_1 - ... - x_{jL}\tilde{\beta}_L) = 0. \tag{3.30}$$

The last represents the l-th row of the normal equations:

$$\mathbf{X}^T\mathbf{Y} = (\mathbf{X}^T\mathbf{X})\hat{\boldsymbol{\beta}}, \tag{3.31}$$

that are satisfied by the least squares estimates. If $(\mathbf{X}^T\mathbf{X})$ is invertible, i.e. if the design matrix is of full rank, then the least squares estimates are:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \tag{3.32}$$

Overdetermined models

If the design matrix $(\mathbf{X})$ is rank deficient, then $(\mathbf{X}^T\mathbf{X})$ is singular, and it is not invertible. Hence, the model is said to be overparameterised: there are infinitely many least squares estimates $\hat{\boldsymbol{\beta}}$ satisfying the normal equations. In this case it becomes necessary to impose constraints on the estimates, or by inverting $(\mathbf{X}^T\mathbf{X})$ using a pseudoinverse technique which effectively implies a constraint. Independently of the adopted method, it is important to remember that the estimates are related to the particular constraint or pseudoinverse method chosen. Therefore, when the statistical results have to be used to make inferences, it is meaningful to consider functions of the parameters that are uninfluenced by the particular constraint chosen.

SPM adopts a the pseudoinverse function implemented in MatLab that gives the Moore-Penrose pseudoinverse. Hence, if $(\mathbf{X}^T\mathbf{X})^-$ denotes the pseudoinverse of $(\mathbf{X}^T\mathbf{X})$, we can use $(\mathbf{X}^T\mathbf{X})^-$ in the eq. (3.32) , obtaining:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{Y} = \mathbf{X}^-\mathbf{Y}. \tag{3.33}$$

Inference

Now we want to derive the t-statistic which can be used to test for a linear combination of effects (contrasts).

The residual variance σ^2 can be estimated by the residual sum of squares divided by the number of degrees of freedom:

$$\hat{\sigma}^2 = \frac{\mathbf{e}^T\mathbf{e}}{J-p} \sim \sigma^2\frac{\chi^2_{J-p}}{J-p}, \tag{3.34}$$

where $p = \mathrm{rank}(\mathbf{X})$. A derivation of this result is reported in [28]. If $\mathbf{X}$ is full rank then it is possible to show that the parameter estimates are normally distributed:

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}\big(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\big). \tag{3.35}$$

From the last follows that for a column vector $\mathbf{c}$ containing L weights:

$$\mathbf{c}^T\hat{\boldsymbol{\beta}} \sim \mathcal{N}\big(\mathbf{c}^T\boldsymbol{\beta}, \sigma^2\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}\big). \tag{3.36}$$

Furthermore, $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are independent. Prespecified hypotheses regarding linear combinations of the model parameters $\mathbf{c}^T\boldsymbol{\beta}$ can be assessed using a t_{J-p} Student's t-distribution with $J-p$ degrees of freedom:

$$\frac{\mathbf{c}^T\hat{\boldsymbol{\beta}} - \mathbf{c}^T\boldsymbol{\beta}}{\sqrt{\hat{\sigma}^2\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}}} \sim t_{J-p}. \tag{3.37}$$

For example, if our initial hypothesis is represented by $\mathcal{H} : \mathbf{c}^T\boldsymbol{\beta} = d$, it can be assessed by evaluating the following:

$$T = \frac{\mathbf{c}^T\hat{\boldsymbol{\beta}} - d}{\sqrt{\hat{\sigma}^2\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}}}, \tag{3.38}$$

and computing a p-value by comparing T with a t-distribution with $J-p$ degrees of freedom.

If the model is overparameterized, then there are infinitely many parameter sets describing the same model. Therefore, when examining linear compounds $\mathbf{c}^T\boldsymbol{\beta}$ of the parameters we are forced to consider only those linear combinations that are invariant over the space of possible parameters. In [28] are thoroughly mathematically characterized such contrasts, in order to determine whether a linear compound is a proper contrast or not.

3.5.2 Two sample t-test

The two sample t-test is a particular case of the general linear model. We begin from two independent groups of random variables Y_{j1} and Y_{j2}. The current test assumes:

$$Y_{qj} \overset{iid}{\sim} \mathcal{N}(\mu_q, \sigma^2) \qquad q = 1, 2, \tag{3.39}$$

$$\mathcal{H} : \mu_1 = \mu_2, \tag{3.40}$$

where $\mathcal{H}$ is the null hypothesis and q labels the two groups. The index j refers to the data points in both groups. The standard statistical way of writing the model is:

$$Y_{qj} = \mu_q + \epsilon_{qj}. \tag{3.41}$$

Moreover, according to the general linear model assumptions, $\epsilon_{qj} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. The eq. (3.41) can be manipulated to obtain an expression closer to the eq. (3.22):

$$Y_{qj} = x_{qj1}\mu_1 + x_{qj2}\mu_2 + \epsilon_{qj}. \tag{3.42}$$

Here the dummy variables indicate group membership:

$$x_{qj1} = \begin{cases} 1 & \text{if } q = 1 \\ 0 & \text{if } q = 2 \end{cases} \qquad x_{qj2} = \begin{cases} 0 & \text{if } q = 1 \\ 1 & \text{if } q = 2 \end{cases}. \tag{3.43}$$

This model leads to a design matrix $\mathbf{X}$ with two columns of dummy variables indicating group membership and parameter vector $\boldsymbol{\beta} = [\mu_1, \mu_2]^T$. Therefore, the null hypothesis in eq. (3.40) is equivalent to $\mathcal{H} = \mathbf{c}^T\beta = 0$ with $\mathbf{c} = [1, -1]^T$. The first column of the design matrix contains n_1 1's and n_2 0's, indicating the measurements from the first group, while the second column contains n_1 0's and n_2 1's for the second group. Thus:

$$\begin{aligned} (\mathbf{X}^T\mathbf{X}) = \begin{pmatrix} n_1 & 0 \\ 0 & n_2 \end{pmatrix} \qquad &\Longrightarrow \qquad (\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} 1/n_1 & 0 \\ 0 & 1/n_2 \end{pmatrix} \\ \Longrightarrow \qquad &\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c} = \frac{1}{n_1} + \frac{1}{n_2}. \end{aligned} \tag{3.44}$$

Now, using the eq. (3.37), we obtain the t-statistic:

$$T = \frac{\hat{\mu_1} - \hat{\mu_2}}{\sqrt{\hat{\sigma}^2(1/n_1 + 1/n_2)}}, \tag{3.45}$$

which is the standard formula for the two-sample t-statistic, with a Student's distribution of $n_1 + n_2 - 2$ degrees of freedom under the null hypothesis.

This statistical model can be refined by adding one or more covariates. If we suppose of having k covariates, in this case we have to add k columns in the design matrix, one for each added covariate. Moreover, the number of degrees of freedom decreases becoming $n_1 + n_2 - 2 - k$.

In our practical case, initial data will be composed by two groups of normalized modulated and smoothed GM image segments, resulting from the VBM pre-processing. Therefore, the situation is more complicated with respect to the framework described above. In group studies we have many voxels and therefore many statistic values. We have to calculate a statistic for each brain voxel that tests for the effect of interest in that voxel. When we have to decide if a brain volume shows any evidence of the effect, we have to take into account that there are many thousands of statistic values. This is the multiple comparison problem in neuroimaging. Random field theory is a recent branch of mathematics that can be used to solve this problem [30].

3.5.3 The Family Wise Error rate correction

Usually we calculate a statistic with the purpose of deciding whether this represents convincing evidence of the effect we are interested in. Within the classical statistical inference framework, we test the statistic against the null hypothesis, which corresponds to the hypothesis that there is no effect. If the statistic is not compatible with the null hypothesis, we can conclude that there is an effect. In order to make a test against the null hypothesis, we can compare our statistic value to a null distribution, which is the distribution of statistic values if there is no effect. Using the null distribution, we can evaluate how probably it is that our statistic could have come about chance. For example if we find that our result has a 5% chance of resulting from a null distribution, we decide to reject the null hypothesis and to accept the alternative hypothesis that there is an effect. Nevertheless, in rejecting the null hypothesis, we must accept a 5% chance that we take that we are wrong when we reject the null hypothesis. Therefore, 5% represents our expected *type I* error rate.

Since in our studies we have many voxels and therefore many statistic values, our hypothesis refers to the whole volume of statistics in the brain. Evidence against the null hypothesis would be that the entire observed volume of values is unlikely have originated from a null distribution. Hence, the question we are asking is about the family of voxel statistics, and the risk of error that we are prepared to accept is the Family Wise Error rate (FWE), which represents the likelihood that the family of voxel values could have arisen by chance [30].

The FWE correction method requires that we find a threshold to apply to every statistic value, so that any values above the threshold are unlikely arisen by chance. It presents the advantage that if we find voxels above threshold, we can conclude that there is an effect at these voxel location, therefore the test possesses localizing power.

A threshold that can control family-wise error has to take into account the number of tests. In fact, if for example we consider a single t statistic value from a null distribution of 40 degrees of freedom, it has a probability of 1% of being greater than 2.42. We now imagine an experiment of 1000 t values with 40 degrees of freedom. If we look at each statistic, then by chance it will have a 1% probability of being greater of 2.42. This means that we would expect 10 t values in our sample of 1000 to be greater than 2.42. If we find one or more t values above 2.42 in this family of test, it is not a good evidence against the family-wise null hypothesis, which is that all these values have been drown from a null distribution. Therefore, we have to compute a new threshold, such that a family of 1000 t statistic values, there is 1% probability of there being one or more t values above this threshold.

Some neuroimaging analyses use the Bonferroni correction to calculate FWE rates, which can be easily computed by employing simple probability rules and that considers statistic values to be independent one from each other [30]. However, most MRI data have some degree of spatial correlation, therefore there are fewer independent values in the statistic volume than there are voxels. It is very common to smooth the images before statistical analysis and, when we apply to our MRI data a smoothing kernel such as a Gaussian, each value in the image is replaced with a weighted average of itself and its neighbours. Smoothing has the effect of blurring the image, and reduces the number of independent observations.

Family-Wise Error rate estimation through Random Field Theory

Random field theory (RFT) can be used to solve the problem of finding the height threshold for a smooth statistical map which gives the required family-wise error rate. The application of RFT proceeds in stages.

- Firstly, the smoothness (spatial correlation) of the statistical map has to be estimated.
- Then, the smoothness values are inserted in the RFT equation, to give the expected Euler characteristic (EC) at different thresholds.
- Finally, it is possible to calculate the threshold at which we would expect 5% of equivalent statistical maps arising under the null hypothesis to contain at least one area above threshold.

If we assume that the data were independent before smoothing, we can take the smoothness as entirely resulting from the smoothing we have applied. Therefore, the smoothing can be expressed as the width of the Gaussian kernel. We can use its FWHM to calculate the number of *resels* in the image. 'Resel' is a term introduced by Worsley [31] to measure the number of 'resolution elements' in a statistical map. This can be considered to be similar to the number of independent observations. A resel is defined as a block of values (in our situation, pixels) having the same size as the FWHM. For example, in a 2D image smoothed with a Gaussian kernel of 10x10 pixels, there are 100 resels.

The Euler characteristic (EC) can be taken as the number of blobs in an image after thresholding. Therefore, it is a property of an image after it has been thresholded. The expected EC (E[EC]), approximately corresponds to the probability of finding an above threshold blob in our statistic image. Therefore, the probability of a family-wise error is approximately equivalent to the E[EC], i.e. $P^{FWE} \approx E[EC]$. Now, if we know the number of resels in our image, we can compute the expected EC at any given threshold. For a 2D image, its formula is given by Worsley [31]. If we indicate with R the number of resels, and with Z_t the Z score threshold, we can write:

$$E[EC] = R(4log_e 2)(2\pi)^{-\frac{3}{2}} Z_t e^{-\frac{1}{2}Z_t^2}. \tag{3.46}$$

Analyses of functional imaging lead to three dimensional statistical images. The same principles of RFT apply in three dimensions, even if the equation for E[EC] is different, it still depends only on the resels of the image. For simplicity, we have only considered a random field of Z scores, i.e. numbers drawn from a normal distribution. Moreover, there are now equivalent results for t, F and χ^2 random fields. For example, SPM8 software uses formulae for t and F random fields to calculate corrected thresholds. Since it is not exactly correct the assumption according to which the smoothness is the same as any explicit smoothing we have applied, algorithms that can calculate the smoothness from the images themselves have to be applied. In practice, smoothness is evaluated using the residual values from the statistical analysis as described in [32] and [33].

There are two assumptions underlying RFT. Firstly, the error fields have to represent a reasonable lattice approximation to an underlying random field with a multivariate Gaussian distribution. Secondly, it is assumed that these fields are continuous. If the data have been sufficiently smoothed and the general linear model correctly specified then the RFT assumptions will be met and the RFT threshold is more accurate than the Bonferroni [30].

Chapter 4

Support Vector Machines

Multivariate approaches are taking place as alternative methods to standard univariate analyses such as VBM, in order to be capable of detecting subtle and spatially distributed differences in brain anatomy that are difficult to appreciate using conventional mass-univariate methods. In contrast to the last, pattern recognition techniques take into account specific inter-regional dependencies, using these information to classify data into one or more classes, for example patients and controls.

In this chapter describe the Support Vector Machines (SVMs) analytic approach [34], a multivariate technique whose importance is actually increasing in investigating about structural brain abnormalities in patients with respect to controls. Moreover, the SVM pattern recognition technique allows to investigate on the predictive capability of MRI scan, instead VBM does not provide predictive value, limiting its diagnostic impact [39].

Since in our study we have implemented linear kernel SVMs, we will principally focus on the theoretical principals their are based on and we will outline the optimization techniques adopted by Joachims and implemented within the SVM-Light software package (`http://www-ai.cs.uni-dortmund.de/svm_light`) [38].

The SVM classification performance will be evaluated in terms of the area under the receiver operating characteristic curve (AUC) [35], obtained in a cross-validation framework. These procedures will be described in the second part of the chapter.

4.1 Linear Support Vector Machines

4.1.1 The separable case

We can begin talking about the simplest case that regards linear machines trained on separable data. In the machine learning classification framework, each image is considered as a point in a high dimensional space, which dimension is equal to the number of voxels in the image (d). We will face with the simplest task of classifying the images into two classes, for example patients and controls [36].

Each image can be seen as a vector of d elements ($\mathbf{x}_i \in \mathbf{R}^d$) labelled with $y_i \in \{-1, 1\}$, where $i = 1, ..., l$ and l represents the number of images composing the sample. The classification problem can be viewed as a task of finding a separating hyperplane that divides the positive from the negative examples.

The classification procedure consists of two phases: training and testing. Firstly we have to divide the dataset into two groups: the training set and the test set. During the training, the algorithm estimates a hyperplane that separates the examples contained in the training set according to their labels. We can indicate each provided example with $\{\mathbf{x}_i, y_i\}$. The decision function that has been learned from the training data then can be used during the testing phase to predict the class of a new test example.

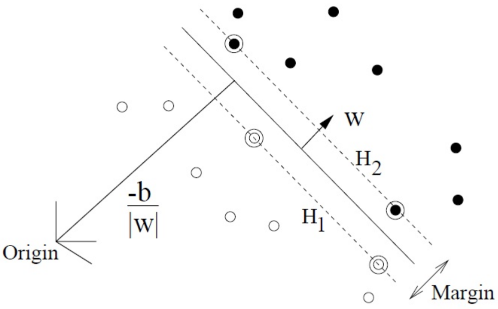


Figure 4.1: Linear separating hyperplanes for the separable case. The support vectors are circled [36].

We now suppose of having some hyperplane which separates the positive from the negative examples. The points $\mathbf{x}$ which lie on the separating hyperplane satisfy the following:

$$\mathbf{w} \cdot \mathbf{x} + b = 0, \tag{4.1}$$

where $\mathbf{w}$ is normal to the hyperplane and $|b|/\|\mathbf{w}\|$ represents the distance from the hyperplane to the origin. Let d_+ (d_-) be the shortest distance from the separating hyperplane to the closest positive (negative) training point. We can also define the 'margin' of a separating hyperplane to be $d_+ + d_-$. The support vector algorithm searches for the separating hyperplane characterized by the largest margin.

Suppose now that all the training data satisfy the following relations:

$$\mathbf{x}_i \cdot \mathbf{w} + b \geq +1 \qquad \text{for } y_i = +1, \tag{4.2}$$

$$\mathbf{x}_i \cdot \mathbf{w} + b \leq -1 \qquad \text{for } y_i = -1, \tag{4.3}$$

that can be summarized:

$$y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 \geq 0 \qquad \forall i. \tag{4.4}$$

The points which satisfy the eq. (4.2) lie on the hyperplane $H_1 : \mathbf{x}_i \cdot \mathbf{w} + b = +1$ with normal $\mathbf{w}$ and distance from the origin $|1 - b|/\|\mathbf{w}\|$. The points for which the eq. (4.3) holds lie on the hyperplane $H_2 : \mathbf{x}_i \cdot \mathbf{w} + b = -1$ with normal $\mathbf{w}$ and perpendicular distance from the origin $|-1 - b|/\|\mathbf{w}\|$. Therefore $d_+ = d_- = 1/\|\mathbf{w}\|$ and the margin is simply equal to $2/\|\mathbf{w}\|$. Moreover H_1 and H_2 are parallel and no training points fall between them. Thus we have to find the couple of hyperplanes which maximize the margin, therefore we have to minimize $\|\mathbf{w}\|^2$ subject to constraints in eq. (4.4). The solution for a typical two dimensional case will have the form shown in figure 4.1. Those training points which lie on one of the hyperplanes H_1, H_2, and whose elimination would change the final solution computed, are called support vectors. These are indicated in figure 4.1 by extra circles.

We can switch to a Lagrangian formulation of the problem, in fact the constraints in eq. (4.4) can be interpreted as constraints on the Lagrangian multipliers. This Lagrangian formulation is the one that is used in the actual training and testing algorithms, in fact it can be easily generalized to the nonlinear case.

We now introduce positive Lagrangian multipliers α_i, where $i = 1, ..., l$, one for each inequality constraint in the eq. (4.4). In order to build the Lagrangian function, we have to subtract the constraints equations multiplied for the α_i from the objective function, the one that we want to minimize. Therefore it gives the Lagrangian:

$$L_P \equiv \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{l} \alpha_i y_i(\mathbf{x}_i \cdot \mathbf{w} + b) + \sum_{i=1}^{l} \alpha_i. \tag{4.5}$$

To solve our problem we have to minimize L_P with respect to $\mathbf{w}$, b, and simultaneously imposing that the first derivatives of the Lagrangian with respect of all the α_i vanish, all subject to the constraints $\alpha_i \geq 0$ (we now call this particular constraints $\mathcal{C}_1$). This represents a convex quadratic programming problem and those points which satisfy the constraints also form a convex set. According to Fletcher [41], this means that we can equivalently solve the 'dual' problem: maximize L_P, subject to the constraints that the gradient of L_P with respect to both $\mathbf{w}$ and b vanish, and also to the constraints $\alpha_i \geq 0$ (we can call these particular constraints $\mathcal{C}_2$). This dual formulation of the problem is called Wolfe dual and possesses the property that the maximum of L_P, subject to constraints $\mathcal{C}_2$ occurs for the same values of $\mathbf{w}$, b and α_i for which L_P is minimized, subject to the constraints $\mathcal{C}_1$.

Minimize the Lagrangian with respect of $\mathbf{w}$ and b leads to the conditions:

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i, \tag{4.6}$$

$$\sum_i \alpha_i y_i = 0. \tag{4.7}$$

Because these are equality constraints in the dual formulation, we can substitute them into the eq. (4.5), obtaining:

$$L_D = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_i y_i y_j \mathbf{x}_i \mathbf{x}_j, \tag{4.8}$$

where P indicates primal and D means dual. Support vector training takes place by maximizing L_D with respect to the α_i subject to constraints in eq. (4.7) and positivity of the α_i, leading to the solutions reported in eq. (4.6). There is a Lagrange multiplier for every training point. In the solution, those points for which $\alpha_i > 0$ are called support vectors, and lie on one of the hyperplanes H_1 or H_2. All other training points have $\alpha_i = 0$ and lie on H_1 or H_2, or on that side of H_1 or H_2, such that the strict inequality in eq. (4.4) holds. In these case the support vectors are the critical elements of the training set. Hence, if all other training points were removed (or moved around, but so not to cross H_1 or H_2), and the training algorithm was repeated, the same separating hyperplane would be found.

4.1.2 The Karush-Kuhn-Turker Conditions

The Karush-Kuhn-Tucker (KKT) conditions play a central role in the theory of constrained optimization. For the minimization problem of L_P formulated above, the KKT conditions are [41]:

$$\frac{\partial}{\partial w_\nu} L_P = w_\nu - \sum_i \alpha_i y_i x_{i\nu} = 0 \qquad \nu = 1, ..., d, \tag{4.9}$$

$$\frac{\partial}{\partial b} L_P = - \sum_i \alpha_i y_i = 0, \tag{4.10}$$

$$y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 \geq 0 \qquad i = 1, ..., l, \tag{4.11}$$

$$\alpha_i \geq 0 \qquad \forall i, \tag{4.12}$$

$$\alpha_i(y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1) = 0 \qquad \forall i. \tag{4.13}$$

The last conditions are satisfied at the solution of any constrained optimization problem (convex or not) [36]. The problem for SVMs is convex, and for convex problems, the KKT conditions are necessary and sufficient for $\mathbf{w}$, b, α to be a solution [41]. Therefore, solving the SVM problem is equivalent to finding a solution to the KKT conditions. This fact results in many approaches to finding the solution [38]. Finding the solution for real world problems usually requires numerical methods, in fact are really rare the cases where the solution can be found analytically. The support vector optimization problem can be solved analytically only when the number of training data is very small, or for the separable case when it is known beforehand which of the training data become support vectors [36].

4.1.3 The non-separable case

The above algorithm for separable data, when applied to non-separable data, will find no possible solution, in fact the Lagrangian shows a growing arbitrarily large [36]. In this case is introduced a further cost through positive slack variables ξ_i, $i = 1, ..., l$, added in the constraints in eqs. (4.2) and (4.3) [34]:

$$\mathbf{x}_i \cdot \mathbf{w} + b \geq +1 - \xi_i \qquad \text{for } y_i = +1, \tag{4.14}$$

$$\mathbf{x}_i \cdot \mathbf{w} + b \geq -1 + \xi_i \qquad \text{for } y_i = -1, \tag{4.15}$$

$$\xi_i \geq 0 \qquad \forall i. \tag{4.16}$$

An error occurs when ξ_i exceeds unity, hence $\sum_i \xi_i$ is an upper bound on the number of training errors. Therefore we can assign an extra cost for errors by changing the objective function to be minimized from $\|\mathbf{w}\|^2/2$ to $\|\mathbf{w}\|^2/2 + C(\sum_i \xi_i)^k$, where C is a parameter to be chosen by the user, and a larger C corresponds to assigning a higher penalty to errors. This is a convex programming problem for any positive integer k. Formulating the Wolfe dual problem [36, 41] for the choice $k = 1$, it is possible to show that the only difference from the optimal plane case is that the α_i now have an upper bound of C. In this case the Wolf problem becomes maximize:

$$L_D \equiv \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \mathbf{x}_j, \tag{4.17}$$

with respect to the following:

$$0 \leq \alpha_i \leq C, \tag{4.18}$$

$$\sum_i \alpha_i y_i = 0. \tag{4.19}$$

The solution is given by:

$$\mathbf{w} = \sum_{i=1}^{N_S} \alpha_i y_i \mathbf{x}_i, \tag{4.20}$$

where N_S is the number of support vectors. The current situation is displayed is figure 4.2.

