Carlos Andrés Méndez Guerrero

Image Based Biomarkers from Magnetic Resonance Modalities: Blending Multiple Modalities, Dimensions and Scales.

Ph.D. Thesis

April 24, 2013

Università degli Studi di Verona
Dipartimento di Informatica

Advisor:
Prof. Gloria Menegaz

Series N°: **TD-05-13**

Università di Verona
Dipartimento di Informatica
Strada le Grazie 15, 37134 Verona
Italy

a mis padres,
que siempre han sido el mejor ejemplo.

Contents

1

Introduction

Medical imaging is arguably one of the most impacting technologies in modern society. Its mostly noninvasive nature has substantially contributed to improve the quality of experience in disease prevention and treatment [65]. Medical Imaging gives support to the medical diagnosis, medical treatment or follow-up as well as to medical and biological research. Currently, it is a solid part of electronic medical systems and it is pervasive across medical institutions including health services, hospitals, universities and research centres. Medical imaging has contributed to improve illness diagnosis and treatment across a wide range of conditions. With the advent of MRI and CT technology, which involve digital computing and advanced electronics, a quantum leap in medical imaging technology was made and a new branch in science was born: digital medical imaging. It embraces several areas of science and technology including, conventional medicine, electronics, digital image and signal processing. The latter is fundamental in modern medical imaging since it contributes to the automatic enhancement of sensed information and more critically to its understanding. Different tasks have been automated with different degrees of success. Several techniques have been applied to these automated tasks, which cover some domains including signal processing, statistics, pattern recognition and machine learning. One of the main advantages of machine learning methods is that they are able to automatically find non-obvious, complex relationships between data that, otherwise, are usually found by an extensive knowledge of the problem. Generalization models can then be much more easily inferred from these relationships [65]. The successful analysis and processing of medical imaging data is a multidisciplinary work that requires the application and combination of knowledge from diverse fields, such as medical engineering, medicine, computer science and pattern recognition and classification.

In this framework the overall objective of this thesis is to investigate non-supervised processing techniques for classification of image based biomarkers, with special emphasis on clinical application for diagnosis and therapy assessment, particularly related to cancerous tissue.

The word biomarker can be defined as any detectable biological feature or parameters that provides information about its source. More specifically it is used to denote anatomic, physiologic, biochemical, or molecular parameters detectable with imaging methods used to establish the presence or severity of disease. By this

definition, much of imaging can be thought of as a biomarker, and certainly MR methods fall in to this definition. However, the utility of an imaging biomarker, especially in decision making about cancer and its treatment, requires more than detection; it requires understanding of the methods strengths and weaknesses.

The development and use of biomarkers offer the prospect of more efficient clinical studies and improvement in both diagnosis and therapy assessment [168]. As a general term, it applies to all detection modalities. An imaging biomarker is a biological feature detectable by imaging modalities. In the medical context it refers to a feature of an image that represents a particular aspect of the patient under the imaging procedure. Successful use of biomarkers is most likely when [158]: (a) the presence of an imaging marker is closely linked with the presence of a target disease; (b) detection and/or measurement of the biomarker is accurate, reproducible, and feasible over time; and (c) measured changes are closely linked to success or failure of the therapy being evaluated.

Imaging biomarkers are often based on the morphology, physiology or metabolism and can include examples such as the following:

- Morphology: tumor diameter, volume, lesion number, tumor burden, infiltration, texture (e.g. solid, necrotic).
- Physiology: tissue vascularity/perfusion, microvascular permeability (angiogenic activity), diffusivity.
- Metabolism: bone scintigraphy, PET/SPECT, MR spectroscopy, biochemical markers.

With respect to cancer research and therapy, the most important imaging biomarkers are related to angiogenesis, the process by which tumors develop a circulatory blood supply, which results in the development of vascular networks that are both structurally and functionally abnormal. Despite it's growing acceptance and support, the various analysis methods employed have considerable influence on the interpretation of derived parameters and their value as potential biomarkers.

1.1 Objectives

Among the different possible scenarios, this thesis will focus on Magnetic Resonance Imaging (MRI) modalities, with particular emphasis on Dynamic Contrast-Enhanced MRI (DCE-MRI) and Diffusion MRI. The overall objective consists in *(i) to propose new processing methodologies for the integration of different MRI modalities, (ii) to identify, extract and characterize features and metrics allowing a good data representation of the multi-parametric MRI volumes, (iii) to implement unsupervised pattern recognitions tools enabling the classification/characterization of tissues, and (iv) to validate the results from both the technical and clinical perspective.*

The main approach to be investigated in this work is the the classification through similarity/dissimilarity representations, a novel approach to pattern recognition in which objects are characterized by relations to other objects instead of

by using features. This dissimilarity representation may be well suited for medical imaging and able to represent the relationship among the (multidimensional, multimodal) data at hand. The application of these pattern recognition techniques to the field of medical imaging follows the idea of getting close to the subjective way a human medical expert explicitly recognize characteristics and classify new images judging similarities from an ideal prototype.

It is worth mentioning that our data cover a wide spectrum of types such as scalars, tensors and temporal series. The project will mostly deal with brain and breast, even though other data could also be considered.