The primal Lagrangian can be written according to its definition:

$$L_P = \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_i \xi_i - \sum_i \alpha_i \{y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 + \xi_i\} - \sum_i \mu_i \xi_i, \tag{4.21}$$

where the μ_i are Lagrange multipliers introduced to enforce positivity of the ξ_i. The KKT conditions for this problem are ($i = 1, ..., l$ and $\nu = 1, ..., d$) [36]:

$$\frac{\partial L_P}{\partial w_\nu} = w_\nu - \sum_i \alpha_i y_i x_{i\nu} = 0, \tag{4.22}$$

$$\frac{\partial L_P}{\partial b} = -\sum_i \alpha_i y_i = 0, \tag{4.23}$$

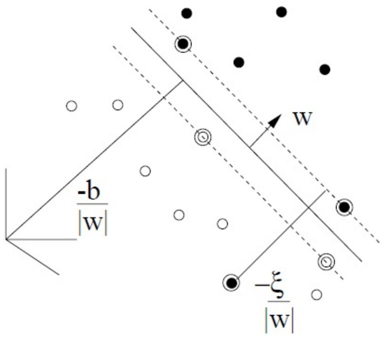


Figure 4.2: Linear separating hyperplanes for the non-separable case. The support vectors are circled. [36].

$$\frac{\partial L_P}{\partial \xi_i} = C - \alpha_i - \mu_i = 0, \tag{4.24}$$

$$y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 + \xi_i \geq 0, \tag{4.25}$$

$$\xi_i \geq 0, \tag{4.26}$$

$$\alpha_i \geq 0, \tag{4.27}$$

$$\mu_i \geq 0, \tag{4.28}$$

$$\alpha_i\{y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 + \xi_i\} = 0, \tag{4.29}$$

$$\mu_i \xi_i = 0. \tag{4.30}$$

We can use eqs. (4.29) and (4.30) to determine the threshold b. Moreover, we can notice that combining eq. (4.24) with eq. (4.29) we obtain that $\xi_i = 0$ if $\alpha_i < C$. Therefore we can simply take any training point for which $0 < \alpha_i < C$ to use eq. (4.29) (with $\xi_i = 0$) to calculate b.

4.2 An introduction to Nonlinear Support Vector Machines

As seen above, the only way in which the data appear in the training problem (see equations (4.17)-(4.19)) is in the form of dot products, $\mathbf{x}_i \cdot \mathbf{x}_j$. Therefore we now suppose to map the data to a new Euclidean space $\mathcal{H}$ (possibly infinite dimensional), using a mapping function Φ:

$$\Phi : \mathbf{R}^d \mapsto \mathcal{H}. \tag{4.31}$$

Therefore, the training algorithm will only depend on the data though dot products in $\mathcal{H}$, of the form $\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$. Hence, if we define a 'kernel function' K:

$$K(\mathbf{x}_i, \mathbf{x}_i) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j), \tag{4.32}$$

we would only need to employ K in the training algorithm in order to compute a solution and we would never need to explicitly know what Φ is. If one replaces $\mathbf{x}_i \cdot \mathbf{x}_j$ with $K(\mathbf{x}_i, \mathbf{x}_j)$ everywhere in the training algorithm, its application will produce support vectors that live in $\mathcal{H}$.

In order to use the machine in the test phase we need $\mathbf{w}$, that also lives in $\mathcal{H}$. During test phase an SVM is used by computing dot products of a given test point $\mathbf{x}$ with $\mathbf{w}$, or more precisely by evaluating the sign of the following:

$$f(\mathbf{x}) = \sum_{i=1}^{N_S} \alpha_i y_i \Phi(\mathbf{s}_i) \cdot \Phi(\mathbf{x}) + b = \sum_{i=1}^{N_S} \alpha_i y_i K(\mathbf{s}_i, \mathbf{x}) + b, \tag{4.33}$$

where the $\mathbf{s}_i$ are the support vectors. Therefore, again we can avoid the explicit calculation of $\Phi(\mathbf{x})$ and use $K(\mathbf{s}_i, \mathbf{x}) = \Phi(\mathbf{s}_i) \cdot \Phi(\mathbf{x})$ instead.

To be sure of obtaining a unique solution, the kernel function has to be opportunely chosen. In particular it has to satisfy the Mercer's condition [34, 36]. It says that there exists a mapping Φ and an expansion:

$$K(\mathbf{x}, \mathbf{y}) = \sum_i \Phi(\mathbf{x})_i \Phi(\mathbf{y})_i, \tag{4.34}$$

if and only if, for any $g(\mathbf{x})$ such that:

$$\int g(\mathbf{x})^2 d\mathbf{x} \qquad \text{is finite}, \tag{4.35}$$

then

$$\int K(\mathbf{x}, \mathbf{y}) g(\mathbf{x}) g(\mathbf{y}) d\mathbf{x} d\mathbf{y} \geq 0. \tag{4.36}$$

For specific cases it may be not easy to verify whether Mercer's condition is satisfied.

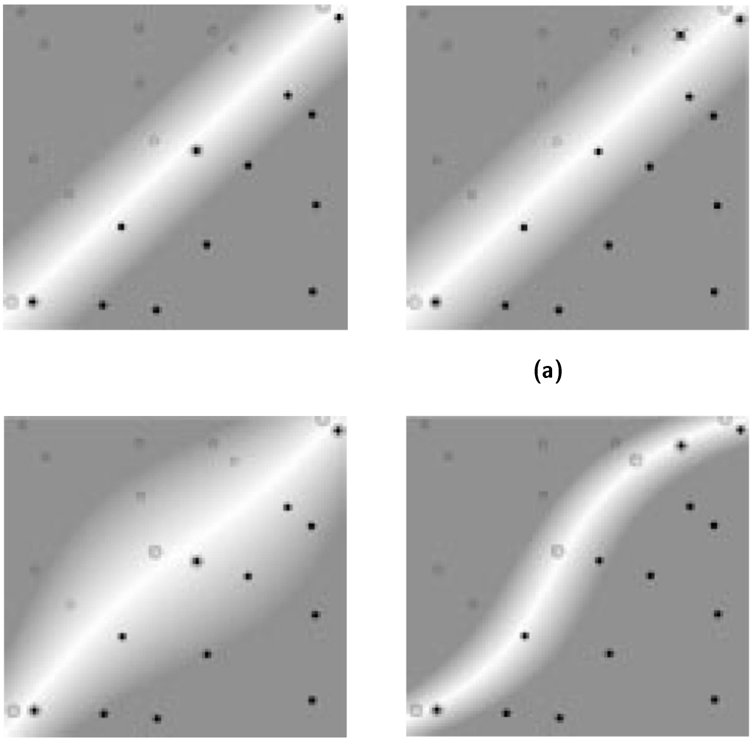
In general, if one uses a kernel for which Mercer's condition does not hold, there may exist data such that the Hessian is indefinite, and for which quadratic programming problem will have no solution (the dual objective function can become arbitrarily large) [36]. However, even for kernels that do not satisfy Mercer's condition, one may check if the training data have a positive semidefinite Hessian, in which cases the training will converge perfectly well.

Some examples showing the difference between linear and nonlinear kernel SVMs applications to a two-class pattern recognition problem are reported in figure 4.3. Figures 4.3c and 4.3d show results for the same pattern recognition problem displayed in 4.3a and 4.3b respectively. Notice that, even though the number of degrees of freedom is higher when a cubic kernel (a polynomial of degree p in the data $K(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y} + 1)^p$, where $p = 3$ in the case of a cubic kernel) is chosen, for the linearly separable case (figure 4.3c), the solution is roughly linear, indicating that the capacity is being controlled. Moreover notice that the linearly non-separable case (figure 4.3b) has become separable (figure 4.3d) [36].

The optimization method currently used is an active set method that combines gradient and conjugate gradient ascendent. Firstly the objective function is computed and also its gradient, at very little extra cost. In phase 1 the search directions $\mathbf{s}$ are fixed along the gradient. Then, the nearest face along the search direction is found. If the dot product of the gradient there with $\mathbf{s}$ indicates that the maximum along $\mathbf{s}$ lies between the current point and the nearest face, the optimal point along the search direction is computed analytically, and phase 2 can begin. Otherwise the algorithm jumps to the new face and repeat phase 1. In phase 2 conjugate gradient ascendent is done, until a new face is encountered or the stopping criterion is met. The algorithm stops when the fractional rate of increase of the objective function falls below a tolerance. A good manner to verify if the algorithm is working is to check that the solution satisfies all the KKT conditions for the primal problem. In fact these represent the necessary and sufficient conditions that the solution be optimal. The KKT conditions are equations (4.22) through eq. (4.30), with dot products between data vectors replaced by kernels wherever they appear [36]. Furthermore, a more detailed description of the optimization algorithm developed for SVM-Light package by Joachims is given in [38].

4.3 Discrimination maps and Recursive Feature Elimination

As we will see later, in the present study we exclusively used a linear kernel SVM to reduce the risk of overfitting the data and to make possible the direct extraction of the weight vector as an image (the SVM discrimination map). We remark that the weight vector is



(a) (b)

(c) (d)

Figure 4.3: Four examples of a two-class pattern recognition problem. The two classes are denoted by black and grey circles respectively. The background colour shows the shape of the decision surface. Figures 4.3a and 4.3b regard linear SVMs, separable and non-separable cases respectively. Moreover, figures 4.3c and 4.3d are obtained using a cubic kernel for the same cases respectively of figures 4.3a and 4.3b [36].

a linear combination of the support vectors and is normal to the estimated hyperplane. Moreover the linear kernel only has one parameter (C) that controls the trade-off between having zero training errors and allowing misclassifications.

When we use linear kernel SVMs to distinguish between two groups, the weight vector normal to the hyperplane indicates the direction along which the two groups of images differ most [39]. Therefore, we can use this vector to generate a map of the most discriminating regions, so called discrimination map. Because the SVM classifier is naturally multivariate, this map can be interpreted as the spatial pattern by which the groups differ.

In order to identify the the most discriminating set of voxels, SVM recursive feature elimination (SVM-RFE) can be applied [42, 43]. This method iteratively removes features (voxels) from the data set with the purpose of removing the major number of non-informative features as possible while retaining most discriminating features. Having chosen a feature ranking criterion to remove a subset of the least informative features, an SVM classifier is trained at each iteration with the aim of evaluating the behavior of classification accuracy as a function of the number of voxels in the data. There is some evidence that SVM-RFE can produce improvements in the predictive accuracy for functional MRI (fMRI) data [43]. In contrast, we will use this method to remove non-informative features from the weight vector and to find the most parsimonious weight representation possible. Therefore, each feature subset does not depend on the test data because the weight vector has been computed exclusively from the training data and the resulting estimate of accuracy will be unbiased, since the weight vector group map has been constructed using all subject data [39].

4.4 Evaluation of learning methods

Evaluation is the key to making real progress in learning methods. To determine which ones to apply in a particular situation we need systematic ways to evaluate how different methods work and to compare one with another.

4.4.1 Error rate

For classification procedures, it is natural to quantify the classification performance in terms of the *error rate*. If the classifier correctly predicts the class of a instance then it is counted as a *success*. Instead, if not, it is counted as an *error*. Therefore the error rate can be considered as the proportion of errors made over a whole set of instances, and it measures the overall performance of the classifier [35].

Obviously, we are interested on the future performance on new data, not the past performance on old data. Hence, performance on the training set does not represent a good indicator of performance on an independent test set. If the old data was used during the learning process to train the classifier, the error rate on old data cannot be considered to be a good indicator of the error rate on new data.

The error rate on the training data is called the *resubstitution error*, since it is calculated by resubstituting the training instances into a classifier that was constructed for them. Although it is not a reliable predictor of the true error rate on new data, it is often usually to know.

Instead, to evaluate the performance of a classifier on new data, we have to estimate the error rate on data that are not been used during the learning phase. This independent dataset is called *test set*. It is important that the test data was not used in any way to create the classifier. If the *training set* is used to create the classifier, then the test data is used to calculate the error rate of the final classification method. The accuracy of the error estimate can be quantified statistically [35].

Table 4.1: Different outcomes of a two-class prediction.

		Predicted class	
		yes	no
Actual class	yes	TP	FN
	no	FP	TN

In the two-class classification case with classes *yes* and *no*, a single prediction has the four possible outcomes shown in table 4.1. The *true positives* (TP) and the *true negatives* (TN) are correct classifications. A *false positive* (FP) occurs when the outcome is incorrectly predicted as positive when it is actually negative. A *false negative* (FN) occurs when the outcome is incorrectly predicted as negative when it is actually positive. Therefore we can define the true positive rate (tp) and the false positive rate (fp):

$$tp = \frac{TP}{TP + FN} \tag{4.37}$$

$$fp = \frac{FP}{FP + TN} \tag{4.38}$$

Moreover we can define the overall success rate as the number of correct classifications divided by the total number of classifications:

$$\text{success rate} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.39}$$

Finally, the error rate is one minus the last:

$$\text{error rate} = 1 - \text{success rate} = \frac{FN + FP}{TP + TN + FP + FN} \tag{4.40}$$

Physicians talk about the *sensitivity* and *specificity* of diagnostic tests. Sensitivity refers to the proportion of people with disease who have a positive test result, that is tp. Specificity refers to the proportion of people without disease who have a negative test result, which is $1 - fp$. Sometimes the product of these two quantities is used as an overall measure:

$$\text{sensitivity} \ \times \ \text{specificity} = tp(1 - fp) = \frac{TP \cdot TN}{(TP + FN) \cdot (FP + TN)} \tag{4.41}$$

In a multiclass prediction, the result on a test set is often displayed as a two-dimensional *confusion matrix* with a row and a column representing each class. Each matrix element shows the number of test examples for which the actual class is the row and the predicted class is the column. Good results correspond to large numbers in the main diagonal and small, ideally zero, off-diagonal elements [35].

4.4.2 Cross-validation

Since in general we have a limited amount of data for training and testing, the classification procedure uses a certain amount for testing and the reminder for training. In practical terms, it is common to hold out one-third of data for testing and use the remaining two-thirds for training. Both the samples used for testing and for training have to be representative subsets of the underlying problem. Therefore a random sampling have to be done in such a way as to guarantee that each class is properly represented in both training and test sets. This procedure is called *stratification*, and we might speak of *stratified holdout*. A general way to reduce any bias caused by the particular sample chosen for holdout is to repeat the whole process, training and testing, many times with different random samples. In each iteration a certain proportion of the data is randomly selected for training, possibly with stratification, and the reminder used for testing. The error rates on the different iterations are averaged to estimate an overall error rate. This is the *repeated holdout* method of the error rate estimation.

In a single holdout procedure, one could consider swapping the roles of testing and training data and average the two results, thus reducing the effect of not consistent representation in training and test sets. However, it is only possible with a 50:50 split between training and test data, which is generally not ideal, in fact it is preferable to use more than half the data for training even at expense of test data. A simple variant of this basic technique is a statistical method called *cross-validation*. Within this procedure, one decides to divide the data in a fixed number of *folds*, or partitions of the dataset. For example, we can decide of dividing the data into three approximately equal partitions and each in turn is used for testing and the remaining is used for training. That is, we are using two-thirds for training and one-third for testing and then we can repeat the procedure three times so that, in the end, every instance has been used exactly once for testing. This is called *three-fold cross-validation*, and if stratification is applied as well, it is called *stratified cross-validation*. The last only was an example, but there are several ways of measuring the error rate of a learning scheme on a particular dataset. Two that are particularly important are *leave-one-out* cross-validation and the *bootstrap*.

Leave-one-out

Leave-one-out cross-validation is a n-fold cross-validation, where n represents the number of instances in the dataset. Each instance in turn is left out, and the classifier is trained on all the remaining instances $(n-1)$. The results of all n judgments are averaged, and that average represents the final error estimate.

This algorithm possesses two advantages. Firstly, the greatest possible amount of data is used for training at each iteration, which presumably increases the chance that the classifier is an accurate one. Secondly, the procedure is deterministic: no random sampling is involved. There is no way of repeating it, the same result will be obtained at each iteration.

Nevertheless, this method has a high computational cost, because the entire learning procedure has to be repeated n times, and this is really expensive for large datasets. Another disadvantage is that this method cannot be stratified. Stratification involves getting the correct proportion of examples in each class into the test set, and this is impossible when the test set contains only a single example.

However, leave-one-out cross-validation seems to offer a chance of obtaining the maximum out of a small dataset, resulting in an accurate estimate as possible.

A variant of leave-one-out that we have adopted in the current study is the commonly used leave-two-out cross-validation approach, which provides a relative unbiased estimate of the true generalization performance [39]. In each trial observations from one subject from each group are used to train the classifier.

The bootstrap

The bootstrap is based on the statistical procedure of sampling with replacement. In the precedent method, whenever a sample was taken from the dataset to form a training or test set, it was drown without replacement. Therefore, the same instance, once selected, could not be chosen again. Instead, the idea of the bootstrap method is to sample the dataset with replacement to form a training set.

A particular variant is called the *0.632 bootstrap*, where a dataset of n instances is sampled n times, with replacement, to give another dataset of n instances. Since there

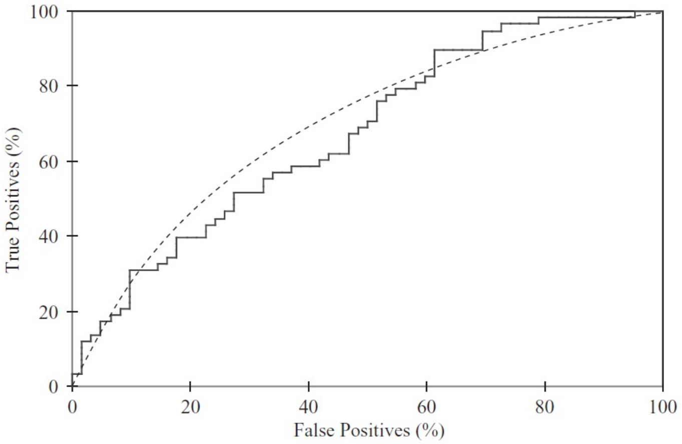


Figure 4.4: An example of ROC curve [35].

must be some instances in the original dataset that have not been picked, they can be used to form the test set. We can evaluate the chance that a particular instance will not be picked after n extractions with replacement. It has a $1/n$ probability of being picked each time and therefore a probability $1 - 1/n$ of not being picked. Hence we have to multiply these probabilities for the number of extractions, obtaining:

$$p = \left(1 - \frac{1}{n}\right)^n \approx e^{-1} = 0.368 \tag{4.42}$$

The last gives the chance of a particular instance not being picked at all. Therefore, for a reasonably large dataset, the test set will contain about the 36.8% of the instances and the training set the 63.2% (this is the reason why this method is called 0.632 bootstrap).

However, this method usually leads to a pessimistic estimate of the true error rate, because the training set contains only 63% of instances, which is not great deal compared, with the 90% of 10-fold cross-validation. Often the whole bootstrap procedure is repeated several times and the results averaged.

4.4.3 ROC curves

Diagnostic accuracy of data mining schemes can be evaluated through graphical techniques such as *ROC curves* (*receiver operating characteristic curves*)[46, 47]. These curves represent the performance of a classifier plotting the sensitivity (tp) versus 1-specificity, i.e. the false positive rate (fp). A ROC curve estimates and reports all of the combinations of sensitivity and specificity that a diagnostic test is able to provide. The northwest corner is the better place to be, in fact here there are the maximum number of true positives and the minimum number of false positives, as shown in figure 4.4 (ROC curve is the jagged line). In general, one point is better than another if tp is high, fp is low, or both. The point (0,1) represents a perfect classification. The diagonal line $y = x$ represents the effect of randomly guessing a class.

It can be estimated by applying cross-validation. A simple approach is to collect the predicted probabilities for all the various test sets (of which there are 10 in a 10-fold cross-validation), along with the true class labels of the corresponding instances, and generate a single ranked list based on this data. This assumes that the probability estimates from the classifiers built from the different training sets are all based on equally sized random samples of the data [35].

ROC curves can be summarized in a single quantity: the *area under the curve* (AUC). The last is largely employed because larger is the area then better is the model. The area can also be interpreted as the probability that the classifier ranks a randomly positive instance above a randomly chosen negative one. Moreover, the meaning of AUC has been proved to be the probability that a random pair of positive/diseased and negative/non-diseased individuals would be correctly identified by the diagnostic test [44].

Chapter 5

Magnetic Resonance Imaging data acquisition, preprocessing and volumetric analysis

While the VBM is a standard and well-established mass-univariate method, the SVM multivariate analysis has been rarely implemented to investigate brain MRI data. An improvement of our knowledge on the pattern classification approach is necessary to be achieved, both to assess its exploratory capability and to point out advantages and disadvantages with respect to the more largely used VBM approach. Despite these methods are potentially suitable to investigate a large variety of neurological and neuropsychiatric disorders, in the present study they have been employed with the purpose of detecting neuroanatomical and gender-related abnormalities in children with autism spectrum disorders (ASD). In fact, the differences in the neuroanatomy of young children with ASD are an intriguing and still poor investigated issue.

The preliminary step both in VBM and SVM analyses is the image preprocessing described in this chapter. The data acquisition parameters, the demographic characteristics and the global volumetric analysis of the subjects involved in this study are also presented and discussed in this chapter.

5.1 Magnetic Resonance Imaging data acquisition

A total of 152 structural MRI scans were selected. Specifically, our dataset was composed by 76 ASD children and 76 matched controls in the 2-7 year age range. The matching was based on age, sex and non verbal intelligence quotient (NVIQ). In fact, both groups of patients (ASD) and controls (C) were composed by 38 males and 38 females, of which 19 males and 19 females with $\mathrm{NVIQ}<70$ (ASDm-DD, ASDf-DD, Cm-DD and Cf-DD respectively) and 19 males and 19 females with NVIQ ≥ 70 (ASDm-noDD, ASDf-noDD, Cm-noDD and Cf-noDD respectively), where m stands for male, f stands for female, DD means idiopathic developmental delay ($\mathrm{NVIQ}<70$) and noDD regards children without developmental delay (NVIQ $\geq$ 70).

MRI data were acquired using a GE 1.5 T Signa Neuro-optimized System (General Electric Medical Systems) at IRCCS Stella Maris Foundation, fitted with 40 mT/m high-speed gradients. The standard MR protocol for children included FSE T2-wighted, FLAIR, DWI, SE T1-weighted sequences and a single voxel 1H MR spectrum. Furthermore, a whole-brain fast spoiled gradient recalled acquisition in the steady-state T1-weighted series (FSPGR) were collected in the axial plane with repetition time 12.4 ms, echo time 2.4 ms, inversion time 700 ms, flip angle of 10°, yielding 124 contiguous 1.1 mm axial slices with an in-plane resolution of 1.1×1.1 mm^2 [40].

The research protocol was approved by the Institutional Review Board of the Clinical Research Institute for Child and Adolescent Neurology and Psychiatry.

5.2 Voxel-Based Morphometry-DARTEL image preprocessing

The T1-weighted volumetric images were analyzed with SPM8 package (Statistical Parametric Mapping, Wellcome Department of Imaging Neuroscience, London, UK, `http://www.fil.ion.ucl.ac.uk/spm`), according to the VBM protocol [13] with modulation. Moreover, we implemented the DARTEL algorithm [26], where a diffeomorphic warping is applied, in order to obtain an accurate inter-subject registration and to generate a study-specific template. The VBM preprocessing involves the following steps [40]:

1. checking for scanner artifacts and gross anatomical abnormalities for each subject of our dataset;
2. setting the origin of each image to the anterior commissure;
3. SPM8 segmentation of brain tissues through the most recent tool 'New Segment' [14];
4. importing the parameter files produced by the tissue segmentation in the DARTEL procedure to generate a study-specific template, and to obtain the brain tissues segmented according to it;
5. affine registration of segmented brain tissues into the MNI space;
6. checking for homogeneity across the sample and using standard smoothing (8 mm isotropic Gaussian kernel).

After this preprocessing we obtained smoothed modulated normalized data (in the MNI space) to be used for the volumetric analysis (see Section 5.3), for the voxel-wise two-sample t-test VBM analysis (see Chapter 6) and as input features in the SVM pattern recognition approach (see Chapter 7). The aim of the modulation step was to render the VBM statistics reflective of the absolute amount of GM in different regions rather than the relative concentration of GM (the proportion of grey matter to other tissue types within a region). [17].

5.3 Participant characteristics and volumetric analysis

Group differences were evaluated by employing self-written Matlab codes for gray matter (GM), white matter (WM), cerebrospinal fluid (CSF) absolute volumes, and total intracranial volume (TIV), calculated as the sum of GM, WM and CSF volumes. The brain regional absolute volumes obtained in the brain segmentation step of the VBM-DARTEL preprocessing were considered. The analysis of variance (ANOVA) was applied to identify any significant between-group difference in global tissue volumes. Table 5.1 shows details of the entire dataset, composed by 76 males and 76 females; tables 5.2 and 5.3 analyse separately these two groups and finally tables 5.4 and 5.5 show the characteristics of the two subgroups of DD and noDD in male and female subsets. The means and the standard deviations of age, NVIQ, GM, WM, CSF volumes and TIV are displayed, and the between-groups statistics are reported.

There are no meaningful differences between groups on age and NVIQ. Moreover, results of ANOVA for brain regional volume group differences revealed the following results on $p < 0.05$: no significant differences between ASD subjects and controls in the subgroups of

Table 5.1: Sample characteristics and global volume group differences in the entire dataset.

Variable	Subject group, mean ± SD [range]		ANOVA	
	ASD (n=76)	C (n=76)	F	p value
Age, months	53 ± 17 $[25 - 88]$	53 ± 18 $[22 - 89]$	0.0002	0.99
NVIQ	71 ± 22 $[30 - 113]$	73 ± 23 $[35 - 112]$	0.133	0.72
GM, ml	662 ± 67	629 ± 80	7.39	0.007
WM, ml	424 ± 47	400 ± 55	8.84	0.003
CSF, ml	225 ± 25	218 ± 34	2.41	0.123
TIV, ml	1311 ± 134	1247 ± 162	7.18	0.008

Table 5.2: Sample characteristics and global volume differences in male group: the characteristics of the subsamples of subjects with and without developmental delay (DD) are separately reported.

Variable	Subject group, mean ± SD [range]		ANOVA	
	ASDm (n=38)	Cm (n=38)	F	p value
Age, months	53 ± 16 $[27 - 87]$	53 ± 17 $[24 - 88]$	0.006	0.94
NVIQ	71 ± 21 $[39 - 113]$	74 ± 23 $[43 - 112]$	0.25	0.62
GM, ml	699 ± 62	664 ± 83	4.25	0.04
WM, ml	449 ± 43	423 ± 56	5.38	0.02
CSF, ml	238 ± 25	232 ± 38	0.53	0.47
TIV, ml	1386 ± 123	1319 ± 168	3.9	0.052

DD and noDD have been found; GM and WM volumes were found to be higher in ASD subjects than controls in the complete dataset and in the subgroups of male and female data. In addition these statistical analyses pointed out meaningful differences in TIV in the entire dataset, and in CSF volume and TIV in the female subgroup.

These statistical tests allowed whole-brain volume comparisons between groups. However the VBM-voxel-wise two-sample t-test analysis could detect any brain local regions in which GM volume is significantly higher in ASD subjects as compared to control group.

Table 5.3: Sample characteristics and global volume differences in female group: the characteristics of the subsamples of subjects with and without developmental delay (DD) are separately reported.

Variable	Subject group, mean ± SD [range]		ANOVA	
	ASDf (n=38)	Cf (n=38)	F	p value
Age, months	53 ± 18 $[25 - 88]$	53 ± 19 $[22 - 89]$	0.0025	0.96
NVIQ	71 ± 22 $[30 - 103]$	71 ± 24 $[35 - 100]$	0.0004	0.98
GM, ml	625 ± 48	594 ± 60	5.9	0.018
WM, ml	400 ± 38	376 ± 44	5.9	0.017
CSF, ml	212 ± 17	203 ± 19	5.1	0.027
TIV, ml	1237 ± 100	1174 ± 120	6.1	0.016

Table 5.4: Sample characteristics and global volume differences in the subsets of ASD and Controls male groups.

Variable	Subject group, mean ± SD [range]		ANOVA		Subject group, mean ± SD [range]		ANOVA	
	ASDm-DD (n=19)	Cm-DD (n=19)	F	p value	ASDm-noDD (n=19)	Cm-noDD (n=19)	F	p value
Age, months	55 ± 16 [27 − 82]	52 ± 13 [29 − 72]	0.25	0.62	52 ± 16 [34 − 87]	54 ± 21 [24 − 88]	0.08	0.78
NVIQ	53 ± 9 [39 − 68]	52 ± 7 [43 − 75]	0.09	0.77	89 ± 13 [70 − 113]	95 ± 10 [78 − 112]	2.35	0.13
GM, ml	702 ± 61	661 ± 85	2.98	0.09	696 ± 65	668 ± 83	1.34	0.26
WM, ml	453 ± 44	424 ± 62	2.72	0.11	445 ± 43	421 ± 51	2.55	0.12
CSF, ml	238 ± 22	228 ± 31	1.13	0.29	238 ± 30	236 ± 44	0.01	0.91
TIV, ml	1393 ± 121	1314 ± 171	2.71	0.11	1378 ± 167	1325 ± 169	1.24	0.27

Table 5.5: Sample characteristics and global volume differences in the subsets of ASD and Controls female groups.

Variable	Subject group, mean ± SD [range]		ANOVA		Subject group, mean ± SD [range]		ANOVA	
	ASDf-DD (n=19)	Cf-DD (n=19)	F	p value	ASDf-noDD (n=19)	Cf-noDD (n=19)	F	p value
Age, months	47 ± 18 [25 − 83]	51 ± 18 [26 − 77]	0.5	0.47	59 ± 16 [36 − 88]	56 ± 20 [22 − 89]	0.4	0.5
NVIQ	53 ± 14 [30 − 69]	50 ± 9 [35 − 65]	0.5	0.47	90 ± 10 [71 − 103]	93 ± 10 [78 − 100]	0.8	0.4
GM, ml	616 ± 59	584 ± 53	3.2	0.08	633 ± 35	605 ± 66	2.7	0.1
WM, ml	393 ± 44	369 ± 41	3.0	0.09	406 ± 31	384 ± 47	2.9	0.1
CSF, ml	211 ± 20	200 ± 16	3.5	0.07	214 ± 13	206 ± 22	1.7	0.2
TIV, ml	1221 ± 121	1153 ± 107	3.4	0.07	1253 ± 73	1195 ± 132	2.8	0.1

Chapter 6

Voxel-Based Morphometry analysis of grey matter segments

VBM statistical analysis was applied to the entire dataset and then to the male and female subgroups. In all these cases the ASD versus control subjects were given as input groups to the voxel-wise two-sample t-test , in order to detect neuroanatomical abnormalities in patients and healthy subjects.

6.1 Statistical analysis of the entire dataset

The regional grey matter (GM) volumes of the entire dataset, composed by 76 males and 76 females, were compared between the two groups of ASD and controls using the VBM-DARTEL analysis. In order to quantify the comparison, the conventional voxel-wise two-sample t-test VBM analysis (see Section 3.5.3) in SPM8 was applied, inserting as covariates Sex, Age and NVIQ, thus ending up with 147 degrees of freedom. The resulting design matrix is shown in figure 6.1.

The usual stringent significance threshold $p < 0.05$ and the family-wise error rate (FWE) correction with an extent threshold of 10 voxels were employed. Moreover, an absolute threshold mask of 0.1 on GM was used to avoid possible edge effects around the border between GM and WM.

A significant volumetric between-group difference has been found in the GM of the entire dataset, as can be seen in figure 6.2. The resulting regions showing larger local GM volume in the ASD group compared to control subjects are listed in table 6.1, where both MNI and Talairach[1] (TAL) coordinates of each blob centroid are reported. The involved brain regions are characterized also in terms of Brodmann Areas[2] and of the number of voxels contained in the cluster. Moreover, the localization in the right (R) hemisphere or in the left (L) one is specified. The opposite contrast, searching for brain regions having larger GM volume in the control group compared to controls, gave no significative results, consistently to the GM volumetric analysis results displayed in table 5.1.

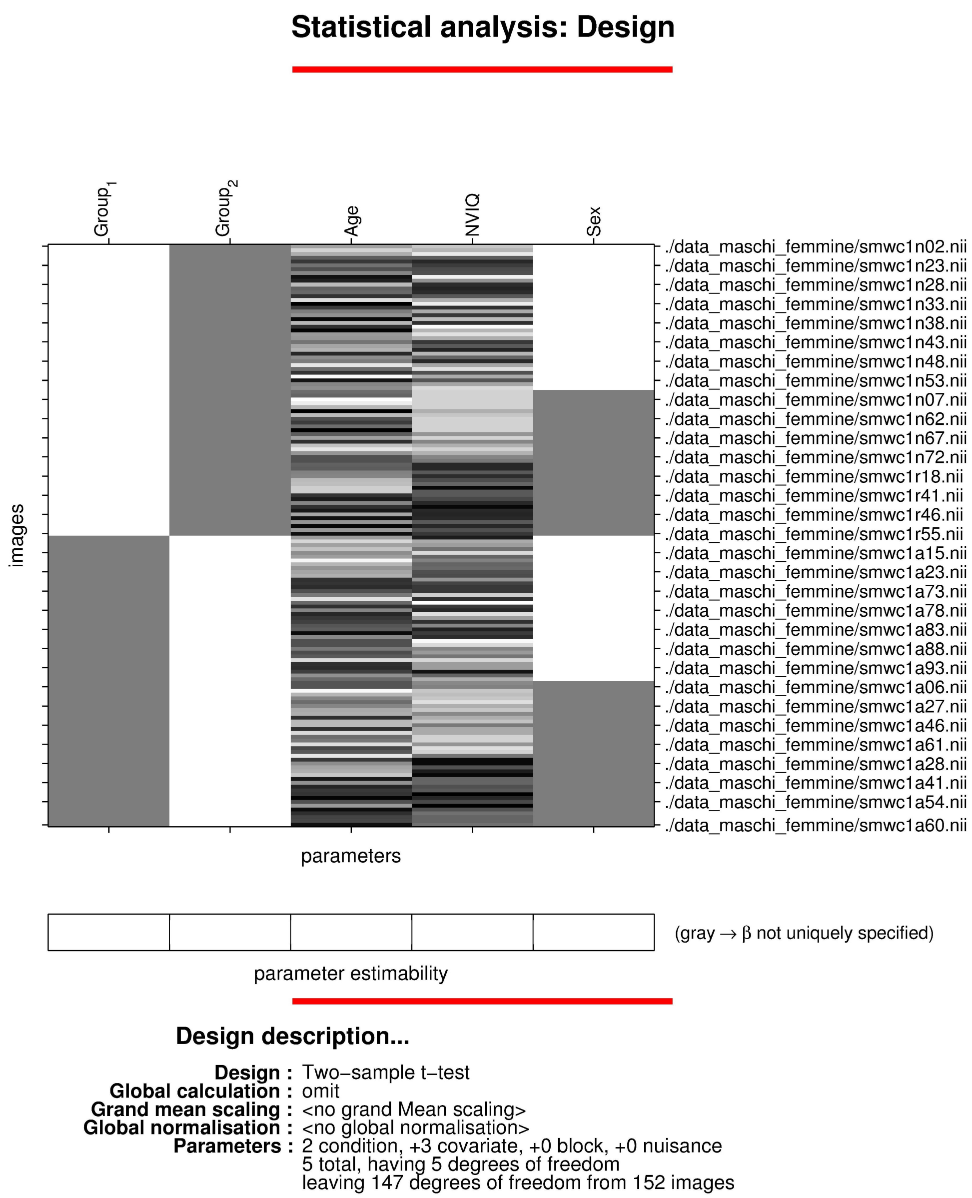
6.2 Statistical analysis of the male subgroup

Normalized modulated and smoothed GM image segments of male subjects, beforehand obtained according to the male plus female study specific template generated adopting the

---

[1] The Talairach coordinate system of the human brain is commonly used to describe the location of brain structures independently of individual differences in the size and overall shape of the brain. Distances in Talairach coordinates are measured from the anterior commissure as origin.

[2] A Broadmann Area is a region of the cerebral cortex defined based on its cytoarchitectural organization of neurons.



Design description...

Design : Two−sample t−test
Global calculation : omit
Grand mean scaling : <no grand Mean scaling>
Global normalisation : <no global normalisation>
Parameters : 2 condition, +3 covariate, +0 block, +0 nuisance
5 total, having 5 degrees of freedom
leaving 147 degrees of freedom from 152 images

Figure 6.1: Two-sample t-test design matrix for the entire dataset. Group$_1$=control subjects, Group$_2$=ASD subjects.

Table 6.1: VBM analysis results for the entire dataset.

		Number of voxels	MNI coordinates (x,y,z)	Talairach coordinates (x,y,z)
Limbic lobe, Parahippocampal gyrus, BA 35	L	284	$(-17, -21, -17)$	$(-16, -20, -13)$
Limbic lobe, Parahippocampal gyrus, Hippocampus	R	79	$(27, -16, -18)$	$(24, -15, -13)$
Sub-lobar, Caudate, Caudate tail	R	14	$(39, -27, -9)$	$(35, -26, -6)$
Frontal lobe, Superior frontal gyrus, BA 10	L	244	$(-27, 53, 4)$	$(-26, 47, 12)$
Temporal lobe, Superior temporal gyrus, BA 22	R	38	$(69, -49, 16)$	$(63, -49, 15)$
Parietal lobe, Precuneus, BA 31	L	1190	$(0, -45, 34)$	$(-2, -47, 30)$
Limbic lobe, Cingulate gyrus, BA 23	R	10	$(5, -22, 31)$	$(2, -25, 30)$

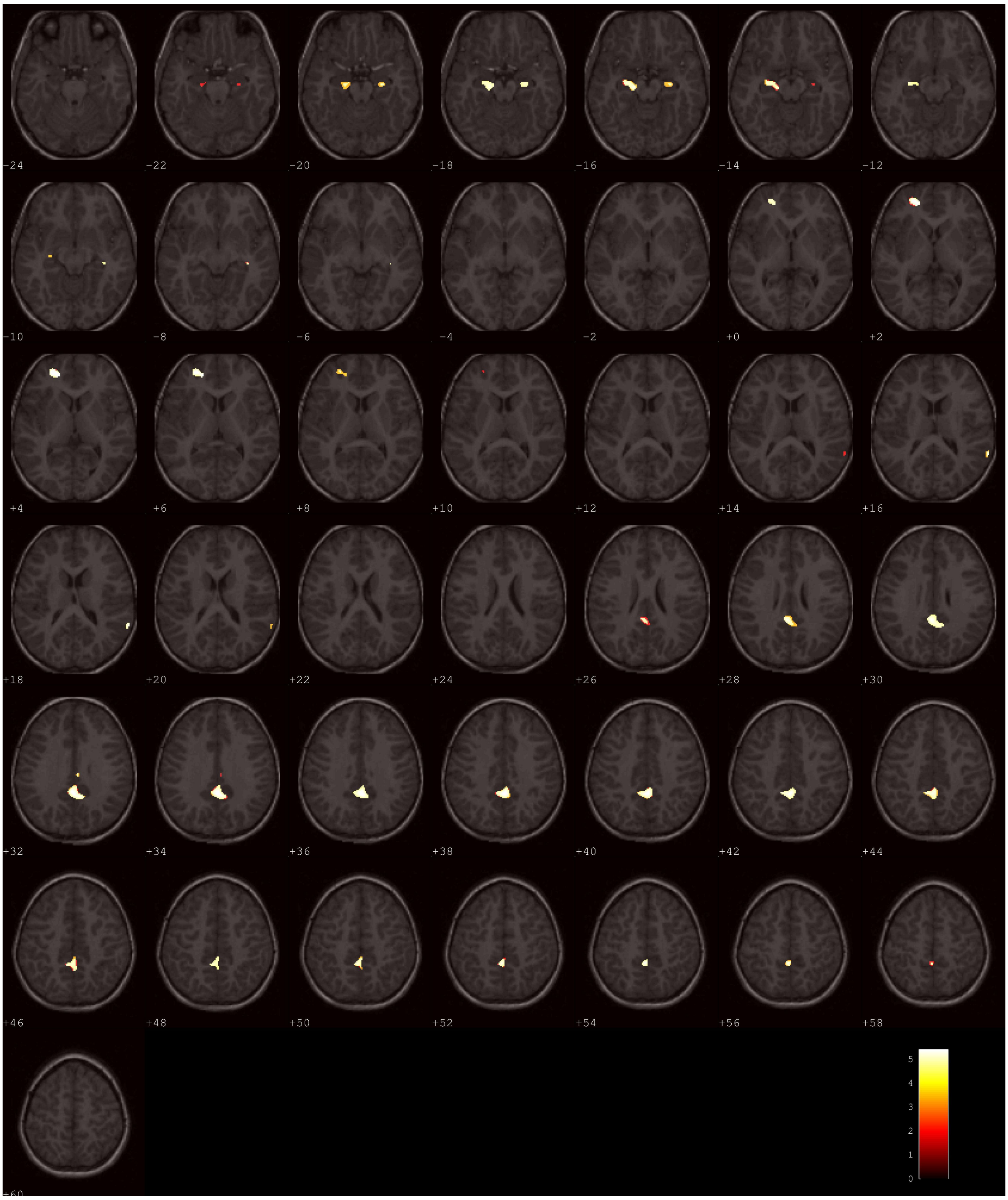


Figure 6.2: Brain regions showing larger local GM volume in the ASD group compared to control subjects are overlaid to a representative structural MR image normalized to the MNI space. These regions are a VBM result with $p < 0.05$, FWE corrected and with an extent threshold of 10 voxels.

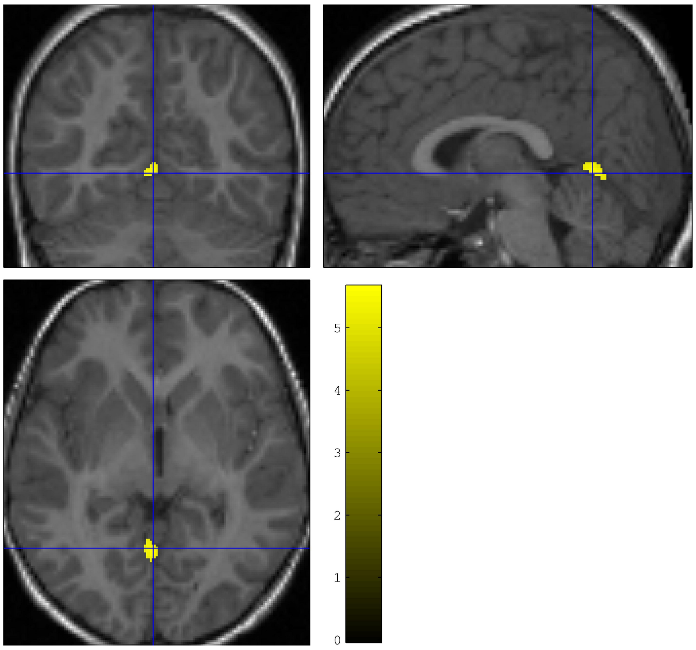


Figure 6.3: Brain region showing larger local GM volume in the male ASD group compared to male control subjects is overlaid to a representative structural MR image normalized to the MNI space. This region is a VBM result with p< 0.05, FWE corrected and with an extent threshold of 10 voxels. This blob is composed by 115 voxels and is localized in the Left Cerebellum, Anterior Lobe, Culmen (MNI coordinates: (-3,-63,-2) ; Talairach coordinates: (-4,-60,-3)

DARTEL algorithm, were entered into a voxel-wise two-sample t-test analysis in SPM8 to compare ASD and controls subgroups. Age and NVIQ were employed as covariates, thus ending up with 72 degrees of freedom, and performing the stringent significance threshold $p < 0.05$. The statistical analysis was family-wise error rate (FWE) corrected and an extent threshold of 10 voxels was applied. An absolute threshold mask of 0.1 on GM was performed to avoid possible edge effects around the border between GM and WM. A significant volumetric between-group difference was found in the GM of the male data set, as shown in figure 6.3. The brain blob found represents a brain region in which there is a greater GM volume in ASD group with respect to control group.

The opposite contrast, searching for brain regions having larger GM volume in the control group compared to controls, gave no significative results, reliably with respect to GM volumetric analysis results displayed in table 5.2.

6.3 Statistical analysis of the female subgroup

Normalized modulated and smoothed GM image segments of female subjects, obtained according to the male plus female study specific template generated according to the DARTEL procedure, were entered into a voxel-wise two-sample t-test analysis in SPM8 with the aim of comparing ASD subgroup with respect to controls. Age and NVIQ covariates were employed, thus ending up with 72 degrees of freedom, and performing the stringent significance threshold $p< 0.05$. The statistical analysis was family-wise error rate (FWE) corrected and an extent threshold of 10 voxels was used. An absolute threshold mask of 0.1 on GM was used to avoid possible edge effects around the border between GM and WM. None significant volumetric between-group difference was found in the GM of the female dataset employing both the contrasts.

6.4 Considerations

By extending the significance level in both the statistical analyses of the male and female subgroups, we reobtained some of those regions that have been detected for the entire data set. When the stringent FWE correction is employed, several regions are not identified as significative ones because of the restricted number of data contained in the subsets.

Chapter 7

Pattern classification approach to grey matter segments analysis

The preprocessed GM segments obtained using the SPM8 software were analyzed using the SVM pattern classification approach, that presents the advantage of intrinsically taking into account interregional correlations.

Then the SVM recursive feature elimination (SVM-RFE) procedure has been implemented both to reduce the large number of features in the classification problem and to enhance the discrimination capability. The SVM-RFE allows also to localize the most discriminant voxels and to visualize them in a discrimination map. However, the pattern classification method was not employed to predict the class membership of undiagnosed subjects, but as a figure of merit allowing to determine an optimal threshold on the discrimination maps, where possible between-group structural differences are encoded.

This pattern classification approach has been implemented both for the entire dataset and for the male and female subgroups, in order to individuate those regions that characterize each group. Since do not exist automatic and standard tools that aid SVMs implementation like those available to apply VBM, all the examinations have been actuated using self-written MatLab codes and employing the SVM-Light software package [38] to train and validate our data.

7.1 Support Vector Machine classification of grey matter segments

The normalized modulated and smoothed GM image segments regarding the entire dataset and obtained from the VBM-DARTEL preprocessing were given in input to the support vector machine (SVM) as vectors of features. Each feature represents the amount of GM in each voxel, as the modulation step has been performed in the SPM8 software.

Linear kernel SVMs have been applied to counteract the data overfitting, that may occurr as the number of fatures/voxels is very large (about 4×10^5), compared to the 152 patterns of the dataset. In our analysis each patient is matched to a control with respect to sex, age and NVIQ score, hence we have 76 couples of data.

The SVMs have been trained according to the leave-pair-out cross-validation (LPO-CV) procedure, thus excluding one couple of matched subjects from the training set at each iteration, and validating the trained SVM on it. The classification performance has been evaluated estimating the area under the receiver operating characteristic curve (AUC), which meaning has been proved to be the probability that a random pair of positive/diseased and negative/non-diseased individuals would be correctly identified by the diagnostic test [44]. Linear kernel SVMs depend only on one parameter (C), that holds power on the trade-off between having zero training errors and allowing for misclassifications.

In the current study we employed the SVM-Light software package [38]. Using the heuristic C value computed by SVM-Light, we obtained a low discrimination performance of SVM trained with all features/voxels of the GM image segments: AUC=0.63.

Therefore we try to improve it by a C-value optimization procedure. This parameter was varied as follows:

$$C = 2^k \qquad \text{where} \quad k = -19, \ldots, 10. \tag{7.1}$$

Then, we selected the one which gives the maximum AUC. In this way we obtained AUC_{opt}=0.63. Therefore, the AUC already reached the maximum value when the heuristic C value was used.

7.2 Discrimination maps and Support Vector Machine Recursive Feature Elimination

By employing linear kernel SVMs to distinguish between two groups, the weight vector ($\mathbf{w}$) normal to the separating hyperplane is computed. Then, we can use this vector to generate a map of the most discriminating regions, so called discrimination map. Because the SVM classifier is naturally multivariate, this map can be interpreted as the spatial pattern by which the groups differ. The discrimination map has been evaluated here by employing the optimized C value.

In order to identify the set of voxels holding the highest discrimination power, hence to achieve a better AUC, we implemented the SVM recursive feature elimination (SVM-RFE) procedure [42, 43]. This feature-selection method has been implemented by iteratively removing from the data set those features which have absolute value of weight vector component ($|w_i|$) under a threshold T_j. The last is defined as follows:

$$T_j = min|w_i| + j(max|w_i| - min|w_i|)/N, \qquad \text{with} \quad j = 0, \ldots, N, \tag{7.2}$$

and N indicates how finely the AUC versus the number of retained voxels is sampled [40]. The LPO-CV has been implemented on the retained data, optimizing on the C value at each step j, hence has been possible to give an estimate of the classification performance of the SVMs in correspondence of the reduced set of features/voxels.