The thesis is organized as follows:

Chapter 2 Background on relevant MRI modalties. Due to the central role played by diffusion imaging in this thesis, in this chapter the different acquisition techniques and their potential applications are briefly illustrated. Furthermore, an introduction to DCE-MRI is presented, as well as details on the information processing models.

Chapter 3 In this chapter the relevant information of unsupervised classification is described. We start focusing on the important issue of data representation, followed by unsupervised classification and cluster ensembles. The chapter ends with the issue of validation in unsupervised classification.

Chapter 4 Here the principal approaches for multi-modal MRI data combination are reviewed. At the end of the chapter we make an overview of the proposed methodologies that will be detailed in the next chapters.

Chapter 5 This chapter proposes two strategies for the clinical assessment of heterogeneity inside tumoral lesions. As a case study we present results obtained with real clinical datasets of breast ductal carcinoma, evaluated both by comparison to a typical feature-based approach as well as by their clinical significance assessed by medical experts.

Chapter 6 In the last methodological chapter we present a multi-view approach to multi-modal MRI combination. This method is described and analyzed with synthetic datasets.

Chapter 7 Conclusions and future perspectives.

1.2 Summary of main contributions

The main contributions of this thesis are the following:

- A protocol for the multi-modal integration of the information provided by multi-modal MRI using a combined dissimilarity vectorial space. This protocol was shown and validated for DCE-MRI and DWI-MR for evaluating tumor heterogeneity.
- The use of the dissimilarity-based representation paradigm to overcome the limitations imposed by the dissimilar nature of multi-modal MRI.
- A study with clinical datasets of breast ductal carcinoma, assessed from the methodological point of view as well as by medical experts.
- The extension of the *multi-view* notion, which served to formulate a novel approach for multi-modal MRI fusion for unsupervised classification, shown with DTI-MR and DCE-MRI.
- The integration of manifold learning techniques to account for the complex high-dimensional geometric structure of the Diffusion Tensor Imaging data.
- The use of Cluster Ensembles as an alternative to the problem of multi-modal data fusion.

2

Principles of MRI and imaging applications

2.1 Overview

In order to better expose the contribution of our work we provide in this chapter the theoretical basics of the principal MRI modalities we will use in subsequent chapters; Diffusion MRI and Dynamic Contrast-Enhanced MRI. Furthermore, we discuss the application of these concepts to the acquisition of MR images, highlighting the potentialities as well as the limits of the different techniques.

This Chapter is essentially subdivided into 3 parts. In the first sections we overview the physics of diffusion and the mathematical aspects behind the acquisition of diffusion MR images. In the second part we concentrate more on the applicability of diffusion MRI to medical investigations, by highlighting its importance and impact in the clinical practice. In this part we also highlight the limitation of the current clinically used methods, i.e. Diffusion Weighted (DWI) and Diffusion Tensor Imaging (DTI), by investigating the problems first from the practical point of view and then by going back to the mathematical framework with the formal description of the solutions proposed in literature to overcome these limitations. We conclude the Chapter with a section dedicated to Dynamic Contrast-Enhanced MRI in which we detail its utility and the challenges derived from the different processing techniques. We make emphasis on the two main approaches; pharmacokinetic modelling and semi-quantitative voxel-wise analysis.

2.2 Diffusion MRI

2.2.1 Basis of the Diffusion phenomenon

Diffusion is an essential physical process for the normal functioning of living systems. For example, the transport of metabolites into cells is facilitated by diffusion. This phenomenon, omnipresent in the water in living tissue, has the potential, through diffusion-weighted magnetic resonance imaging, to provide insights into cell physiology, cell structure and potentially the connections of the living human brain.

Diffusion is a mass transport process arising in nature, which results in molecular or particle mixing without requiring bulk motion. Diffusion should not be

confused with convection or dispersion, which are different transport mechanisms that require bulk motion to carry particles from one place to another. Imagine carefully introducing a drop of colored fluorescent dye into a jar of water. Initially, the dye appears to remain concentrated at the point of release, but over time it spreads radially, in a spherically symmetric profile. This mixing process takes place without stirring or other bulk fluid motion.

2.2.2 Fick's Law

The physical law that explains this phenomenon is called Ficks first law, which relates the diffusive flux to any concentration. Given a local concentration of particles $n(\mathbf{r}, t)$, Fick suggested that the flux of particles J may be written

$$\mathbf{J} = -D\nabla n(\mathbf{r}, t) \tag{2.1}$$

where the constant of proportionality, D, is called the "diffusion coefficient". Total particle number number conservation requires that the time rate of change of $n(\mathbf{r}, t)$ is simply related to the local flux divergence, $-\frac{\partial}{\partial n}J = \frac{\partial n}{\partial t}$, leading to

$$\frac{\partial n}{\partial t} = D\nabla^2 n \tag{2.2}$$

Equations 2.1 and 2.2 are known respectively as Fick's first and second laws.