The behavior of AUC versus the number of retained voxels is shown in figure 7.1. Clearly this procedure allows an improvement of the classification performance, in fact we obtained a much better maximum value: AUC_{max}=0.9. We chose the point of the figure in which AUC=0.87, therefore the relative difference between AUC and AUC_{max} was about 3%, to achieve a compromise between the number of retained voxels and the classification performance. In the selected point the number of voxels used for the classification is 1876, corresponding to the 0.4% of the total amount of GM voxels.

The discrimination maps obtained at the corresponding threshold value on $|w_i|$ encode the anatomical information about the higher discriminating sets of voxels. Given two groups, patients and controls, with labels +1 and −1, respectively, a positive value in the discrimination map means relatively higher GM matter volume in patients than in controls and a negative value means relatively higher GM volumes in controls than in patients [39]. The regions where GM is greater in ASD subjects with respect to controls are shown in figure 7.2, whereas the superimposition of the latter and the regions where the weights w_i have negative values is reported in figure 7.3. These results have been obtained with AUC=0.87. The best of the involved brain areas, including the BA information and both the MNI and Talairach coordinates are reported for positive and negative results in tables 7.1 and 7.2, respectively.

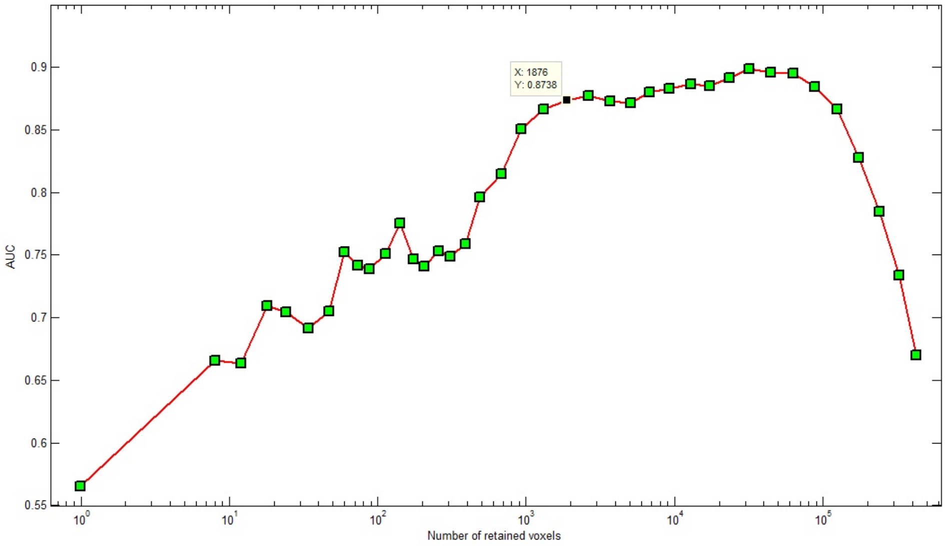


Figure 7.1: AUC versus the number of voxels with values of $|w_i|$ over threshold T_j, considered for the SVM classification of the GM segments. The maximum AUC value ($\mathrm{AUC_{max}}$=0.9) is obtained considering about 32000 voxels (7.5 % of GM voxels) in the SVM training. The point marked represents the conditions chosen for the analysis (AUC=0.87, number of retained voxels 1876, about the 0.4 % of GM voxels).

Table 7.1: Brain areas where GM is greater in ASD group with respect to control group (AUC=0.87).

		Number of voxels	MNI coordinates (x,y,z)	Talairach coordinates (x,y,z)
Temporal lobe, Inferior temporal gyrus, BA 20	R	40	$(56, -29, -19)$	$(51, -27, -15)$
Cerebellum, Posterior lobe	R	29	$(35, -69, -13)$	$(31, -65, -13)$
Temporal lobe, Fusiform gyrus, BA 37	L	48	$(-43, -62, -6)$	$(-41, -59, -7)$
Temporal lobe, Sub-gyral, BA 37	R	41	$(53, -51, -9)$	$(48, -49, -7)$
Temporal lobe, Superior temporal gyrus, BA 22	L	78	$(-52, -19, -8)$	$(-49, -19, -6)$
Temporal lobe, Superior temporal gyrus, BA 41	R	77	$(49, -38, 7)$	$(44, -38, 8)$
Occipital lobe, Middle occipital gyrus, BA 19	L	56	$(-27, -81, 18)$	$(-26, -79, 12)$
Frontal lobe, Superior frontal gyrus, BA 10	L	253	$(-26, 52, 4)$	$(-25, 47, 12)$
Frontal lobe, Superior frontal gyrus, BA 10	R	79	$(28, 50, 10)$	$(25, 44, 18)$
Temporal lobe, Superior temporal gyrus, BA 39	R	74	$(50, -52, 24)$	$(45, -53, 22)$
Occipital lobe, Precuneus, BA 31	L	87	$(-27, -69, 31)$	$(-26, -68, 25)$
Frontal lobe, Middle frontal gyrus, BA 9	L	25	$(-25, 42, 23)$	$(-25, 36, 28)$
Parietal lobe, Precuneus, BA 31	R	116	$(8, -49, 32)$	$(6, -50, 29)$
Parietal lobe, Precuneus, BA 31	L	80	$(-6, -45, 34)$	$(-7, -47, 30)$
White matter, Parietal lobe, Sub-gyral	R	16	$(35, -38, 42)$	$(30, -41, 39)$

Table 7.2: Brain areas where GM is greater in control group with respect to ASD group (AUC=0.87).

		Number of voxels	MNI coordinates (x,y,z)	Talairach coordinates (x,y,z)
Occipital lobe, Fusiform Gyrus, BA 18	L	49	$(-24, -99, -5)$	$(-24, -94, -10)$
Temporal lobe, Inferior temporal gyrus, BA 37	L	33	$(-53, -57, -4)$	$(-50, -54, -6)$
Temporal lobe, Inferior temporal gyrus, BA 37	R	340	$(50, -52, 4)$	$(45, -51, 4)$
Parietal lobe, Precuneus, BA 7	R	15	$(31, -66, 33)$	$(27, -66, 28)$
Parietal lobe, Angular gyrus, BA 39	L	65	$(-31, -56, 38)$	$(-31, -57, 33)$
Frontal lobe, Middle frontal gyrus, BA 6	L	171	$(-38, 14, 42)$	$(-37, 8, 42)$
Parietal lobe, Inferior parietal lobule, BA 40	R	15	$(54, -37, 45)$	$(48, -40, 42)$
Frontal lobe, Superior frontal gyrus, BA 6	R	11	$(18, 35, 47)$	$(15, 27, 50)$
Parietal lobe, Precuneus, BA 7	R	57	$(23, -56, 54)$	$(20, -59, 48)$

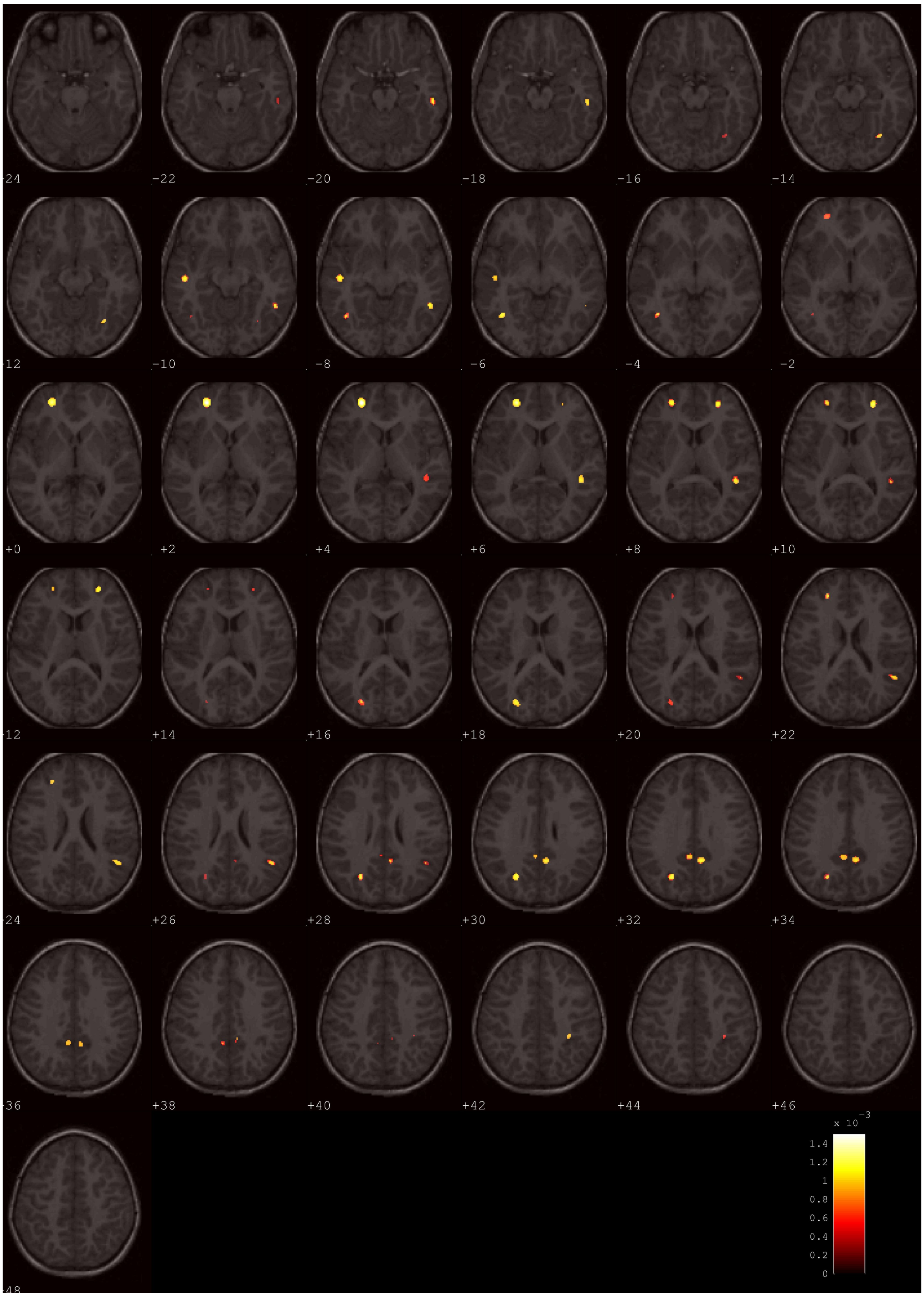


Figure 7.2: Discrimination map overlaid to a representative structural MR image. The voxels in red scale are the ones holding the highest discrimination power (AUC=0.87) when the number of retained voxels is about the 0.4 % of the total amount of GM voxels, and correspond to the brain areas where GM is greater in ASD group with respect to controls.

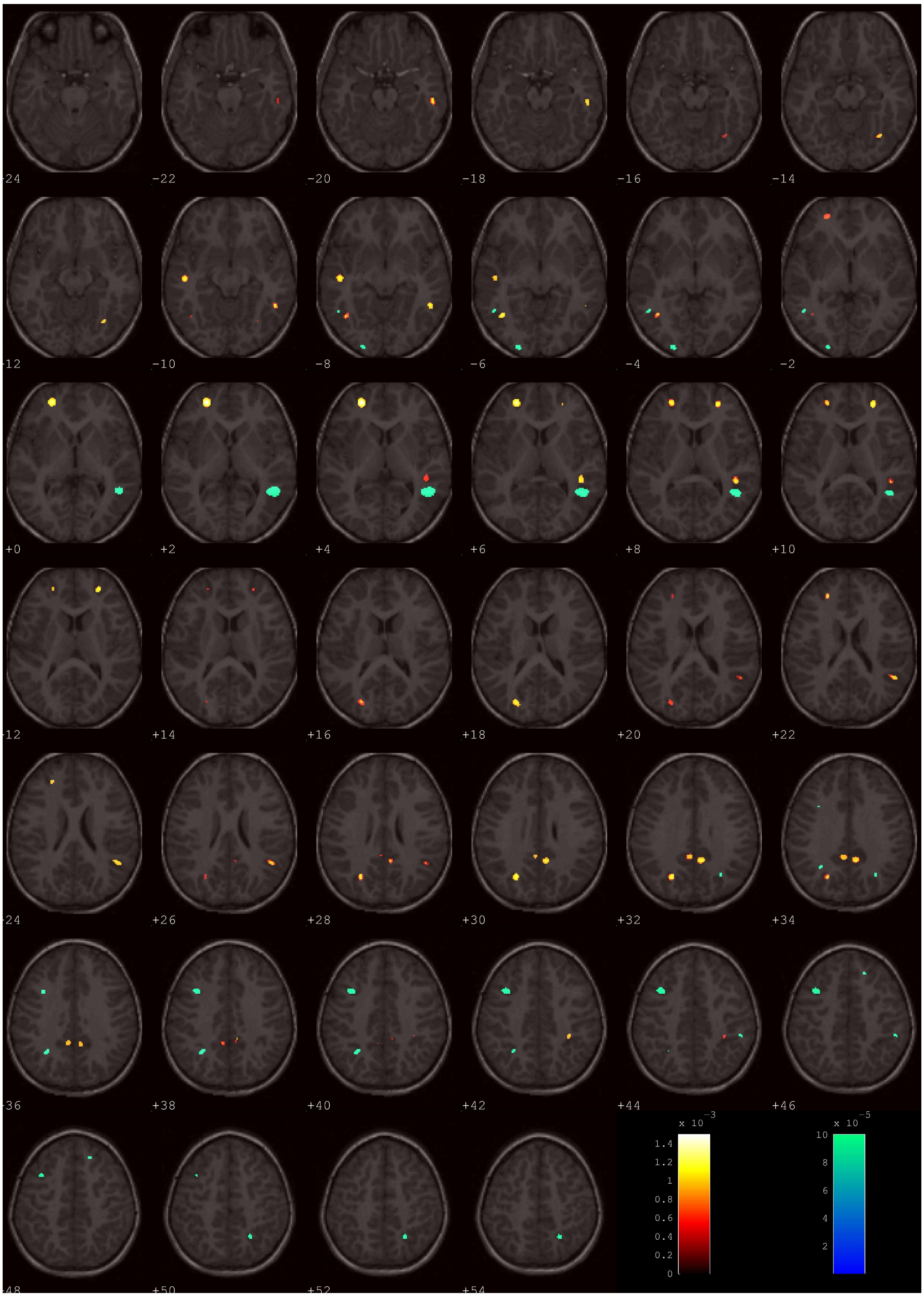


Figure 7.3: Discrimination map overlaid to a representative structural MR image. The voxels in red and blue scales are the ones with the highest discrimination power (AUC=0.87), when the number of retained voxels is about the 0.4 % of the total amount of GM voxels. Voxels in red scale correspond to the brain areas in which GM is greater in ASD group with respect to controls, instead voxels in blue scale represent the brain areas in which GM is greater in controls with respect to ASD group.

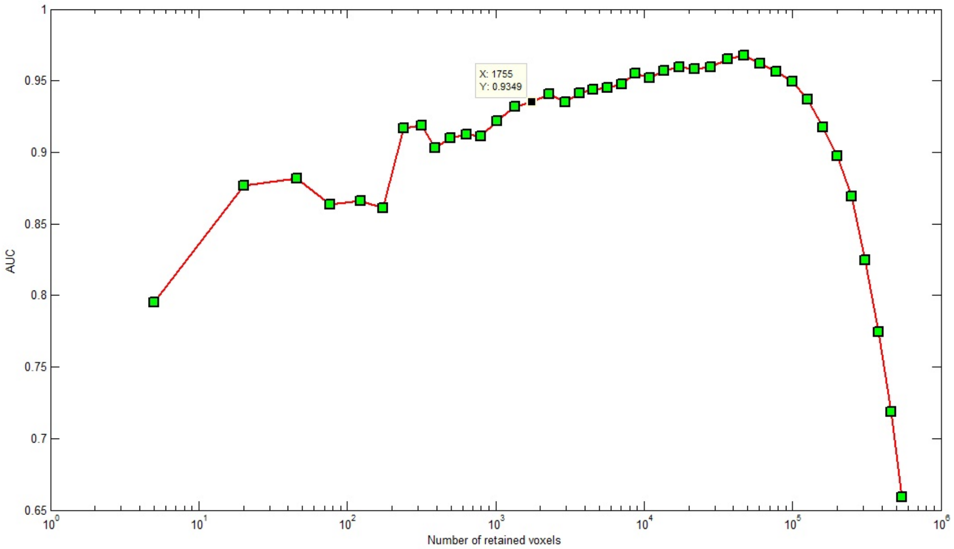


Figure 7.4: AUC versus the number of voxels with values of $|w_i|$ over threshold T_j, considered for the SVM-RFE classification GM segments of male subjects. The maximum AUC value ($\mathrm{AUC_{max}}$=0.97) is obtained considering about 46970 voxels (8.6 % of GM voxels) in the SVM training. The point marked represents the conditions chosen for the analysis (AUC=0.93, number of retained voxels 1755, about the 0.3 % of GM voxels).

7.3 Pattern classification of male subgroup grey matter segments

The same procedure described above in the case regarding the entire data set has been applied to the reduced data set of male subgroup with the aim of detecting those regions typical of male subjects.

We firstly evaluated the linear kernel SVMs classification performance adopting the heuristic C value, thus obtaining AUC=0.58.

Then, in order to achieve a more suitable classification accuracy we employed the C value optimization procedure. Then, in correspondence of the optimum C value we obtained: $\mathrm{AUC_{opt}}$=0.68.

The resulting optimized C value was used to construct the discrimination map. The last was employed to actuate the SVM-RFE procedure with the C optimization at each step j. The resulting AUC with respect to the number of retained voxels at each step is displayed in figure 7.4. The best classification performance ($\mathrm{AUC_{max}}$=0.97) is obtained when the number of retained features is about the 8.6 % of the total amount of GM voxels. Nevertheless, in order to employ a number of voxels comparable to those used in the case of SVM-RFE applied to the entire data set, we chose as operative point the one with the 0.3% of the entire amount of voxels, achieving to AUC=0.93. The resulting significative regions that contain greater GM in ASD with respect to controls group are listed in table 7.3 and shown in figure 7.5.

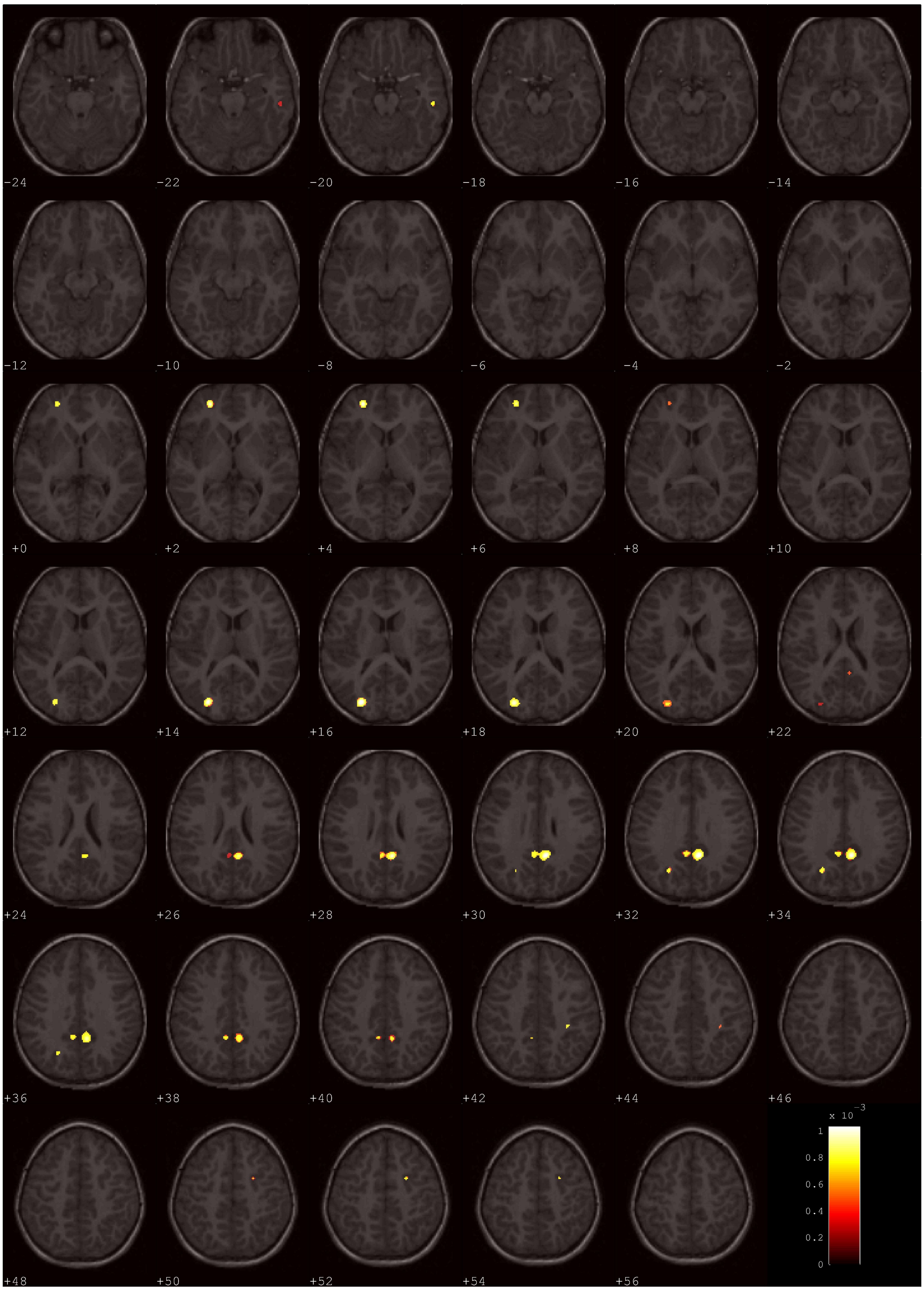


Figure 7.5: Male discrimination map overlaid to a representative structural MR image. The voxels in red scale are the ones with the highest discrimination power (AUC=0.93), when the number of retained voxels is about the 0.3% of the total amount of GM voxels, and correspond to the brain areas where GM is greater in ASD group with respect to control group.

Table 7.3: SVM-RFE resulting brain areas where male GM is greater in ASD group with respect of controls group obtained with AUC=0.93.

		Number of voxels	MNI coordinates (x,y,z)	Talairach coordinates (x,y,z)
Temporal lobe, Inferior temporal gyrus, BA 20	R	19	$(56,-26,-20)$	$(51,-25,-15)$
Frontal lobe, Medial frontal gyrus, BA 9	L	125	$(-27,52,3)$	$(-25,46,11)$
Occipital lobe, Middle occipital gyrus, BA 19	L	215	$(-29,-84,16)$	$(-28,-82,11)$
Limbic lobe, Cingulate gyrus, BA 31	R	624	$(4,-48,32)$	$(2,-49,29)$
Parietal lobe, Precuneus, BA 7	L	46	$(-26,-66,34)$	$(-26,-66,28)$
Frontal lobe, Sub-gyral, BA 6	R	15	$(24,2,52)$	$(21,-5,51)$

7.4 Pattern classification of female subgroup grey matter segments

The same procedure described above in the cases regarding the entire data set and after used for male subgroup has been applied to the reduced data set of female subgroup with the aim of detecting those regions typical of female subjects.

We firstly evaluated the linear kernel SVMs classification performance adopting the heuristic C value, thus obtaining AUC=0.58.

In order to achieve a more suitable classification accuracy we employed the C value optimization procedure and we employed the optimum C value to construct the discrimination map obtaining AUCopt=0.62.

With the purpose of identifying the voxels having the highest discrimination power we implemented the SVM-RFE procedure in the same way discussed above, applying the C value optimization procedure at each step j. The behavior of AUC versus the number of retained voxels is shown in figure 7.6. It was chosen the point where AUC had the maximum value: AUC_{max}=0.83 with 352 voxels, about the 0.07 % of the total number of GM voxels. In table 7.4 are listed the blobs obtained with AUC_{max} (see also figure 7.7). In this case we chose to display the results relative to so few retained features in order to preserve a good classification accuracy; in fact for the female subgroup, the AUC is significatively lower than those obtained for the entire data set and for the male subgroup. This finding is not surprising as the studies investigating volumetric brain differences within female ASD subjects and matched controls suggest various and largely unreplicated findings [40].