As illustrated in Figure 2.1, Ficks first law embodies the notion that particles flow from regions of high concentration to low concentration (thus the minus sign in equation 2.1) in an entirely analogous way that heat flows from regions of high temperature to low temperature, as described in the earlier Fouriers law of heating on which Ficks law was based. In the case of diffusion, the rate of the flux is proportional to the concentration gradient as well as to the diffusion coefficient. Unlike the flux vector or the concentration, the diffusion coefficient is an intrinsic property of the medium, and its value is determined by the size of the diffusing molecules and the temperature and microstructural features of the environment. The sensitivity of the diffusion coefficient on the local microstructure enables its use as a probe of physical properties of biological tissue. On a molecular level diffusive mixing results solely from collisions between atoms or molecules in the liquid or gas state. Another interesting feature of diffusion is that it occurs even in thermodynamic equilibrium, for example in a jar of water kept at a constant temperature and pressure. This is quite remarkable because the classical picture of diffusion, as expressed above in Ficks first law, implies that when the temperature or concentration gradients vanish, there is no net flux. There were many who held that diffusive mixing or energy transfer stopped at this point. We now know that although the net flux vanishes, microscopic motions of molecule still persist; it is just that on average, there is no net molecular flux in equilibrium.

2.2.3 The propagator description

There are two ways to begin with, in order to describe basic diffusion [67]: either a phenomenological approach starting with Ficks laws and their mathematical solutions, as we have described above, or a physical and atomistic one, by considering

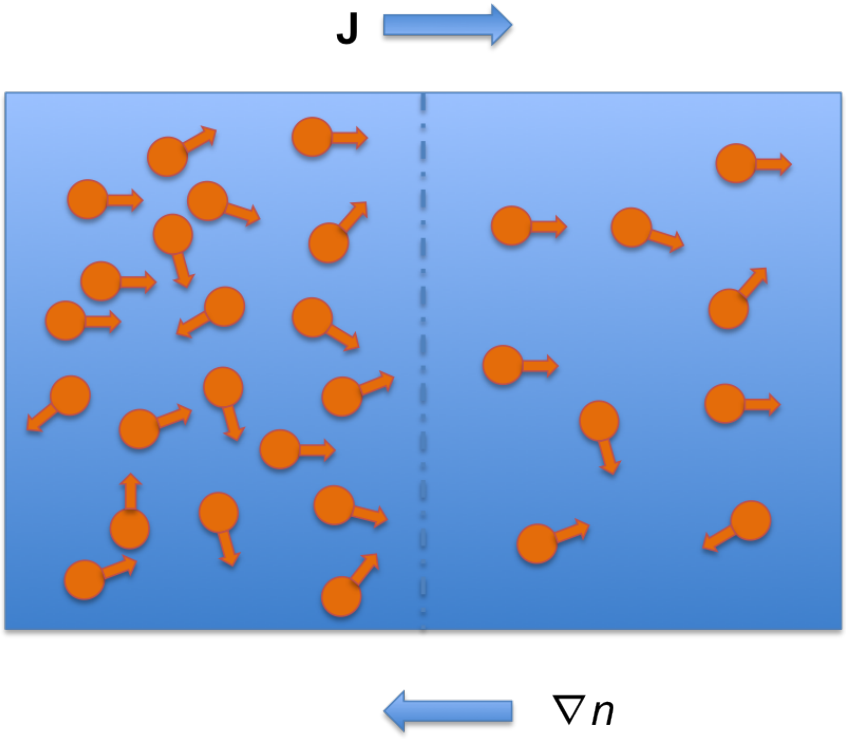


Fig. 2.1: According to Ficks first law, when the specimen contains different regions with different concentrations of molecules, the particles will, on average, tend to move from high concentration regions to low concentration regions leading to a net flux (**J**).

the "random walk" of the diffusing particles. While this last approach was rather straightforward in gases thanks to Maxwells kinetic theory of gases, the first one follows the historical development of diffusion studies in solid materials under a gradient of chemical potential. People began to be concerned with an atomic scale approach first of all with the electrical conductivity of ionic crystals, and later with the Kirkendall effect which was observed in several inter-diffusion systems. As diffusion processes depend on atom (ion) jumps whose occurrence is dictated by atomic defects (vacancies or interstitials), a description based on atom movements became compulsory. The never-ending movement of particles in suspension in a fluid was discovered by a Scottish botanist, Robert Brown, who was observing with his microscope the "swarming" motion in the fluid of small particles extracted from living pollen grains. He noticed that this motion was quite general in fresh pollen grains, as well as in dried ones. Browns experiments revealed such motion to be a general property of matter in this state. The name "Brownian motion" has been coined in honor of Brown to qualify the random walk of microscopic particles in suspension in a fluid.

The mathematical form of Brownian motion was derived a little bit later (in 1905) by Albert Einstein. He was the first to understand, contrarily to many scientists of his time, that the basic quantity was not the average velocity of the particles, but their mean square displacement in a given time. Trajectories are such that velocity is meaningless. Einstein, who was unaware of Browns observation and seeking evidence that would undoubtedly imply the existence of atoms, came to the conclusion that "bodies of microscopically visible size suspended in a liquid will perform movements of such magnitude that they can be easily observed in a microscope" [45] . Einstein used a probabilistic framework to describe the motion of an ensemble of particles undergoing diffusion, which led to a coherent description of diffusion, reconciling the Fickian and Brownian pictures. He introduced the "displacement distribution" for this purpose, which quantifies the fraction of particles that will traverse a certain distance within a particular timeframe, or equivalently, the likelihood that a single given particle will undergo that displacement. In this sense, he rewrote Fick's laws for the diffusion of molecules in a concentration gradient, in terms of diffusion under probability gradients. This step enabled a description of Brownian motion as a stochastic process, but one in which the probability densities obeyed differential equations. The key tool in this description is the conditional probability $P(\mathbf{r}|\mathbf{r}', t)$ that a particle starting at $\mathbf{r}$ at time zero will move to $\mathbf{r}'$ after a time t. Combined with the local particle concentration $n(\mathbf{r}, t)$ one may write

$$n(\mathbf{r}', \mathbf{t}) = \int n(\mathbf{r}, 0) P(\mathbf{r}|\mathbf{r}', t) d\mathbf{r}. \tag{2.3}$$

Since $n(\mathbf{r}', t)$ obeys the Fick's law diffusion equation for arbitrary initial condition $n(\mathbf{r}, 0)$, the conditional probability also obeys the partial differential equation

$$\frac{\partial}{\partial t} P(\mathbf{r}|\mathbf{r}', t) = D\nabla^2 P(\mathbf{r}|\mathbf{r}', t), \tag{2.4}$$

where ∇ is taken to operate on the primed spatial coordinates. Given the initial condition where molecules start at r, $P(\mathbf{r}|\mathbf{r}', 0) = \delta(\mathbf{r}' - \mathbf{r})$, the solution to Equation 2.4 is the Gaussian

$$P(\mathbf{r}|\mathbf{r}', t) = (4\pi D t)^{-3/2} \exp\left(-\frac{(\mathbf{r}' - \mathbf{r})^2}{4Dt}\right). \tag{2.5}$$

At this point we introduce the idea of an ensamble, the set of all replicas of the system (for example, the molecule under consideration) representing each of the accessible states. Hence, we may define the ensemble average, $< A >$, of some property A, as

$$< A >= \sum_s P(s) A(s) \tag{2.6}$$

where s represents a possible state of the system in the ensemble and $P(s)$ is the probability of that state. The Gaussian nature of the conditional probabilty for self-diffusion, represented by Equation 2.4, leads to the important result

$$< (\mathbf{r}' - \mathbf{r})^2 >= 6Dt \tag{2.7}$$

that in one dimension becomes $< (x' - x)^2 >= 2Dt$. This is known as the Einstein equation for diffusion.

It is to be noted that Equation (2.4) is true only for an isotropic medium, where the diffusion is indeed a simple scalar property. In anisotropic media, it is necessary to define a diffusion tensor and rewrite the differential equation

$$\frac{\partial}{\partial t} P(\mathbf{r}|\mathbf{r}', t) = \nabla \cdot [\mathbf{D} \nabla P(\mathbf{r}|', t)] \,. \tag{2.8}$$

where $\mathbf{D}$ is known as the *diffusion tensor*. This diffusion tensor is a 3×3 symmetric positive definite matrix that characterizes diffusion in 3D, it describes how the particle flux in any direction is related to the directional probability gradients.

2.2.4 Adding Diffusion Weighting to a pulse sequence

MRI exploits the fact that the human body is mainly constituted by water molecules, and each molecule has two hydrogen protons. When the scanner applies a powerful magnetic field, the magnetic moments of some of these protons change, aligning with the direction of the magnetic field. A radio frequency is then briefly applied, producing an electromagnetic field, causing the flip of the spin of the aligned protons in the body. After the field is turned off, the protons decay to the original state and the difference in energy between the two states is released as a radio frequency photon. These photons produce the electromagnetic field detected by the scanner. Additional magnetic fields are applied in order to make the field strength depend on the position within the scanned subject, thus making the frequency of the released photons dependent on the position. An image can be constructed since the protons in different tissues return to their equilibrium state at different rates [137].

Almost any MRI impulse sequence can be modified to become sensitive to diffusion. The basics for diffusion weighted were introduced by Stejskal and Tanner [173]. After excitation and before signal sampling, application of a bipolar gradient adds to each spin's precession a positive phase proportional to its average position (along the direction of the gradient) during the first gradient lobe, and a negative phase proportional to its average position during the second lobe. The sum of this phases is related to the difference between these two positions. As shown in Fig. 2.2, the bipolar gradient has no net effect on spins which do not move, i.e. low diffusion regions; they are completely "in phase" after its application.

As shown in Fig. 2.2, if there is spin displacement as a result of Brownian motion, i.e. we are in a high diffusion region, the signal A is attenuated exponentially by the product of the diffusion coefficient D and a factor b which is a function of the diffusion weighting gradients [107], i.e. for rectangular gradients