Table 7.4: SVM-RFE resulting areas of the brain where female GM is greater in ASD group with respect of controls group (AUC=0.83).

		Number of voxels	MNI coordinates (x,y,z)	Talairach coordinates (x,y,z)
Temporal lobe, Superior temporal gyrus, BA 22	L	14	$(-53,-21,-8)$	$(-50,-20,-6)$
Frontal lobe, Superior frontal gyrus, BA 10	R	47	$(27,51,11)$	$(24,45,19)$
Occipital lobe, Precuneus, BA 31	L	66	$(-28,-73,26)$	$(-27,-72,20)$
Frontal lobe, Middle frontal gyrus, BA 9	L	150	$(-25,44,22)$	$(-24,37,27)$
Frontal lobe, Middle frontal gyrus gyrus, BA 6	L	67	$(-29,-6,51)$	$(-29,-11,48)$

7.5 Considerations

The optimization steps involved in the pattern classification approach to grey matter segments analysis described in this chapter have determined an enhancement of the classification performance. The principal improvements are due to the C value optimization application during the SVM classification and to the final combination between the recursive feature elimination algorithm and the C value optimization procedure performed at each features removal step.

In order to instantly realize these improvements, the AUC outline resulting from the whole data set and from the male and female subgroups pattern classifications is shown in

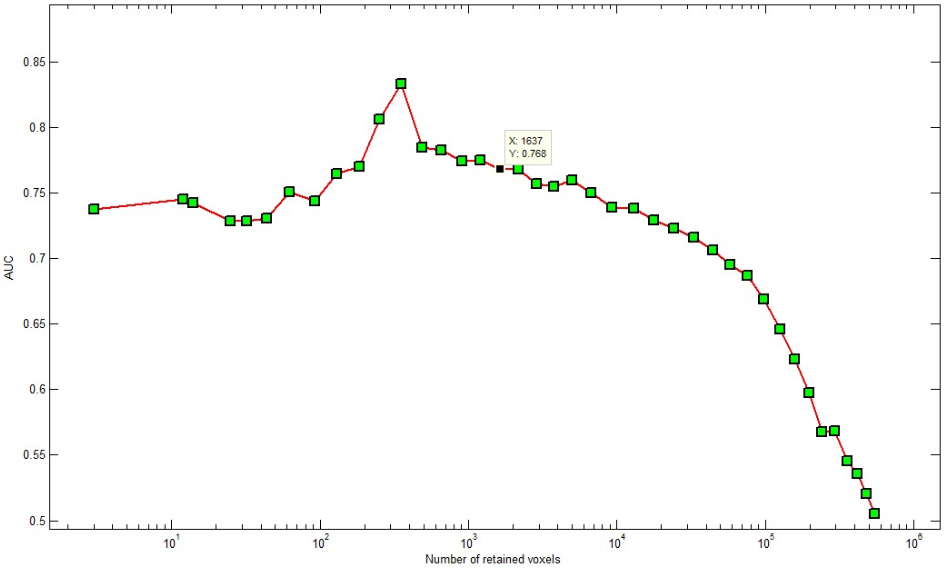


Figure 7.6: AUC versus the number of voxels with values of $|w_i|$ over threshold T_j, considered for the SVM classification of the female GM segments. The maximum AUC value (AUC_{max}=0.83) is obtained when 352 is the number of retained voxels (about 0.07 % of the entire amount of GM voxels) in the SVM training.

figure 7.8. Step 1 regards SVM with the heuristic C value employed, step 2 shows the AUC in correspondence of the optimized C value (AUC_{opt}) and step 3 gives the AUC_{max} resulting from the SVM-RFE application combined with the C value optimization procedure.

Evidently, a greater classification accuracy is achievable for male subgroup with respect to the female one. Although the larger number of subjects contained in the whole group case training set, thus leading to a potentially better generalization ability, a greater discriminating performance is obtained for male subgroup with respect to the entire data set pattern classification.

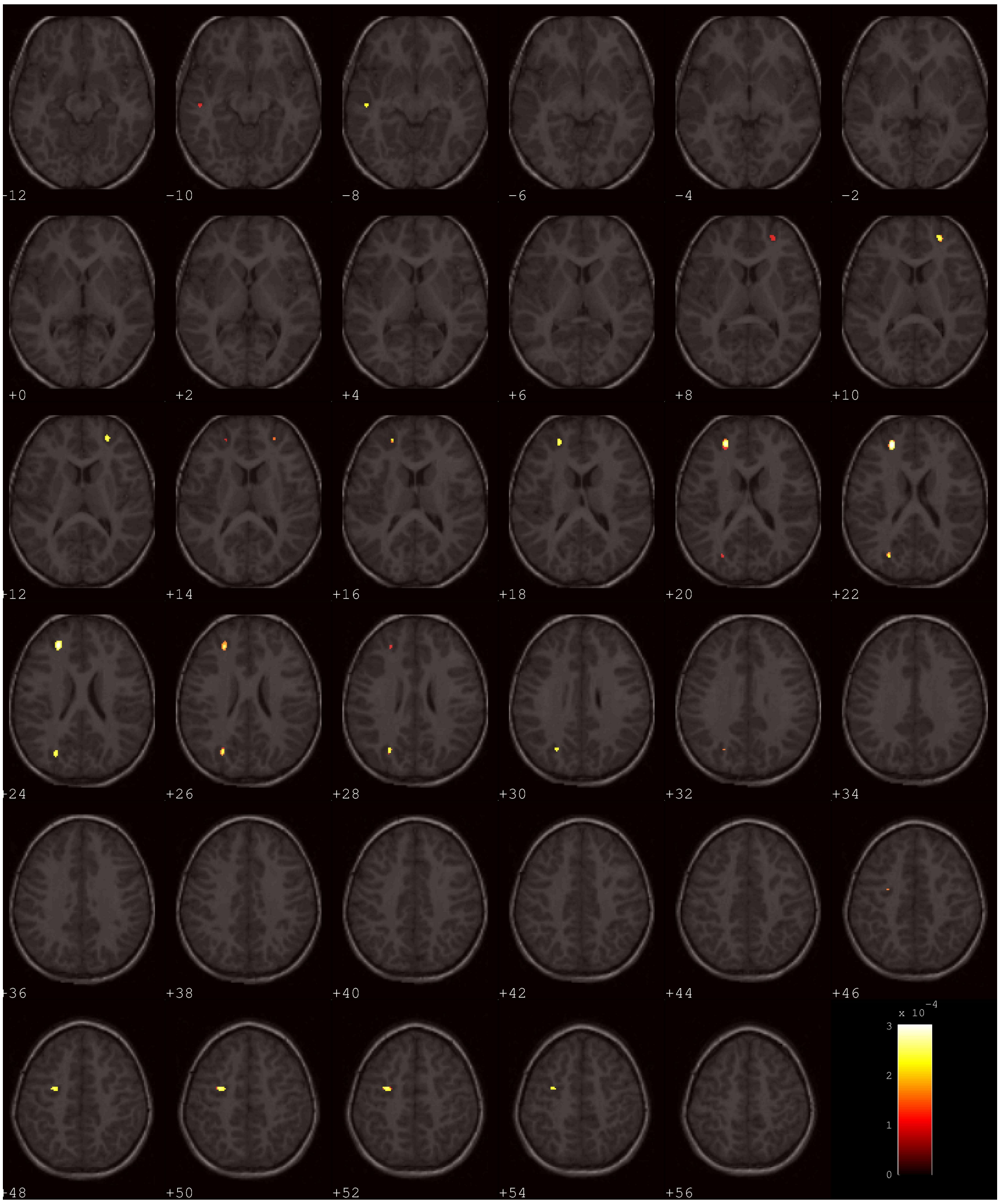


Figure 7.7: Female discrimination map overlaid to a representative structural MR image. The voxels in red scale are the ones with the highest discrimination power (AUC= 0.83), when the number of retained voxels is about the 0.07% of the total amount of GM voxels, and correspond to the areas of the brain where GM is greater in ASD group with respect to control group. Instead, the point indicated in figure (1637 retained voxels) is the one chosen to obtain the discrimination map shown in figure 7.6.

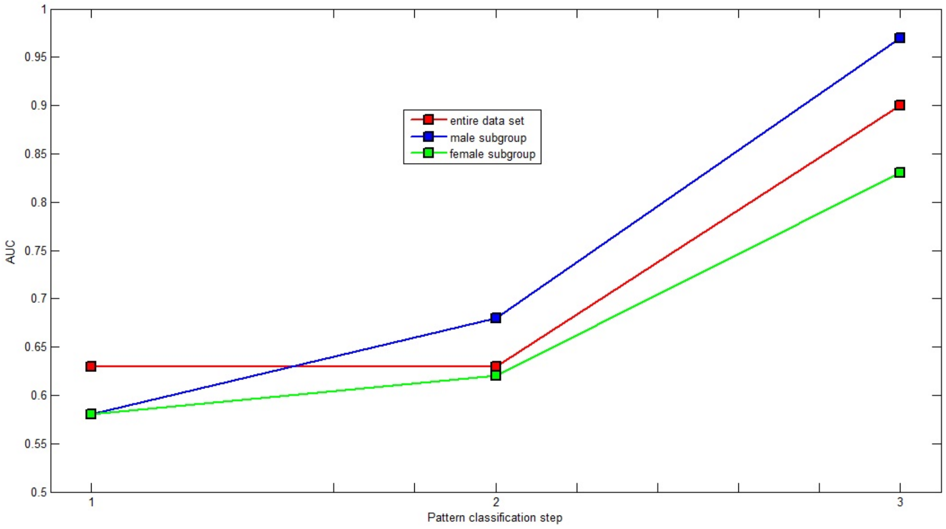


Figure 7.8: AUC versus pattern classification step in the whole group and male and female subsets cases.Step 1 regards SVM with the heuristic C value employed, step 2 shows the AUC in correspondence of the optimized C value ($\mathrm{AUC}_{\mathrm{opt}}$) and step 3 gives the $\mathrm{AUC}_{\mathrm{max}}$ resulting from the SVM-RFE application combined with the C value optimization procedure.

7.6 SVM-RFE applied to preselected grey matter regions of interest

As we have already seen in the current chapter, the employment in our study of the SVM recursive feature elimination algorithm with the aim of classifying ASD and control subjects has improved the classification accuracy. Since it is believed that removing non-informative signal can reduce noise, and can increase the contrast between labelled groups, several feature selection methods are being developed with the purpose of achieving greater predictive accuracies. Nevertheless, not all the feature selection approaches lead to greater classification performances, for example the obtained results depend on the sample size, as discussed in [48], and there are still many questions to answer in this context.

One type of feature selection method is based on using prior knowledge about some interesting regions where neuroanatomical abnormalities of patients with respect to controls can be contained. Such a region of interest (ROI) approach have been employed in the current study with the aim of looking for those brain regions where GM is greater in ASD group with respect to matched controls, when the analysed brain regions are a priori selected by a medical expert. Thus, we could compare the results obtained with a ROI-based analysis with those achieved by a whole-brain SVM approach, with the aim of understanding if there are any brain areas that mostly contribute to the classification when not all the GM voxels are given as input features to the SVMs classifier.

In order to actuate this study we selected 18 ROIs of grey matter from those defined in the Laboratory of Neuro Imaging (LONI) Probabilistic Brain Atlas (LPBA40) (`http://www.loni.ucla.edu/Atlases/LPBA40`) [49], which were generated from anatomic regions of MRI delineated manually from 40 subjects. There are 56 structures in the atlas. A pair of figures showing the brain regions that have been considered in our analysis are 7.9 and 7.10.

Our selected ROIs were overlayed into a unique image and affine registered to our study-specific DARTEL-created template, in order to create a mask to apply on the GM prepro-

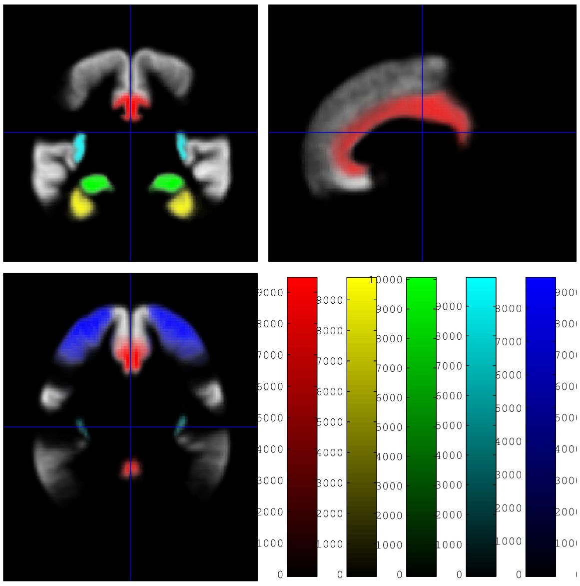


Figure 7.9: 10 selected ROIs of grey matter defined in LONI Probabilistic Brain Atlas: L and R Cingulate Gyri (red scale), L and R Fusiform Gyri (yellow scale), L and R Hippocampi (green scale), L and R Insular Cortices (cian scale), L and R Middle Frontal Gyri (blue scale).

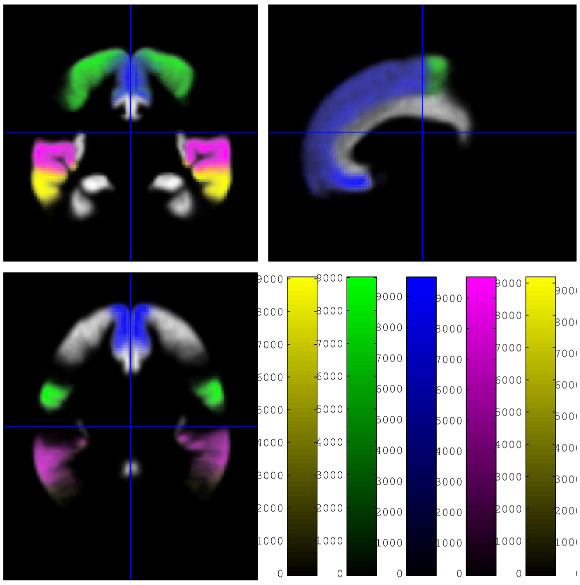


Figure 7.10: 8 selected ROIs of grey matter defined in LONI Probabilistic Brain Atlas: L and R Middle Temporal Gyri (yellow scale), L and R Precentral Gyri (green scale), L and R Superior Frontal Gyri (blue scale), L and R Superior Temporal Gyri (magenta scale).

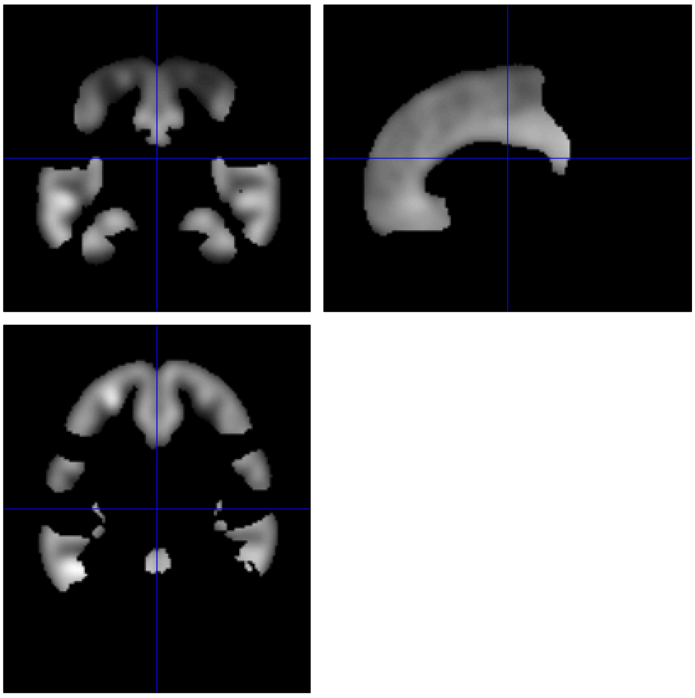


Figure 7.11: An example of GM preprocessed image scan that has been masked on the basis of the selected ROIs.

cessed image segments of our entire data set. In this way we obtained for each initial preprocessed GM scan a subset of grey matter voxels. An example of masked grey matter scan is reported in figure 7.11.

These reduced normalized modulated and smoothed GM image segments, obtained according to ROI-based selection, were given as input to a SVM classifier. The leave-pair-out cross-validation procedure was employed to evaluate the classification accuracy. By using the heuristic C value we evaluated an AUC of 0.582. Then, we decided to optimize this parameter. The optimization procedure leaded to a better classification accuracy: $\mathrm{AUC}_{\mathrm{opt}}=0.692$. The optimized C value was used to build the discrimination map that had to be used within the SVM-RFE algorithm. Using 20 threshold values depending on eq. (7.2), and optimizing on C at each feature elimination step we obtained the curve of AUC versus the number of retained voxels that is shown in figure 7.12. The maximum AUC was $\mathrm{AUC}_{\mathrm{max}}=0.868$, reached when the number of retained features is about the 10 % of the total amount of GM voxels. By choosing as operative point the one containing about the 1 % of the total amount of GM voxels leading to the analysed ROIs, we have obtained the regions showed in figure 7.12 and listed in table 7.5, where the GM of ASD group was greater than that of controls with AUC=0.844. The resulting regions coincide with those already detected in the whole-brain SVM-RFE analysis of the entire data set excepted for the last, that has been obtained only for the whole-brain SVM-RFE analysis of female subgroup. These results are consistent with those regarding the whole-brain pattern classification, but, using the same number of retained voxels we have collected only 8 of the 15 regions of the whole-brain plus another that we have proved to be characteristic of female ASD subjects.

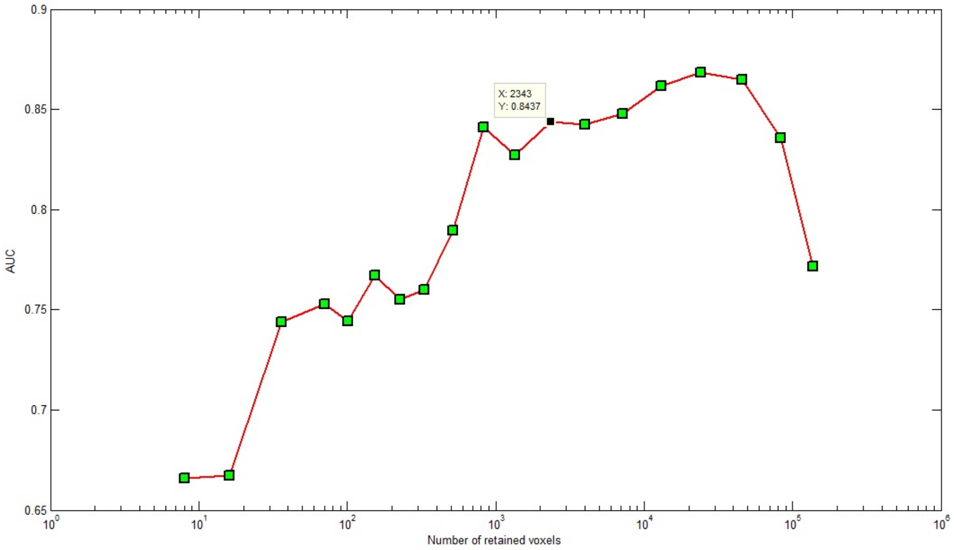


Figure 7.12: AUC versus the number of retained voxels curve, achieved by applying SVM-RFE-ROI-based analysis. The maximum AUC was AUC_{max}=0.868, reached when the number of retained features is about the 10 % of the total amount of GM voxels. Instead, the point indicated (about 1% of total GM voxels), was chosen as operative condition corresponding to AUC=0.844.

Table 7.5: Brain areas where GM is greater in ASD group with respect to control group when SVM-RFE-ROI-based analysis is applied to the entire data set. (AUC=0.844).

		Number of voxels	MNI coordinates (x,y,z)	Talairach coordinates (x,y,z)
Temporal lobe, Inferior temporal gyrus, BA 20	R	91	$(56, -30, -19)$	$(51, -28, -15)$
Temporal lobe, Superior temporal gyrus, BA 39	R	120	$(55, -52, -8)$	$(49, -50, 7)$
Temporal lobe, Superior temporal gyrus, BA 22	L	127	$(-52, -20, -8)$	$(-49, -19, -6)$
Temporal lobe, Fusiform gyrus, BA 37	L	17	$(-44, -62, -6)$	$(-41, -59, -7)$
Frontal lobe, Superior frontal gyrus, BA 10	L	513	$(-26, 51, 7)$	$(-25, 45, 14)$
Temporal lobe, Superior temporal gyrus, BA 41	R	95	$(48, -37, 7)$	$(44, -37, 7)$
Frontal lobe, Superior frontal gyrus, BA 10	R	138	$(28, 51, 10)$	$(24, 44, 18)$
Parietal lobe, Precuneus, BA 31	R	408	$(3, -46, 34)$	$(1, -48, 30)$
Frontal lobe, Middle frontal gyrus, BA 6	L	106	$(-30, -6, 50)$	$(-29, -12, 48)$

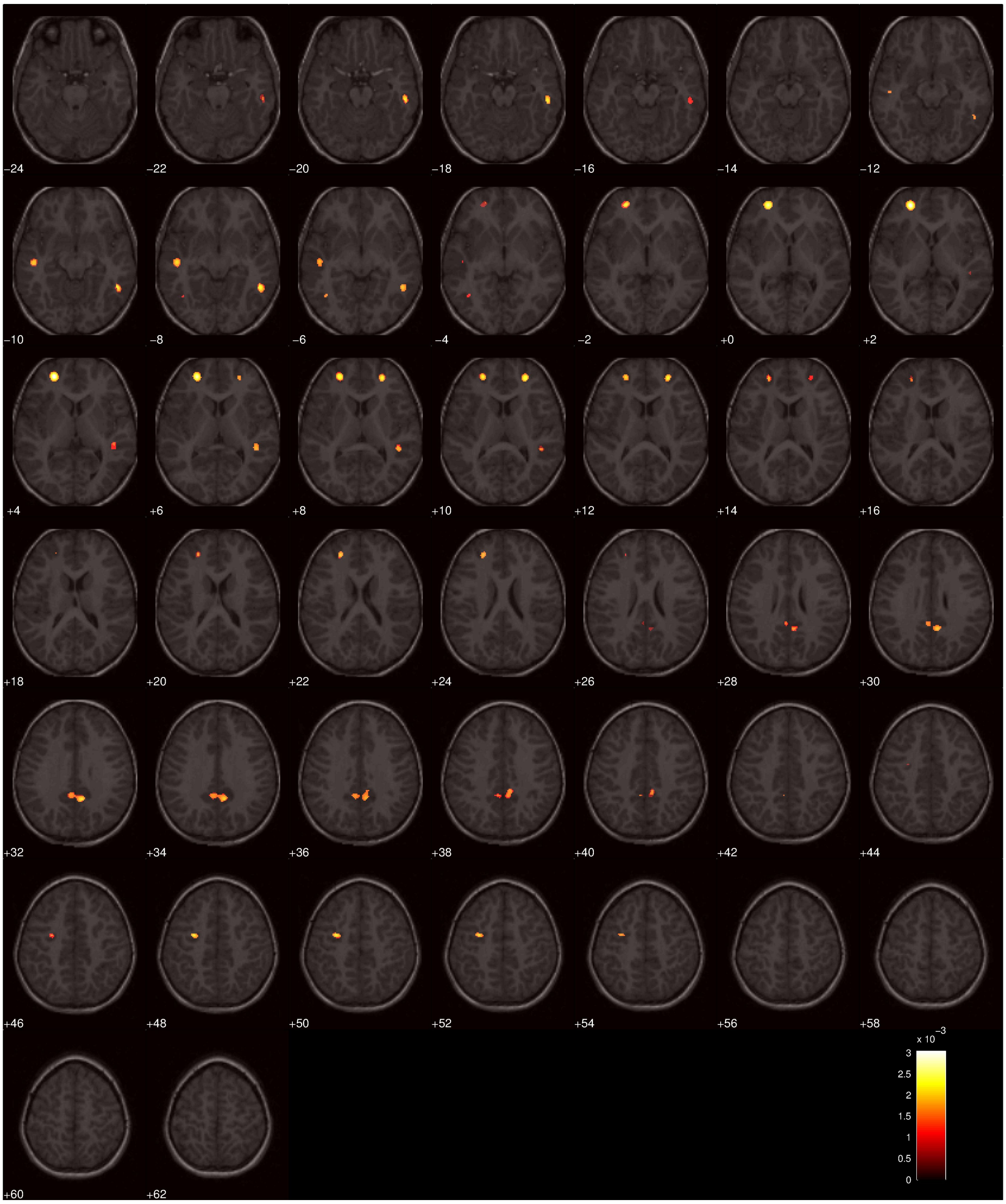


Figure 7.13: Discrimination map obtained from ROI-based analysis overlaid to a structural MR scan. The regions in red scale indicate those brain areas where the GM of ASD group is grater with respect to controls. This map corresponds to AUC=0.844.

Chapter 8

Impact of the training set on the SVM analysis

With the aim of strengthening the SVM-based methods applied to brain data and to guarantee reliability of the results, we assessed the dependency on the population of the training set within the cross-validation procedure. In this way we became able to check for the stability of our results both with respect to the subject chosen to obtain the discrimination maps in the SVM-RFE procedure and with respect to the number of subjects employed during the learning phase. Furthermore, we can evaluate the classification performances for different cross-validation schemes.

8.1 SVM discrimination map update approach

The SVM algorithm that has been implemented in Chapter 7 built the discrimination map by randomly extracting the support vector parameters obtained from one of the LPO-CV steps. Such a procedure, was considered to be reliable since the weight vectors obtained from different cross-validation iterations seemed to be not too distinct one from each other. Nevertheless, a more accurate investigation can be performed, in order to check for the reliability of our above reported results. For this reason, we decided to implement a SVM discrimination map update approach on the entire data set.

Firstly, we collected all the 76 discrimination maps, each obtained from one of the SVM LPO-CV steps. Then, we employed each of these maps to start a SVM-RFE procedure, with the purpose of finally verifying the convergence of all the 76 AUC versus the number of retained features outlines in our region of interest (about under 5000 voxels). We have now to remark that the recursive feature elimination algorithm applied in the above described approach was characterized by a number of retained features at each elimination step that depended on $\min|w_i|$ and $\max|w_i|$ values. Hence, if we change the initial weight vector, then we also modify the number of retained voxels at each SVM-RFE step. Therefore, if we did not apply a different recursive feature elimination algorithm, we would not be able to compare all the discrimination maps at fixed features step. For this reason we implemented a new recursive procedure that at each elimination step removes the n less informative features, where n is fixed at each step but usually it is taken different from a step to each other. In addition, in each recursive step a discrimination map update was employed by using the support vectors obtained from the last LPO-CV iteration within the current SVM-RFE step.

In this way we become able to evaluate the mean and the standard deviation of the 76 discrimination maps at each recursive feature elimination step. Some examples of average maps and their standard deviations when the number of retained voxels is 144000 and 6000

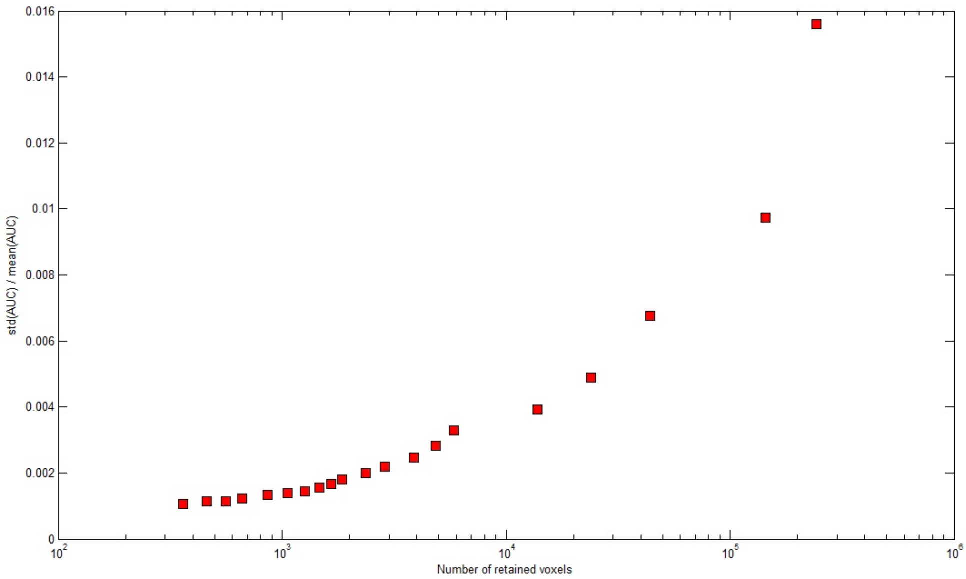
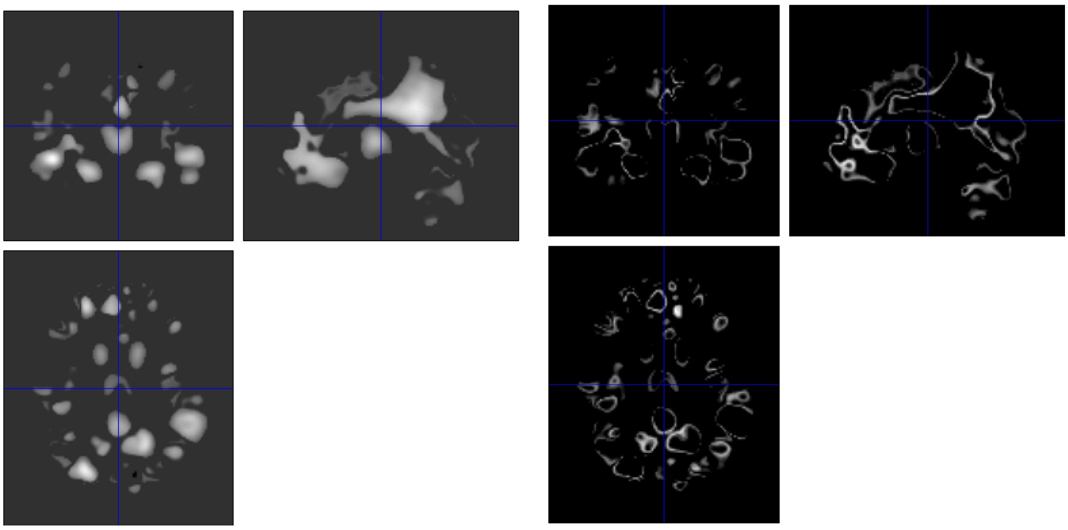


Figure 8.1: AUC relative standard deviation as a function of the number of retained voxels step. Standard deviation is evaluated on the 76 AUC values obtained at each SVM-RFE step from the 76 different initial discrimination maps. In this case the heuristic C value has been employed.

are shown in figure 8.2.

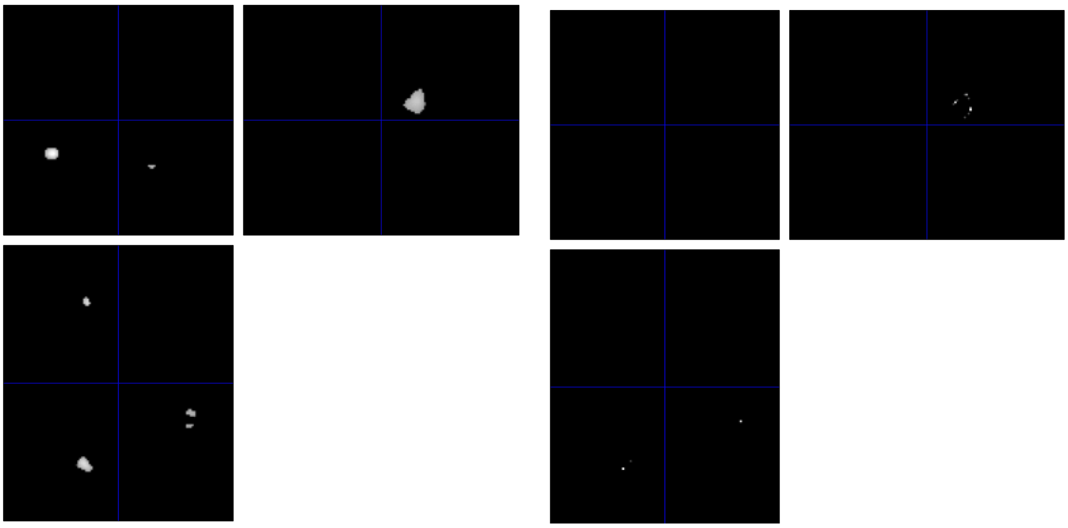
In the current situation, the SVM-RFE procedure was applied without C value optimization, in order to limit the execution time, thus the heuristic C value was employed, entailing a considerable reduction of the AUC. Nevertheless, such a rough procedure allowed us to realize the AUC versus the number of retained features curves convergence in our region of interest and to estimate a superior limit to the AUC variability in correspondence of a fixed number of employed features. Evidently, the AUC variability decreases when the number of features reduces and finally it becomes roughly constant, as shown in figure 8.1, where the relative standard deviation of AUC is reported as a function of the number of retained voxels.

A direct comparison between the discrimination maps (containing $|w_i|$ values) obtained in both the approaches (with and without discrimination map update) when the number of retained voxels is about 1800 is displayed in figures 8.3 and 8.4. From these images it is clear that several equal regions are detected in both the cases even if the discrimination map update approach has been implemented using the heuristic C value, thus the resulting AUC is considerable lower with respect to the one obtained by employing the C value optimization procedure.



(a) Mean of the discrimination maps at about 14400 retained voxels.

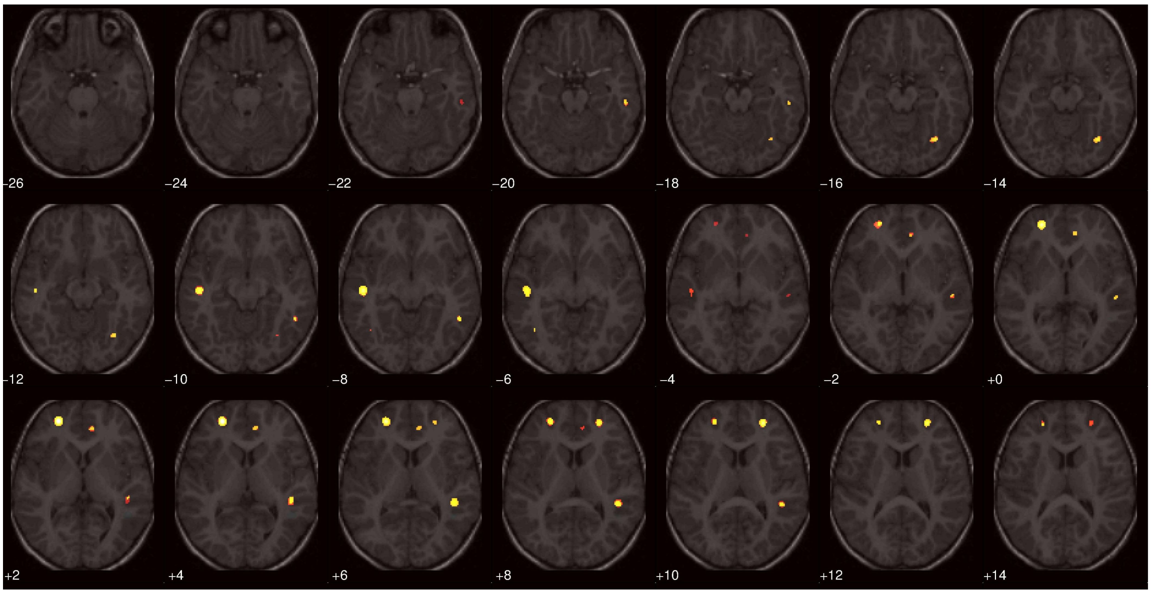
(b) Standard deviation of the discrimination maps at about 144000 retained voxels.



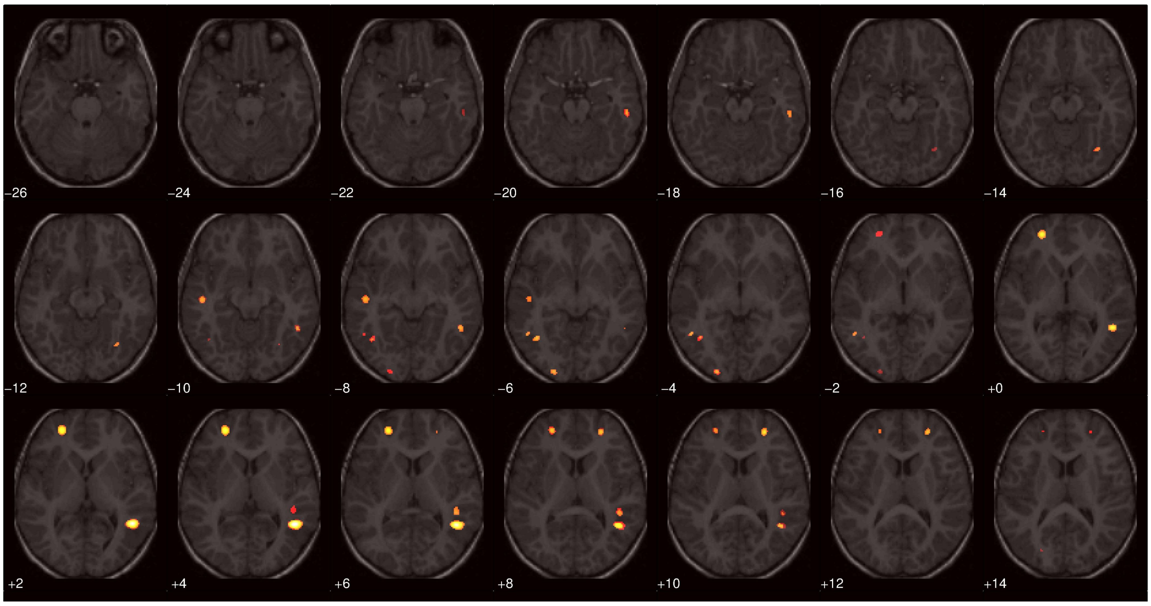
(c) Discrimination maps mean at about 6000 retained voxels.

(d) Standard deviation of the discrimination maps at about 6000 retained voxels.

Figure 8.2: Some examples of mean of discrimination maps (8.2a e 8.2c) and their standard deviations (8.2b and 8.3b) in correspondence of some fixed number of retained voxels.

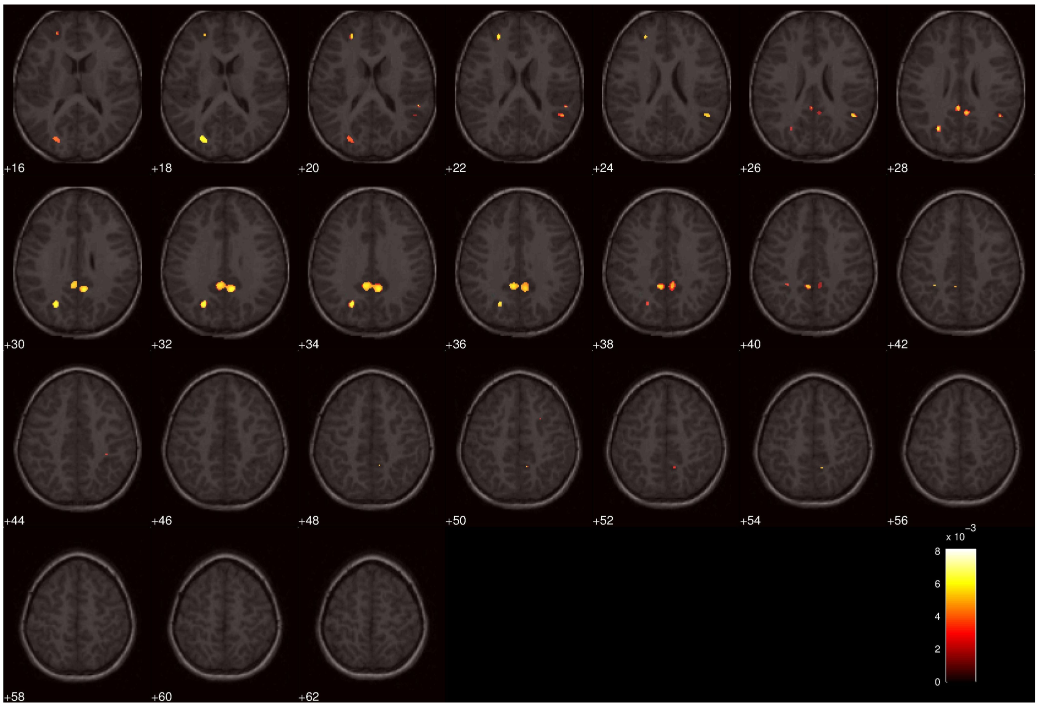


(a)

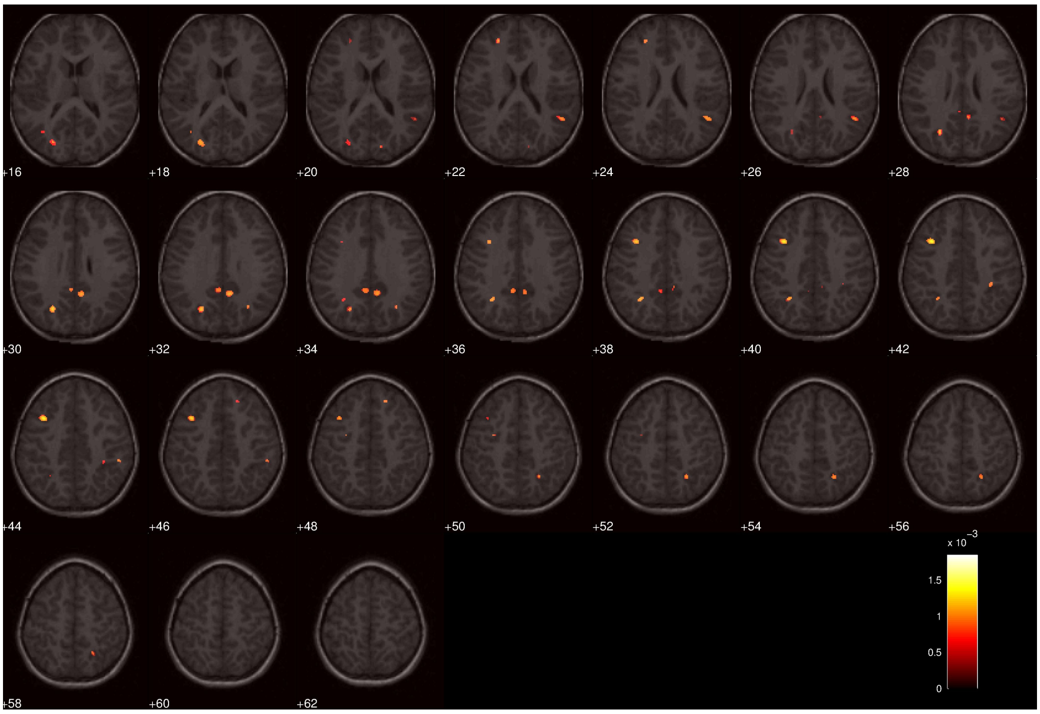


(b)

Figure 8.3: Discrimination maps overlaid to a representative structural MR image (axial coordinate in range (-26, +14)). The voxels in red scale represent the features ($|w_i|$) with the highest discrimination power between ASD group and controls. Figure 8.3a shows the mean of the discrimination maps within the update approach obtained when the number of retained voxels is 1861 and AUC=0.68 (here, the heuristic C value has been employed). Instead, figure 8.3b displays the discrimination map ($|w_i|$) obtained without weight vector update when the number of retained voxels is 1876 and AUC=0.87. The last discrimination map has already been shown in figure 7.3. The chromatic scale is that reported in figure 8.4.



(a)



(b)

Figure 8.4: Discrimination maps overlaid to a representative structural MR image (axial coordinate in range $(+16, +62)$). The voxels in red scale represent the features $(|w_i|)$ possessing the highest discrimination power between ASD group and controls. Figure 8.4a shows the mean of the discrimination maps within the update approach obtained when the number of retained voxels is 1861 and AUC=0.68 (here, the heuristic C value has been employed). Instead, figure 8.4b displays the discrimination map $(|w_i|)$ obtained without weight vector update when the number of retained voxels is 1876 and AUC=0.87. The last discrimination map has already been shown in figure 7.3.

8.2 Impact of the training set size on the SVM analysis

In all analyses reported in Chapter 7, the SVMs have been trained according to the leave-pair-out cross-validation (LPO-CV) procedure, thus excluding one couple of matched subjects (the matching was made according to sex, age and NVIQ) from the training set at each CV iteration and validating the trained SVM on it. Nevertheless we can study how the classification performance changes if we leave more than one couple of subjects at each iteration. In order to actuate this examination we decided to apply leave-2pairs-out, leave-4pairs-out and leave-19pairs-out CV procedures that correspond to divide the whole data set, composed by 76 couples of subjects, in 38, 19 and 4 folds of equal size respectively.

For example, within the leave-4pairs-out CV algorithm the entire data set is firstly randomly divided into 19 subgroups, each composed by 4 pairs of matched subjects. Then, at each cross-validation iteration one of these subgroups is used as validation set and the rest as training set. Since we want that the same instance, once selected, could not be chosen again and that training and validation sets are of the same size at each iteration, the leave-4pairs-out CV procedure will have 19 iterations. These considerations are summarized in table 8.1 for the last and for the others CV algorithms considered.

Table 8.1: Leave-pair-out CV when the number of pairs that are left out varies. In all the cases that are reported here we suppose of having an initial data set composed by 76 pairs of subjects.

CV-procedure	training set		validation set		n. of folds (CV iterations)
	n. of pairs	n. of subjects	n. of pairs	n. of subjects	
leave-1pair-out	75	150	1	2	76
leave-2pairs-out	74	148	2	4	38
leave-4pairs-out	72	144	4	8	19
leave-19pairs-out	57	114	19	38	4

We have to remark that within leave-one-pair-out CV procedure no random sampling was involved, instead, when more than one pair is left out we become able to evaluate the classification performance with respect to the initial random data sampling. In fact there is only one way to sample a group of n instances by picking one instance at a time without replacement for n times.

In this chapter we will report the results obtained by SVMs applied to the entire dataset by adopting leave-2pairs-out, leave-4pairs-out and leave-19pairs-out CV algorithms, in order to compare the obtained results with those relative to the leave-1pair-out CV (reported in Chapter 7), where the greatest possible amount of data was used for training at each iteration.

8.2.1 Leave-2pairs-out cross-validation

SVM classification

We began by evaluating the classification performance of SVMs in correspondence of randomly selected couples of matched subjects to compose training and validation sets within the leave-2pairs-out CV procedure. For this reason we decided to repeat the SVM classification 20 times, each with a different initial randomization of the data set and by employing the heuristic C value. In this way we obtained: AUC$=0.611 \pm 0.001$.

Hence, an optimization of this parameter seems to be necessary. But, in order to be sure that the optimized C value does not depend on the initial randomization, we decided to study how the computed optimized C value is related to the initial random sampling.

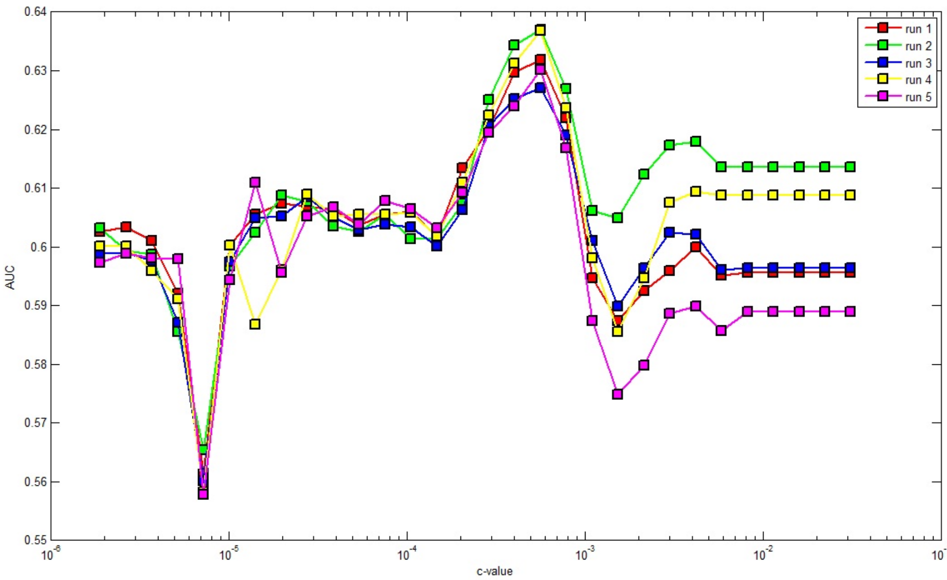


Figure 8.5: AUC versus C value in correspondence of 5 different random samplings, evaluated within leave-2pairs-out CV. The resulting optimized C value is that corresponding to the maximum AUC and is the same in all 5 runs.