$$A = \frac{S_g}{S_0} = \exp(-bD), \tag{2.9}$$

where

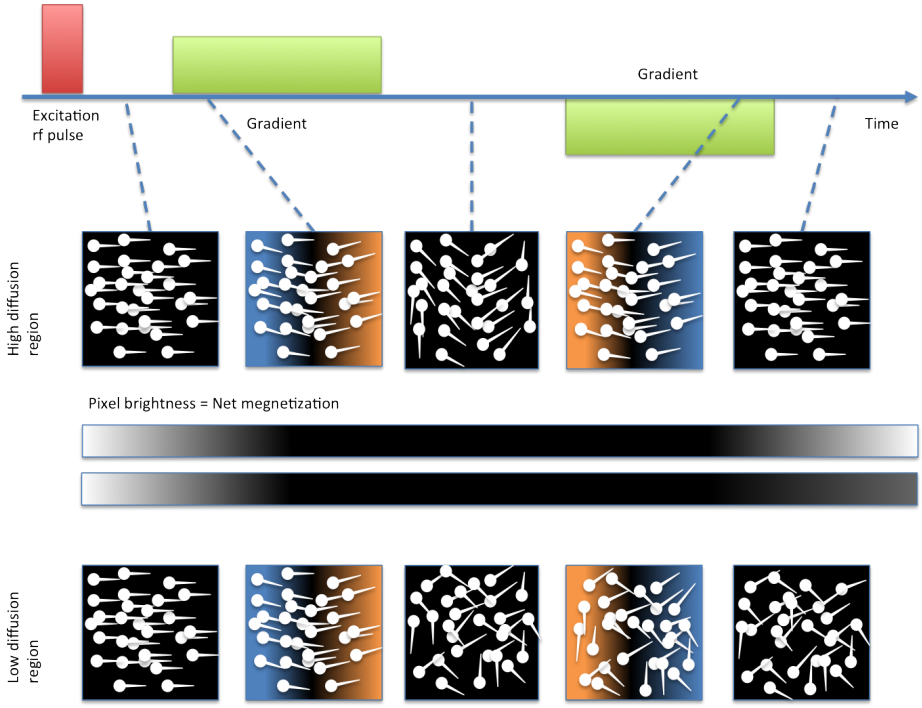


Fig. 2.2: Illustration of the effect of a bipolar gradient on spin phase. After excitation, spins are in-phase. With the application of a positive gradient, the spins experience a stronger magnetic field (blue) to the (e.g.) right, and a weaker magnetic field (shown in orange) to the (e.g.) left. This creates an increase in precessional frequencies of spins on the right compared to that of spins on the left, and spins accrue a phase proportional to their left-right position. When the positive gradient is turned off, all spins precess at the same frequency, but retain their relative phases. The net magnetization is negligible at this time due to phase incoherence of spins. A negative gradient reverses the direction of the spins precessional frequency change, resulting in an equal but opposite phase proportional to their left-right position. In the absence of displacement, i.e. in low diffusion regions, between the first and second gradient lobes, there is no net effect on spin phase, net magnetization, or pixel brightness. In high diffusion regions, when the second gradient is applied, it removes much of the dephasing, but magnetization recovery is incomplete due to diffusion-induced displacement during the bipolar gradient application. Spins in regions with high diffusion have greater phase incoherence and signal loss than spins in regions with low diffusion, as shown in the central bars.

$$b = (\gamma G \Delta)^2 (\Delta - \delta/3), \tag{2.10}$$

S_0 is the signal intensity without the diffusion weighting $b = 0$, S_g is the signal with the gradient g, λ is the gyromagnetic ratio, G is the strenght of the gradient pulse, δ is the duration of the pulse and Δ is the time between the two diffusion-weighting pulses.

2.2.5 Anisotropic diffusion

As we have briefly introduced earlier in this Chapter, while diffusion is a three dimensional process, the molecular mobility may not be the same in all directions. This anisotropy may be due to the physical arrangement of the medium or to the presence of obstacles that impede diffusion in some directions. The result is that diffusion appears different when gradients are put in different directions.

The proper way to address anisotropic diffusion is to consider the diffusion *tensor*. Diffusion is no longer characterized by a single scalar coefficient but by a symmetric tensor, $\mathbf{D}$, a 3×3 symmetric positive definite matrix, which describes molecular mobility along each axis and correlation between displacements along these axes:

$$\mathbf{D} = \begin{pmatrix} D_{xx} & D_{xy} & D_{xz} \\ D_{xy} & D_{yy} & D_{yz} \\ D_{xz} & D_{yz} & D_{zz} \end{pmatrix} \tag{2.11}$$

The motivations behind the symmetry of the diffusion tensor are related to the fact that in the virtual reference frame $\{x', y', z'\}$ that coincides with the principal or main directions of diffusivity, the off-diagonal terms do not exist and the tensor is reduced only to its diagonal terms, $D_{x'x'}$, $D_{y'y'}$, $D_{z'z'}$, which represent molecular mobility along virtual axes x', y', and z', respectively. The signal attenuation A then becomes:

$$A = \exp(-b_{x'x'} D_{x'x'} - b_{y'y'} D_{y'y'} - b_{z'z'} D_{z'z'}) \tag{2.12}$$

where b_{ij} are the elements of matrix $\mathbf{b}$ (which now replaces the $b - value$) expressed in the coordinates of this reference frame.

In practice, however, measurements are made in the *reference* frame $[x, y, z]$ of the gradients, which usually does not coincide with that of the tissue. Therefore, one must must also consider the coupling of non-diagonal terms, D_{ij}, $(i \neq j)$, of the diffusion tensor (now expressed in the gradient frame), which reflect correlation between molecular displacements in perpendicular directions [13]. Therefore Equation (2.12) becomes:

$$A = \exp(-\sum_{i=x,y,z} \sum_{j=x,y,z} \mathbf{b}_{ij} \mathbf{D}_{ij}) \tag{2.13}$$

2.2.6 Geometric representation of the diffusion tensor

An intuitive way to understand the meaning of $\mathbf{D}$ is to perform various thought experiments in which we follow the Brownian motion for an ensemble of "tagged"

water molecules released from the center of a voxel. If we imagine performing this experiments in a jar of water, the rate of diffusive transport will be the same in all directions. Diffusion is then said to be *isotropic* and is completely specified by a single scalar diffusion constant, D. This, diffusion isotropy describes the case in which the molecular diffusivity is independent of the medium's orientation. For a diffusion time, Δ, the translational displacement distribution is spherically symmetric, and surfaces of constant probability or water concentration are concentric spheres. When considering the Einstein equation 2.7, we can construct a sphere whose radius equals the root-mean-squared (rms) displacement of water molecules after diffusion time Δ, a graphical representation of which can be found in the first gliph of Fig. 2.3.