In fact, we repeated the C value optimization algorithm in correspondence of 5 different randomly composed training and testing sets. The obtained results are shown in figure 8.5, where the AUC versus the employed C values are reported in the 5 random situations that have been analysed. The optimized C values resulted to be the same in all 5 cases, confirming that it only depends on the number of features considered for training, as shown in the following section. In this way we obtained $\text{AUC}_{\text{opt}}=0.634\pm0.003$.

SVM-RFE analysis

The resulting optimized C value was used to extract the discrimination map that has been employed to start the SVM-RFE algorithm with C value optimization at each step. It was implemented in such a way the number of retained features at each step was fixed. With the aim of verifying that the optimized C value at each recursive feature elimination step in correspondence of different initial random samplings of the data set does not change, we repeated the SVM-RFE algorithm with 5 fixed steps for 5 runs each different from the others for the initial randomization. The results are shown in figure 8.6, which displays the optimized C value versus the number of retained features for 5 different initial random samplings. Because the C value stability with respect to the randomization seems to decrease when a fewer number of retained features is considered, we decided to intensify the number of retained voxels sampling in the region containing less than 5000 features. Thus, we applied the SVM-RFE algorithm using 10 fixed steps for 6 runs. In this way we obtained the results reported in figure 8.7. We then concluded that the C value is roughly independent of the initial randomization.

We can apply the SVM-RFE algorithm keeping fixed the initial random sampling of the data set. In order to allow a direct final comparison between the results obtained from the application of SVM-RFE by employing different leave-pair-out CV approaches, we decided to implement a recursive feature elimination algorithm that removes a fixed number of voxels at each step, chosen in such a way the number of retained features sampling exactly corresponds to that obtained from SVM-RFE algorithm that has been applied to

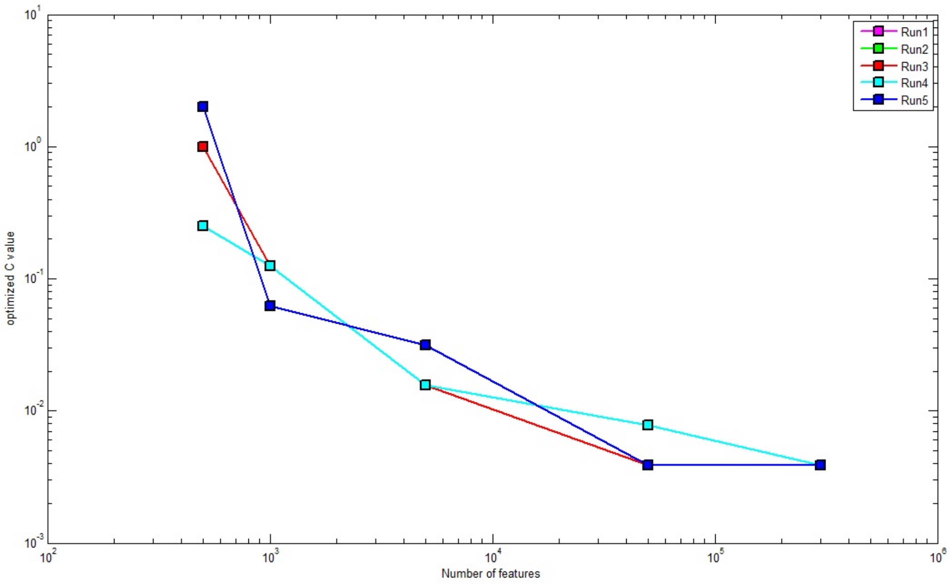


Figure 8.6: Optimized C value versus the number of retained features for 5 different initial random samplings, obtained from SVM-RFE with leave-2pairs-out CV procedure, when 5 steps are chosen.

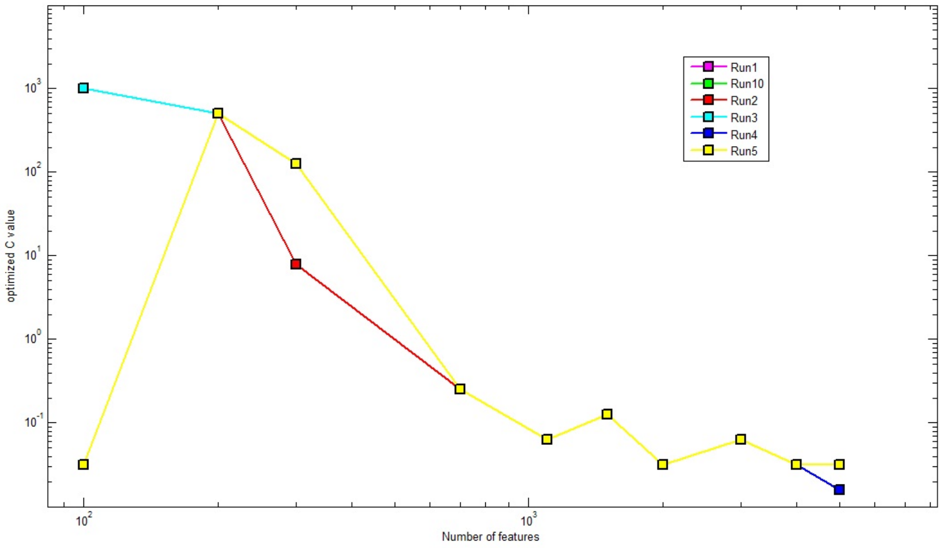


Figure 8.7: Optimized C value versus the number of retained features for 6 different initial random samplings, obtained from SVM-RFE with leave-2pairs-out CV procedure, when 10 steps are chosen (n. of retained voxels $<$5000).

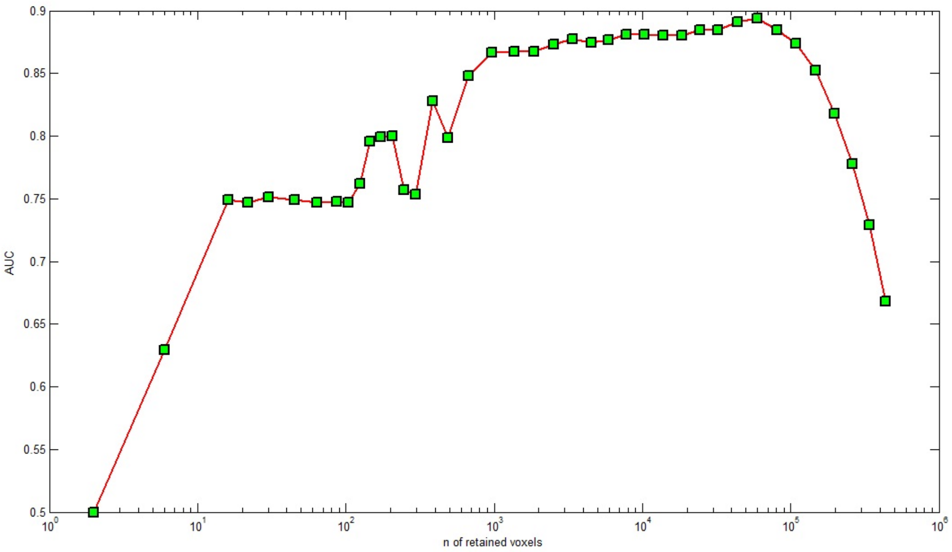


Figure 8.8: AUC versus the number of retained voxels, obtained within the SVM-RFE leave-2pairs-out CV. The maximum AUC was achieved when the number of retained voxel was about the 14% of the total amount of GM voxels (AUC_{max}=0.893).

the entire data set within the leave-1pair-out CV (see Section 7.2). Therefore we obtained the behaviour of AUC versus the number of retained voxels that is shown in figure 8.8. The maximum AUC was achieved when the number of retained voxel was about the 14% of the total amount of GM voxels (AUC_{max}=0.893).

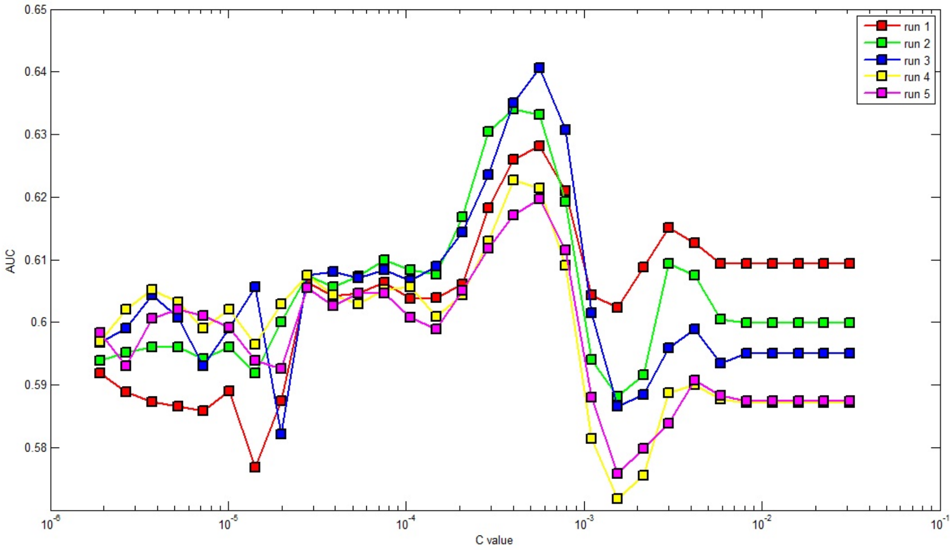


Figure 8.9: AUC versus C value in correspondence of 5 different random samplings, evaluated within leave-4pairs-out CV. The resulting optimized C value is that corresponding to the maximum AUC and is the same in all 5 runs.

8.2.2 Leave-4pairs-out cross-validation

SVM classification

We firstly evaluated the AUC achieved by applying the SVMs within the leave-4pairs-out cross-validation employing the heuristic C value and repeating the the cross-validation for 5 times, each different from the others for the initial randomization. In this way we obtained AUC=0.600 ± 0.002.

Then, we actuated the same optimization procedure described above, repeating it for 5 times by changing the initial randomization. The obtained results are shown in figure 8.9. Also in this case the optimized C value does not change when a different initial randomization is used. The classification accuracy obtained when the optimized C value is employed is $\mathrm{AUC}_{\mathrm{opt}}=0.629\pm0.008$.

SVM-RFE analysis

The above obtained optimized C value has been employed to build the discrimination map that has to be used within the recursive feature elimination procedure. The SVM-RFE algorithm with 5 fixed step was repeated now, as the starting map in the leave-4pairs-out cross-validation procedure in correspondence of 5 different randomly composed training and testing sets, with the aim of verifying that the optimized C value in correspondence of a certain number of retained features does not depend on the initial data set randomization. The results, expressed in terms of optimized C value with respect to the number of retained features are shown in figure 8.10. Even if sometimes the C value is not exactly the same for different runs, we can consider our procedure to be stable with respect to the initial randomization because the number of outliers is really poor.

Then, we have implemented the complete SVM-RFE procedure described in Section 8.2.1 by applying the C value optimization at each step, considering the resulting C value independent from the initial randomization. Therefore, we obtained the usual AUC versus number of retained features outline, that is displayed in figure 8.11. The maximum classification performance was obtained when the number of retained voxels was about the 4% of

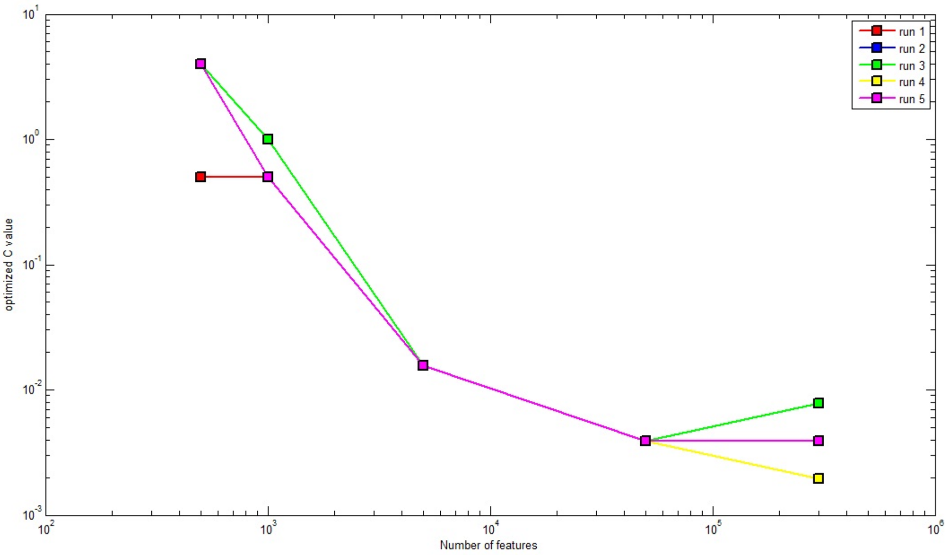


Figure 8.10: Optimized C value versus the number of retained features for 5 different initial random samplings, obtained from SVM-RFE with leave-4pairs-out CV procedure, when 5 steps are chosen.

the total amount of GM voxels with $\mathrm{AUC}_{\max}$=0.893.

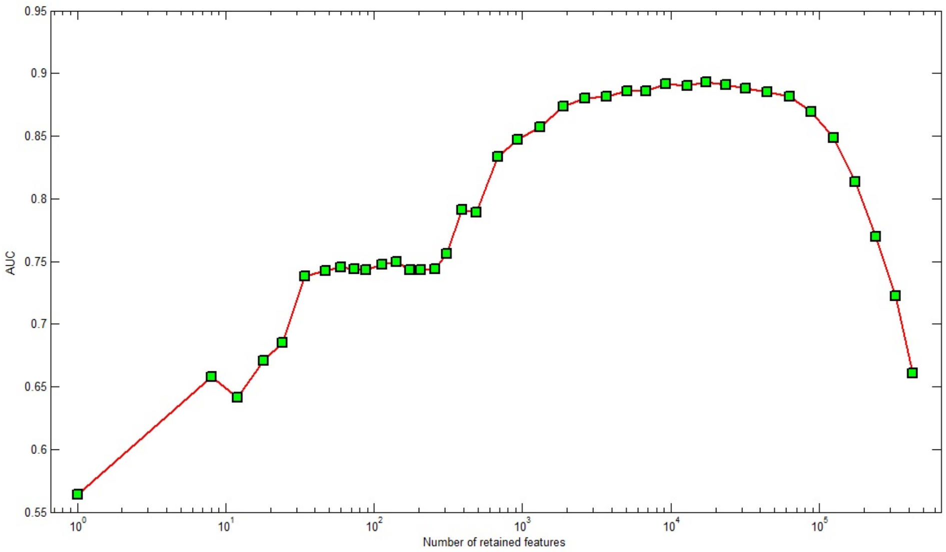


Figure 8.11: AUC versus the number of retained voxels, obtained within the SVM-RFE leave-4pairs-out CV. The maximum AUC was achieved when the number of retained voxel was about the 4% of the total amount of GM voxels (AUC_{max}=0.893).

8.2.3 Leave-19pairs-out cross-validation

Finally, we evaluated the classification performance of SVMs applied to the entire data set by cross-validating using the leave-19pairs-out method. In order to reduce the execution time we considered only one initial randomization, by assuming the previous conclusions about C value stability with respect to the initial data set randomization to be true also in this case.

The AUC value obtained when the heuristic C value has been employed is AUC=0.601. Then, we applied the C value optimization algorithm to improve the classification performance, achieving AUC_{opt}=0.6273. Then, by employing the optimized C value, the discrimination map was built and it was given as input to the complete SVM-RFE algorithm. The resulting AUC versus the number of retained features is shown in figure 8.12. Evidently, a decrease in classification accuracy is reached. Now, the maximum is AUC_{max}=0.804 with a number of retained features of about the 2% of the total amount.

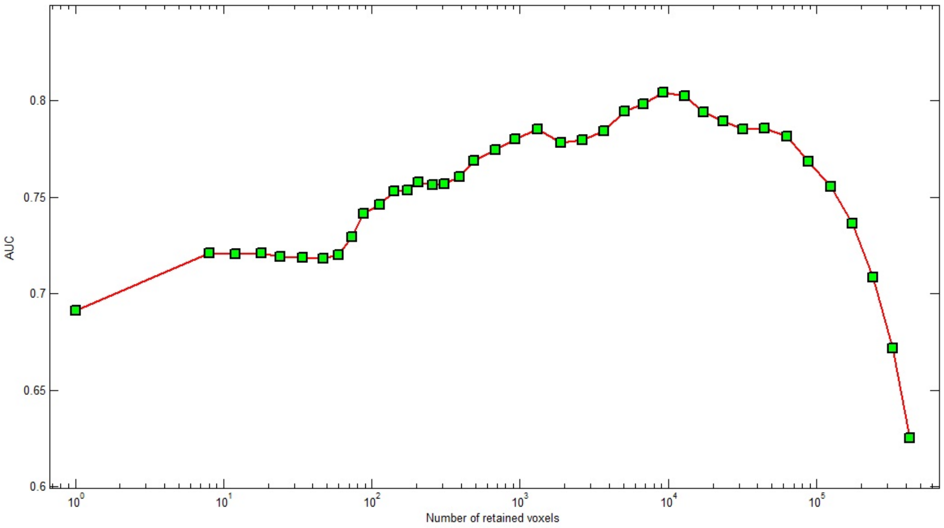


Figure 8.12: AUC versus the number of retained voxels, obtained within the SVM-RFE leave-19pairs-out CV. The maximum AUC was achieved when the number of retained voxel was about the 2% of the total amount of GM voxels (AUC_{max}=0.804).

8.2.4 Comparison of obtained results

In order to make a direct comparison between the SVM-RFE results obtained from the different cross-validation procedures, we can represent all together the AUC versus number of retained voxels in figure 8.13. Despite the trend seems to be the same in all the cases, a sensibly lower classification accuracy is reached when the leave-19pairs-out cross-validation procedure is applied, thus when the training set only contains 57 couples of subjects.

Moreover, we can choose an operative point in the AUC curve where looking for the significative most discriminative regions where ASD grey matter is greater than that of matched controls. In order to be able to compare the most interesting regions obtained from all the cross-validation algorithm, we decided to choose the same operative point that has been considered in Section 7.2) and that leaded to the significative regions shown in figure 7.2. As that operative point the number of retained voxels was about 1800 corresponding roughly to the 0.4 % of the total amount of GM voxels. The resulting AUC versus the number of subjects composing the training set is shown in figure 8.14. Thus, when the number of couples of subjects employed for training remains above 72 (corresponding to 144 subjects), we are sure that the classification performance does not depend on the size of training set and our results can be considered to be stable. Instead, when the number of subjects becomes to decrease, the classification performance gradually gets worse.

In correspondence of these AUC values we can obtain the brain regions where GM is greater in ASD group with respect to controls. When the number of pairs out in the validation set is 1, 2 or 4 the resulting interesting regions in the discrimination maps are the same that have been shown in table 7.1, even if their sizes slightly change. Instead, when 19 pairs are left, thus the training set contains 57 pairs, some regions disappear, as we can see in figure 8.15. However, the leave-19-pairs out cross-validation procedure still leads to a great considerable classification accuracy and several interesting and already obtained regions remain visible (AUC_{max}=0.804, instead AUC in the operative point is 0.778).

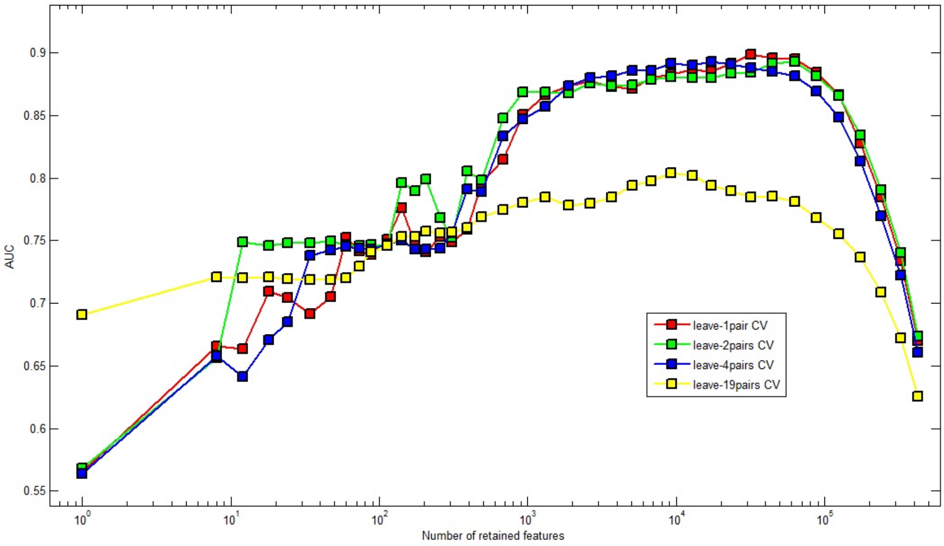


Figure 8.13: Superimposition of AUC versus number of retained voxels curves, obtained from SVM-RFE within leave-1pair out, leave-2pairs-out, leave-4pairs-out and leave-19pairs-out CV procedures.