If we now perform the same experiment in a liquid crystalline medium or a system with microscopically aligned rods, the rate of diffusive transport will no longer be the same in all directions. Diffusion anisotropy implies that the translational displacement probability of the diffusing species is now biased, depending on the medium orientation. In homogeneous, i.e. spatially uniform, anisotropic media, the voxel-averaged displacement distribution is given by:

$$P(\mathbf{r}|\mathbf{r}',t) = (4\pi|\mathbf{D}|t)^{-3/2} \exp\left(-\frac{\left[(\mathbf{r}'-\mathbf{r})^2\right]^T \mathbf{D}^{-1}(\mathbf{r}'-\mathbf{r})^2}{4t}\right) \tag{2.14}$$

where $(\mathbf{r}' - \mathbf{r})$ is the displacement, and $|\mathbf{D}|$ is the determinant of $\mathbf{D}$. The covariance matrix characterizes the shape of the displacement distribution.

In order for Equation (2.14) to tend to zero for large displacements, all quadratic forms of the diffusion tensor have to be positive, i.e. $\mathbf{D}$ has to be a positive definite matrix. Then this equation describes a three-dimensional ellipsoid in displacement space, call the "diffusion ellipsoid" [11], whose size, shape, and orientation embody important features of anisotropic gaussian diffusion. The shape of diffusion can be easily visualized with these ellipsoidal glyphs (squished or stretched spheres). Figure 2.3 illustrates diffusion as anisotropic (cigar shaped), as a planar shaped and isotropic, visualized as a sphere.

The DT can be decomposed, by eigenanalysis, into eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq 0$ and corresponding eigenvectors $\epsilon_1, \epsilon_2, \epsilon_3$. The first vector gives the principal direction of diffusion, the other two span an orthogonal plane to it and the eigenvalues quantify the diffusivity in these directions. When $\lambda_1 \gg \lambda_2$, ϵ_1 is aligned with the preferred diffusion direction of the water molecules in that voxel, and λ_1 is its diffusivity (Fig. 2.4).

If the diffusion coefficient of water at body temperature is a constant, then how can we use it to garner information about tissue microstructure? The answer lies in Einstein equation (2.7), which says that the mean squared displacement is directly proportional to the observation time. We do not measure a diffusion coefficient directly with diffusion MRI. Rather, we infer the diffusion coefficient from observations of the displacements over a given time period. If the diffusing water molecules encounter any hindrances along their random walk, such as cell membranes and macromolecules, the mean squared displacement per unit time will be lower than when observed in "free" water. Thus, when we apply Einstein equation to compute the diffusion coefficient, it will appear that the diffusion

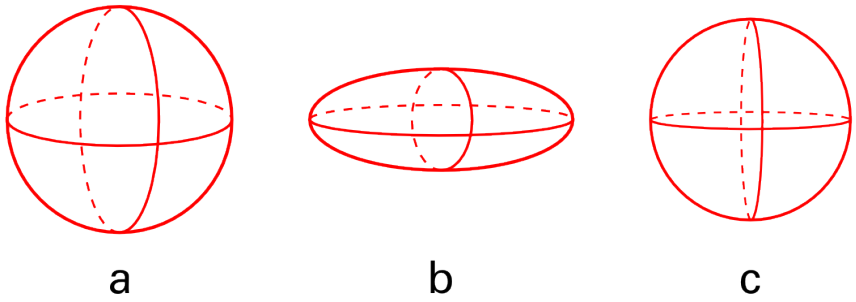


Fig. 2.3: The three stereotypes of Gaussian diffusion in 3D, visualized with ellipsoidal isoprobability surface glyphs with a) isotropic, b) linear or c) planar shape. In DTI, all diffusion shapes are spanned by an interpolation between these three types [137].

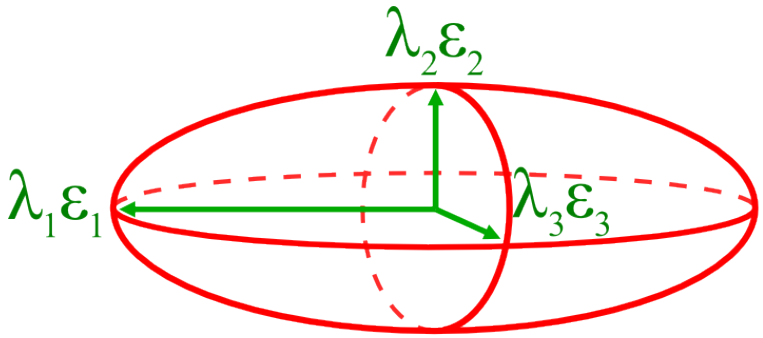


Fig. 2.4: The three stereotypes of Gaussian diffusion in 3D, visualized with ellipsoidal isoprobability surface glyphs with a) isotropic, b) linear or c) planar shape. In DTI, all diffusion shapes are spanned by an interpolation between these three types [137].

coefficient is lower [186]. Thus, we refer to the apparent diffusion coefficient, most frequently abbreviated to "ADC" [19]. The effect of such hindrance in tissue is to make the average apparent diffusion coefficient about four times smaller than in free water. For what concerns the acquisition of this kind of images, diffusion-weighted MR sequences are made sensitive to diffusion by the addition of magnetic field gradients, i.e. the magnetic field is made to vary in a linear manner over the volume of interest.

2.2.7 The set of positive definite matrices as a Riemannian manifold

Positive definite matrices are symmetric matrices with the restriction that their eigenvalues have to be positive. This restriction can be translated into constraints on the values that the entries of the matrix can take. For illustration, let

$$X = \begin{pmatrix} a & c \\ c & b \end{pmatrix} \tag{2.15}$$

be a 2×2 symmetric matrix. X is positive definite if and only if the diagonal elements are positive ($a > 0, b > 0$) and the determinant is positive ($ab - c2 > 0$). The set of triplets (a, b, c) that result in positive definite matrices is an open subset of $\mathbb{R}^3$ and has the shape of a cone. This is illustrated in Figure 2.5. Valid triplets (a, b, c) lay inside the cone. For example, the matrices $X1 = diag(1, 0.1)$ and $X2 = diag(0.1, 1)$ are represented by the triplets $(0.9, 0.1, 0)$ and $(0.1, 0.9, 0)$ respectively. Notice that the cone is convex, so interpolation between any two points of the cone is permitted. This implies averages of positive definite matrices are positive definite, as illustrated by the midpoint point $\overline{X} = diag(0.5, 0.5)$. Extrapolation, however, might result in matrices that are not positive definite: in Figure 2.5, the straight line connecting $X1$ and $X2$ extends beyond the boundaries of the cone.

By means of log transformations taking into account the Riemmanian geometry and constrains [163], a straight line can be traced in the log-space and then taking the matrix exponential results in the hyperbola connecting $X1$ and $X2$. This hyperbola is entirely inside the cone by definition. The midpoint of this line is $\breve{X} = diag(0.3, 0.3)$. For positive numbers, the exponential of the average of the logs of two numbers is the same as the geometric mean. Analogously, $\breve{X}$ can be thought of as a geometric mean of $X1$ and $X2$.

Both the pure straight line and the hyperbola are "straight lines", depending on whether it is traced directly or on the log-space. In fact, both lines are geodesics corresponding to two different geometries defined on the cone. A thorough geometric review of the properties of Diffusion Tensors is presented by Schwartzman [163] and Fletcher [52].

2.3 Applications: diffusion in neural tissue

MRI is a powerful tool for diagnosis of brain disorders and for neuroscience research because of the variety of water properties that can be used to create unique image contrast. In most cases, the contrast due to variable water properties (e.g. T1, T2, magnetization transfer, etc.) that result from interactions of the water with the surrounding macromolecular environment yields an indirect or nonspecific indicator that there is something different about a given tissue (e.g. pathological versus normal, white matter versus grey matter, etc.). One such molecular property amenable to MRI measurement and quantification is *diffusion*, which refers to the random, thermally induced "mobility" or Brownian motion of a molecule over time. The diffusing water molecule samples and interacts with the local environment, and thus, by measuring the degree and direction of water motion, the structure can be inferred. For example, if the water encounters highly ordered barriers such that the distance traveled in one direction is greater than that in another direction in the same amount of time, the diffusion is said to be anisotropic. This fundamental property of anisotropic water diffusion is the physical basis behind the utility of diffusion tensor imaging (DTI) in the brain. Although DTI have led to numerous findings of diffusion parameter differences in distinct brain regions of patient populations versus control subjects, often in regions where conventional MRI is not different, the physical interpretation of the diffusion changes is not straightforward. This is particularly true in most clinical studies where there is no access to pathological specimen comparison.