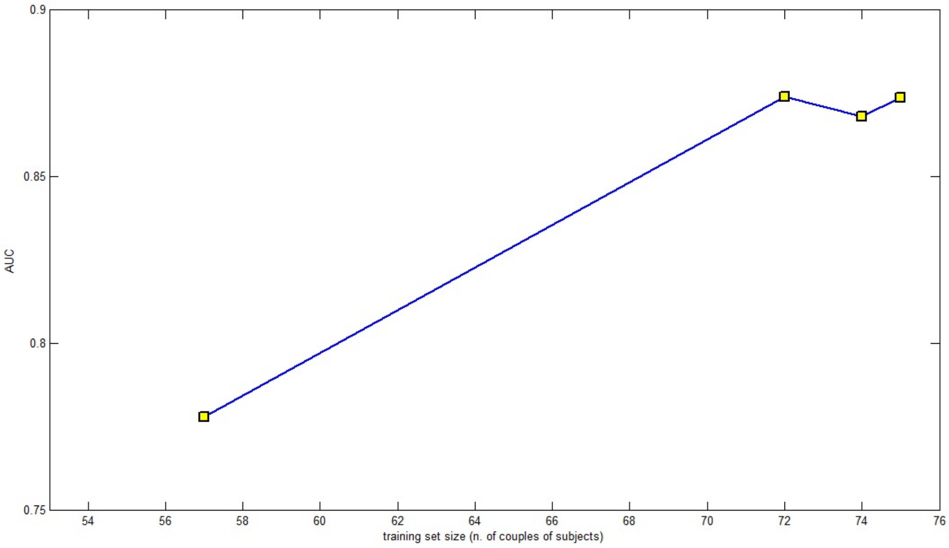
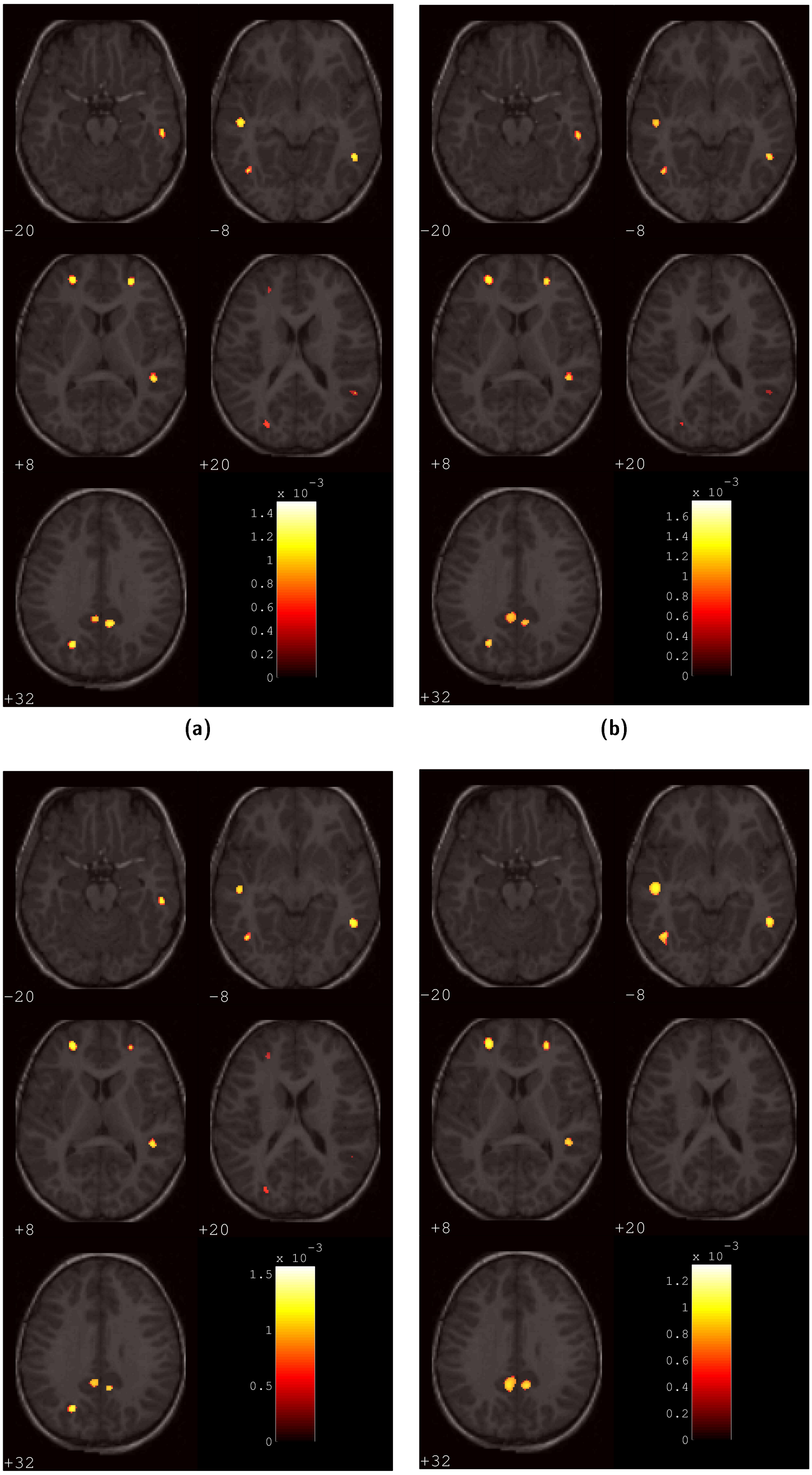


Figure 8.14: AUC versus training set size expressed in terms of number of couples of subjects.



(a) (b)

(c) (d)

Figure 8.15: Discrimination maps overlaid to a representative structural MR image, obtained from SVM-RFE analysis using leave-1pair-out (8.15a AUC=0.874), leave-2pairs-out (8.15b AUC=0.868), leave-4pairs-out (8.15c AUC=0.874) and leave-19pairs-out (8.15d AUC=0.778) cross-validation, and showing those regions where GM is greater in ASD group than controls.

Chapter 9

Discussion of the results and concluding remarks

The aim of this thesis was to compare the widely implemented VBM approach to group studies in neuropsychiatric disorders with the innovative SVM based multivariate techniques. Thus we implemented both the analysis approaches on the same data, composed by 152 male and female subjects, 76 ASD subject and 76 healthy controls matched for sex, age and NVIQ (see Chapters 6, 7, 8) and we compare the results.

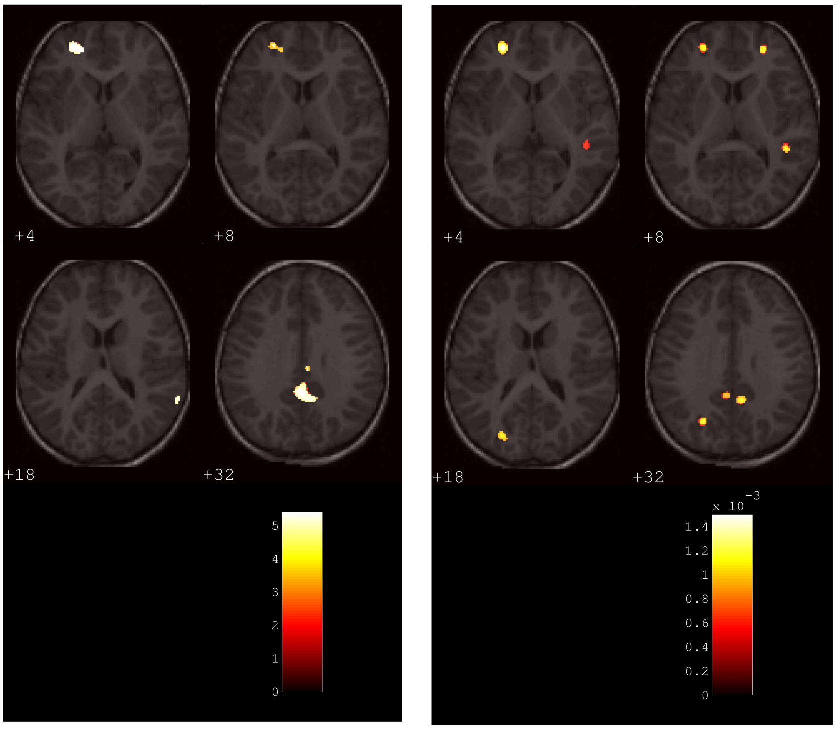
The VBM approach detected interesting between group differences in the analysis of the whole data set, whereas it was no enough sensitive to identify volume differences in the subgroups of male and female subjects (see fig. 6.2).

The implementation of a SVM-based classification method carried out in this thesis correctly discriminated ASD male and female subjects with respect to matched controls with an AUC above the 87 % and with a fraction of retained voxels in the 0.4–29 % range. By choosing as operative point of the system that corresponding to the lower amount of significant features (about 0.4 % of the total amount of GM voxels) we obtained a sensitivity of 82 % and a specificity of 80 %. The discrimination maps at that operative point showed the following main significant regions where the GM of ASD subjects (males and females) is greater than that of the matched control group: Left (L) and Right (R) Superior Frontal Gyri (BA 10); L and R Precunei (BA 31); R Temporo-Parietal Junction (BA 39); L Superior Temporal Gyrus (BA 22); R Superior Temporal Gyrus (BA 41). The separate SVM analyses of male and female subgroups using the whole-group DARTEL-created template revealed gender differences in the following regions where an excess of GM is found in the ASD individuals with respect to control group: L and R Precunei dominate the male ASD subgroup; L and R Superior Frontal Gyri and Middle Frontal Gyrus prevail in the female ASD subgroup.

The consistency with the standard univariate approach was achieved as also the VBM analysis detected some characteristic regions such as the L and R Precunei (BA 31) and the L Superior Frontal Gyrus (BA 10), as shown in fig. 9.1. Moreover, by extending the significance level in the VBM statistical analyses of the entire data set and of the male and female subgroups, we reobtained some of those regions that have been detected within the pattern classification approach.

In this study we demonstrated that multivariate approach based on SVM-RFE could contribute not only to distinguish ASD from control children but also to disentangle the gender specificity of ASD brain alterations [50, 51]. Regional neural differences between male and female ASD children could, in its turn, be related on sex-based differences in the phenotypic expression of ASD. Future studies are then required to specifically investigate this issue.

Nevertheless, our examinations and results have highlighted that multivariate approaches



(a) **(b)**

Figure 9.1: Direct comparison between some of the main regions obtained from VBM (9.1a) and SVM (9.1a) analyses applied to the entire data set.

such as supervised machine learning methods allow the detection of subtle interesting regions and spatially distributed patterns that would be otherwise unobservable by only employing traditional univariate methods. In fact, in recent years, SVM has been successfully applied in the context of disease diagnosis, transition prediction and treatment prognosis, using both structural and functional neuroimaging data [45]. Because of its advantages over traditional analytical techniques and its potentialities in neuroimaging research, standardization and automatization of this method are needed. In order to make it easily implementable and reachable to all research groups, opportunely creating protocols and software tools, such as those actually existing to implement VBM analysis, seem to be necessary.

The increasing interest in the application of pattern classification methods to neuroimaging data to aid diagnosis and prognosis in neurological and psychiatric disorders is justified by a key advantage of these methods over traditional analytical approaches: they allow inferences to be made at single subject level. Such methods could have high translational potential in a clinical setting since they may be used to support autism diagnosis in individual cases.

In our context, improving the performance of the decisional systems by achieving a better AUC could make possible to employ the pattern recognition approach not only to identify brain regions discriminating between patients and controls, but also to predict the class membership of undiagnosed subjects, thus facilitating the early diagnosis of the ASD pathology.

Despite there are several theoretical and practical difficulties to the translational implementation of this approach, the results achieved from our examinations and those of the study published so far are encouraging and inciting the development of computer-based diagnostic and prognostic tools in neurology and psychiatry.

Bibliography

[1] F. Bloch, W. Hansen, and M. Packard. Nuclear induction. *Physical Review*, 69:127, 1946.

[2] E. Purcell, H. Torrey, and R. Pound. Resonance absorbtion by nuclear magnetic moments in a solid. *Physical Review*, 69: 37-38, 1946.

[3] A. Abragam. *The Principles of Nuclear Magnetism.* Clarendon, Oxford, England, 1961.

[4] A. Abragam and B. Bleaney. *Electron Paramagnetic Resonance of transition ions.* Clarendon, Oxford, England, 1970.

[5] A. Carrington and A. McLachlan. *Introduction to magnetic resonance with application to chemistry and chemical physics.* Chapman and all Ltd., 1980.

[6] C.P. Slichter. *Principles of Magnetic Resonance.* Springer-Verlag Berlin Heidelberg New York, 1996.

[7] E. Haacke, R. Brown, M. Thompson, and R. Venkatesan. *Magnetic resonance imaging: physical principles and sequence design.* A John Wiley and Sons, 1999.

[8] F. Bloch. Nuclear induction. *Physical Review*, 70:460-473, 1946.

[9] I. Solomon. Relaxation processes in a system of two spins. *Physical Review*, 99:2, 1955.

[10] A. G. Redfield. On the theory of relaxation processes. *IBM J. Res. Develop.*, 1:19-31, 1957.

[11] A. G. Redfield, Theory of relaxation processes. *Adv. Magn. Res.*, 1:1, 1965.

[12] E. Hahn. Spin echoes. *Physical Review*, 80:580-594, 1950.

[13] J. Ashburner. *VBM Tutorial.* 2010.

[14] J. Ashburner et al. *SPM8 Manual.* 2012.

[15] J. Ashburner. Computational anatomy with the SPM software. *Magnetic Resonance Imaging*, 27:1163-1174, 2009.

[16] D. W. Thompson, *Growth and Form.* Cambridge, UK: Cambridge University Press, 1917.

[17] J. Ashburner and K.J. Friston. Voxel-Based Morphometry-The Method. *NeuroImage*, 11:805-821, 2000.

[18] J. Ashburner and K.J. Friston. Morphometry. In R.S.J. Frackowiak, K.J. Friston, C. Frith, R. Dolan, K.J. Friston, C.J. Price, S. Zeki, J. Ashburner, and W.D. Penny, editors, Human Brain Function. Academic Press, 2nd edition, 2003.

[19] J. Ashburner and K.J. Friston. Spatial normalization using basis functions. In R.S.J. Frackowiak, K.J. Friston, C. Frith, R. Dolan, K.J. Friston, C.J. Price, S. Zeki, J. Ashburner, and W.D. Penny, editors, *Human Brain Function*. Academic Press, 2nd edition, 2003.

[20] J. Ashburner and K.J. Friston. Image segmentation. In R.S.J. Frackowiak, K.J. Friston, C. Frith, R. Dolan, K.J. Friston, C.J. Price, S. Zeki, J. Ashburner, and W.D. Penny, editors, *Human Brain Function*. Academic Press, 2nd edition, 2003.

[21] J. Ashburner and K.J. Friston. Unified segmentation. *NeuroImage*, 26:839-851, 2005.

[22] J. Ashburner and K.J. Friston. Multimodal image coregistration and partioning-a unified framework. *NeuroImage*, 6(3):209-2017, 1997.

[23] J. Ashburner and K. J. friston. Nonlinear spatial normalization using basis functions. *Hum. Brain Mapp.*, 7(4):254-266, 1999.

[24] A.C. Evans, M. Kamber, D. L. Collins and D. Macdonald. An MRIbased probabilistic atlas of neuroanatomy. *Magnetic Resonance Scanning and Epilepsy*, NATO ASI Ser., Ser. A: Life Sci., 264:263-274, 1994.

[25] C. M. Bishop. *Neural Networks for pattern recognition.* Oxford Univ. Press., 1995.

[26] J. Ashburner. A fast diffeomorphic image registration algorithm. *NeuroImage*, 38:95-113, 2007.

[27] W. H. Press, S. A. Teukolosky, W. T. Vetterling and B. P. Flannery. *Numerical Recipes in C* 2nd ed. Cambridge University Press, Cambridge, UK, 1992.

[28] S. Kiebel and A. Holmes. The General Linear Model. In R.S.J. Frackowiak, K.J. Friston, C. Frith, R. Dolan, K.J. Friston, C.J. Price, S. Zeki, J. Ashburner, and W.D. Penny, editors, *Human Brain Function*. Academic Press, 2nd edition, 2003.

[29] C. H. Salmond, J. Ashburner, F. Vargha-Khandem, A. Connelly, D. G. Gadian and K. J. Friston. Distributional assumptions in voxel-based morphometry. *NeiroImage*, 17(2):1027-1030, 2002.

[30] M. Brett, W. Penny and S. Kiebel. Introduction to Random Field Theory. In R.S.J. Frackowiak, K.J. Friston, C. Frith, R. Dolan, K.J. Friston, C.J. Price, S. Zeki, J. Ashburner, and W.D. Penny, editors, *Human Brain Function*. Academic Press, 2nd edition, 2003.

[31] K. J. Worsley, A. C. Evans, S. Marlett and P. Neelin. A three dimensional statistical analysis for CBF activation studies in the human brain. *Journal of Cerebral Blood Flow and Metabolism*, 918:12-900, 1992.

[32] S. J. Kiebel, J. B. Poline, K. J. Friston, A. P. Holmes and K. J. Worsley. Robust smoothness estimation in statistical parametric maps using standardized residuals from the general linear model. *NeuroImage*, 10:756-766, 1999.

[33] K. J. Worsley, M. Andermann, T. Koulis, D. MacDonald and A. C. Evans. Detectiong changes in nonisotropic images. *Human Brain Mapping*, 8:98-101, 1999.

[34] V. N. Vapnik. *The Nature of Statistical Learning Theory.* Springer, 1995.

[35] I. H. Witten and E. Frank *Data Mining: Practical Machine Learning Tools and Techniques* - 2nd ed. Elsevier, 2005.

[36] C. J. C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2:121-167, 1998.

[37] C. W. Hsu, C. C. Chang, C. J. Lin. *A Practical Guide to Support Vector Classification.* 2010.

[38] T. Joachims. Making Large-Scale SVM Learning Practical. In B. Scholköpf, C. J. C: Burges and A. J. Smola (eds.). *Advances in Kernel Methods - Support Vector Learning.* MIT Press, Cambridge, USA, 1999.

[39] C. Ecker, V. Rocha-Rego, P. Johnston, J. Mourao-Miranda, A. Marquand, E. M. Daly, M. J. Brammer, C. Murphy, D. G. Murphy and the MRC AIMS Consortium. Investigating the predictive value of whole-brain structural MR scans in autism: A pattern classification approach. *NeuroImage*, 49:44-56, 2010.

[40] S. Calderoni, A. Retico, L. Biagi, R. Tancredi, F. Muratori, M. Tosetti. Female children with autism spectrum disorder: An insight from mass-univariate and pattern classification analyses. *NeuroImage*, 59(2):1013-1022, 2012.

[41] R. Fletcher. *Practical Methods of Optimization.* John Wiley and Sons, Inc., 2nd edition, 1987.

[42] I. Guyon, J. Weston, S. Barnhill and V. Vapnik. Gene selection for cancer classification using support vector machines. *Mach. Learn.*, 46:389-422, 2002.

[43] F. De Martino, G. Valente, N. Staeren, J. Ashburner, R. Goebel and E. Formisano. Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. *NeuroImage*, 43:44-58, 2008.

[44] D. M. Green, J. A. Swets. *Signal Detection Theory and Psychophysics.* Wiley, New York, 1966.

[45] G. Orrù, W. Petterson-Yeo, A. F. Marquand and G. Sartori. Using Support Vector Machine to identify imaging biomarkers of neurological and psychiatric disease: A critical review. *Neuroscience and Biobehavioral Reviews*, 36:1140-1152, 2012.

[46] C. E. Metz. ROC methodology in radiologic imaging. *Invest. Radiol.*, 21:720–733, 1986.

[47] C. E. Metz. Receiver Operating Characteristic Analysis: A Tool for the Quantitaive Evaluation of Observer Performance and Imaging Systems. *J. Am. Coll. Radiol.*, 3:413-422, 2006

[48] C. Chu, A. L. Hsu, K. H. Chou, P. Bandettini, C. P. Lin and for the Altzheimer's Disease Neuroimaging Initiative. Does feature selection improve the classification accuracy? Impact of sample size and feature selection on classification usig anatomical magnetic resonance images. *NeuroImage*, 60:59-70, 2012.

[49] D. W. Shattuck, M. Mirza, V. Adisetiyo, C. Hojatkashani, G. Salamon, K. L. Narr, R. A. Poldrak, R. M. Bilder, A. W. Toga. Construction of a 3D probabilistic atlas of human cortical structures. *NeuroImage*, 39:1064-1080, 2008.

[50] S. Calderoni, A. Retico, A. Giuliano, L. Biagi, M. Tosetti and F. Muratori. Gender Differences in the Neuroanatomy of Young Children with Autism Spectrum Disorders Detected by Machine Learning Techniques, accepted contribution to the International Meeting for Autism, IMFAR 2013, 2-4 May Donostia-San Sebastián, Basque Country, Spain.

[51] A. Giuliano, S. Calderoni, F. Muratori, L. Biagi, M. Tosetti, A. Retico. Multivariate analysis of structural MRI data to detect gender-related brain abnormalities in children with autism spectrum disorder, accepted contribution to the ECR 2013, 7–11 March Vienna.

Acknowledgments

In primo luogo, desidero ringraziare la mia relatrice, la Dott.ssa Alessandra Retico, che mi ha seguito sin dall'inizio di questo lavoro con competenza, attenzione, costanza e dedizione. Mi ha trasmesso passione, voglia di imparare e migliorare e soprattutto mi ha supportato con pazienza durante la stesura di questa tesi.

Ringrazio la Dott.ssa Sara Calderoni per il grande contributo dato a questo lavoro e per tutte le spiegazioni mediche che con pazienza mi ha fornito.

Un ringraziamento particolare va a tutti i professori del gruppo di Fisica Medica di Pisa, in particolare al Prof. Alberto Del Guerra, senza il quale il nostro corso di laurea magistrale non ci sarebbe. Ringrazio inoltre tutti i professori degli innumerevoli corsi che ho seguito, in modo diverso ognuno è riuscito a insegnarmi tanto. In particolare ringrazio il mio relatore di laurea triennale, il Prof. Francesco Maccarrone, per tutto ciò che mi ha insegnato durante i due anni di laboratorio e nel periodo di tesi.

Un grazie infinito va alle mie migliori amiche Maria, Vania, Ottavia e Cinzia.

Maria, ci conosciamo da quando eravamo bambine e siamo state sempre grandi amiche. Abbiamo condiviso quasi tutte le tappe scolastiche, ci siamo sempre capite al volo e possiamo vantare una collezione infinita di meravigliosi ricordi insieme. Ho sempre saputo di poter contare su di te e so che potrò farlo per sempre. La forte amicizia e la profonda stima reciproca che ci legano non ci abbandoneranno mai. Ti auguro con tutto il cuore il meglio amica mia, e ti ringrazio per la forza che mi hai trasmesso durante il mio percorso universitario.

Vania, sei stata la compagna di banco e di studi perfetta. Mi mancano tantissimo i pomeriggi e le notti passate a studiare tra libri, appunti, coperte sulle gambe, thè e biscotti. Da te ho imparato moltissimo, ma soprattutto mi hai insegnato cosa volesse dire l'amicizia. Sei sempre riuscita a sopportare la mia pesantezza, il mio essere scontrosa e la mia rigidità più di tutti gli altri e hai saputo capirmi e starmi vicino quando ne avevo bisogno. Ti ringrazio per essere venuta ad ascoltarmi alla discussione di laurea triennale e sono certa che vorresti essere qui accanto a me anche oggi.

Cinzia e Ottavia, siete state all'inizio delle perfette compagne di studi, poi siete diventate anche delle amiche eccezionali. Ogni ostacolo universitario, ogni difficoltà emotiva, ogni problema e incomprensione, si alleggerivano improvvisamente quando ne parlavamo tutte e tre insieme. Avete rappresentato una seconda famiglia per me negli ultimi mesi e mi avete ricordato quali fossero i miei veri punti di riferimento anche quando credevo di averli persi. Senza il vostro aiuto non ce l'avrei mai fatta a superare tutto e ad essere qui oggi a scrivere finalmente l'ultima pagina di questa tesi.

Ringrazio le mie coinquiline e vicine di casa: Stefania, Neda, Lisa, Rosalba, Angela, Livia, Ester e Giulia. Senza di voi non ce l'avrei fatta a raggiungere questo traguardo. Avete sempre creduto in me, mi avete sopportato con enorme pazienza quando ero scontrosa e stressata, mi avete ascoltato e capito nei momenti difficoltà. Non dimenticherò mai tutte le nostre cene, gli esperimenti in cucina, le tisane e le chiacchiere serali, avete reso questi cinque anni indimenticabili.

Un saluto affettuoso va a tutti i miei compagni di università e amici di Pisa, tra cui

Giovanni, Riccardo, Matteo, Gabriele, Marta, Antonietta, Piero, Chiara, Elena, Giorgio, Camilla e Guido. Tutte le cene, i picnic in Piazza dei Miracoli, le gite a Gardaland, le giornate passate insieme a studiare in biblioteca, i caffè presi alle macchinette dell'INFN, hanno contribuito ad arricchire questi ultimi cinque anni trascorsi a Pisa e a renderli indimenticabili.

Ringrazio tutti i miei amici di Rutigliano e dintorni, mi avete sempre fatto sentire a casa ogni volta che tornavo da voi. Un abbraccio particolare va a Nicky, Giuseppe, Vanessa, Tania, Marienza, Massimo, Claudio, Antonio, Domenico e Vincenzo.

Ringrazio di cuore tutta la mia famiglia. Un abbraccio fortissimo va alla mia nonna e a zia Carmela. E' da quando sono nata che fate sacrifici per me in modo naturale e senza mai farmeli pesare e se sono riuscita a realizzare molti dei miei desideri lo devo anche a voi. Un grazie va anche a zia Rita e a zio Toti, alle mie tre cuginette e a mio fratello Antonio. E' anche grazie al vostro sostegno, ai vostri preziosi consigli e al vostro affetto che ce l'ho fatta a raggiungere i miei obiettivi.

E infine mamma e papà, è a voi che dedico questo importante traguardo. Avete sempre assecondato le miei passioni e i miei desideri senza limitarmi in alcun modo, e vi sarò sempre riconoscente per tutti i sacrifici che ogni giorno fate per me e per tutti gli insegnamenti che mi avete dato. Tutto questo non sarebbe stato possibile senza il vostro aiuto, il vostro supporto, la vostra fiducia in me, i vostri consigli e il vostro affetto.