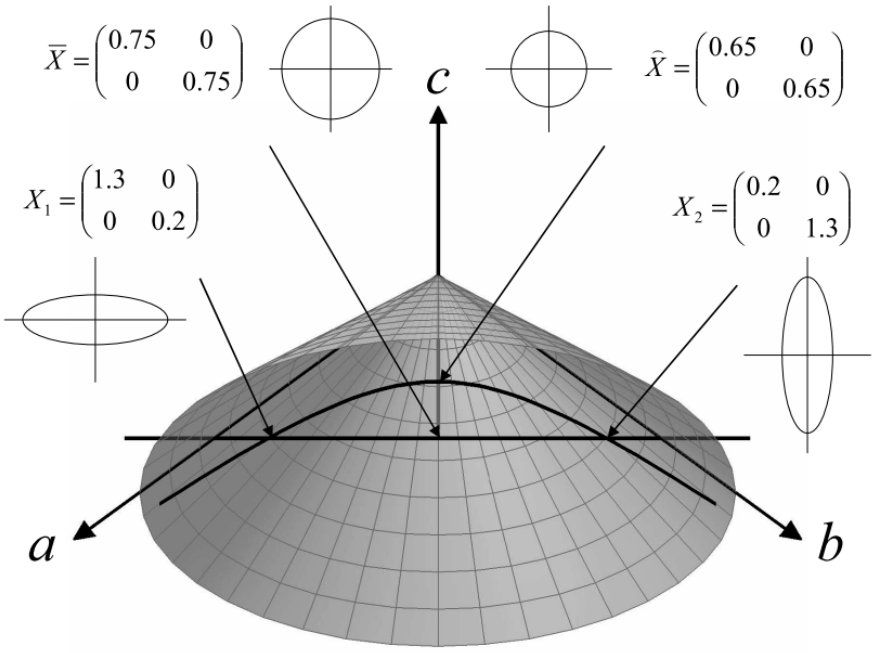


Fig. 2.5: The cone of 2×2 positive definite matrices [163].

2.3.1 Water mobility as a probe of tissue microstructure

Aa early mentioned in this Chapter, molecules can either diffuse equally in all directions, known as *isotropic diffusion*, or they can diffuse preferentially along a particular direction or axis, known as *anisotropic diffusion*. Isotropic diffusion occurs if there a re no barriers to diffusion (e.g. in pure liquid) or if the barriers are randomly oriented, whereas anisotropic diffusion occurs when there are barriers that impede diffusion in certain directions while favoring movement of water along others. Thus the directionality, or lack thereof, of water diffusion can be used as an indirect marker of the status ("integrity") of the tissue microstructure.

A key concept of this physical phenomenon is that there is enough time for the diffusing molecules to *sample* the environment. We have to recall that the Einstein equation 2.7 relates the root mean squared displacement of the molecules with both the diffusion coefficient and the time given for diffusion. At the limit time of very short diffusion times, the molecules do not have time to experience any barriers, hence measurement of their diffusion does not give any indication of their environment. In this case, the mobility of the molecules is a function of the intrinsic diffusion coefficient of the molecule in that solution and the temperature (i.e. energy source for diffusion) (Fig. 2.6). On the other hand, if there is sufficient time for the water molecules to interact with their surroundings, such as the cellular components in tissue that reduce the distance traveled, then the structure can be inferred by measuring the molecular displacement in multiple directions (Fig. 2.6). In this case, MRI measures the apparent diffusion coefficient (ADC)

- apparent because the intrinsic molecular mobility is reduced by these further interactions with the tissue microstructure. At some point, the molecular mean squared displacement may not increase even when using longer diffusion times since the structure does not permit the molecules to travel further (Fig. 2.6).

This "diffusion time" is readily controlled in MRI measurements by the time between the two diffusion-sensitizing gradients; however, the smallest achievable diffusion time for a requisite b-value reflecting the diffusion sensitivity of the pulse sequence is limited by the maximum gradient strength (see Equation 2.10). In most clinical MRI scanner diffusion experiments, the diffusion time is on the order of 40 ms, which corresponds to an root mean squared monodimensional displacement of $\sim 8\mu m$ for water diffusion in the brain, assuming ann ADC of $0.8\text{x}10^{-3}mm^2/s$. Free water at $37°C$ diffuses $\sim 15\mu m$ in 40 ms. As most microstructure within the brain are smaller than these dimensions (e.g. axons are on the order of several μm in diameter), it is clear that MRI (at least in human scans) cannot measure the intrinsic diffusion of water in the brain; on the other hand, as mentioned earlier, this inability is useful because then water diffusion becomes an indicator of the tissue microenvironment.

In order to measure the degree of anisotropy in a well-aligned neural fiber, one must measure the ADC parallel and perpendicular to its length, either by lining up the neural fibers with the gradient axes or by using the tensor model. For cases in which the fibers can be aligned with the magnetic field gradients of the MRI scanner, only two measurements are required (this, in fact, was the case for many early anisotropy studies on excised tissue such as nerve or spinal cord). However, this simplified and time efficient measurement protocol is not possible in more complex situations, such as the intact brain, where not all the fibers can be aligned with the gradient axes. Thus, a minimum of six acquisitions with sensitizing gradients applied (and a non-diffusion-weighted image, the so-called b_0) must be used to measure the diffusion tensor [11].

The causes or biophysical basis of diffusion anisotropy have not been fully elucidated, although most investigators ascribe it to ordered, heterogeneous structures, such as large oriented extracellular and intracellular macromolecules, supermacromolecular structures, organelles, and membranes. Clearly, in the central nervous system (CNS), diffusion anisotropy in white matter is not simply caused by myelin, since several studies have shown that even before myelin is deposited, diffusion anisotropy can be measured using MRI [15]. Thus, despite the fact that increases in myelin are temporally correlated with increases in diffusion anisotropy, structures other than the myelin sheet must significantly contribute to diffusion anisotropy. This is important because the degree of diffusion anisotropy is not a quantitative measure or "stain" of myelin content.

In what follows we briefly overview the clinical applications of the acquisition method we have described so far, i.e. Diffusion Weighted Imaging (DWI) and Diffusion Tensor Imaging (DTI), in order to describe their impact in the clinical practice, but also pointing out their limits and critical aspects. This final conclusion will lead to the final part of this Chapter, with the description of the emerging techniques in the field of diffusion imaging.

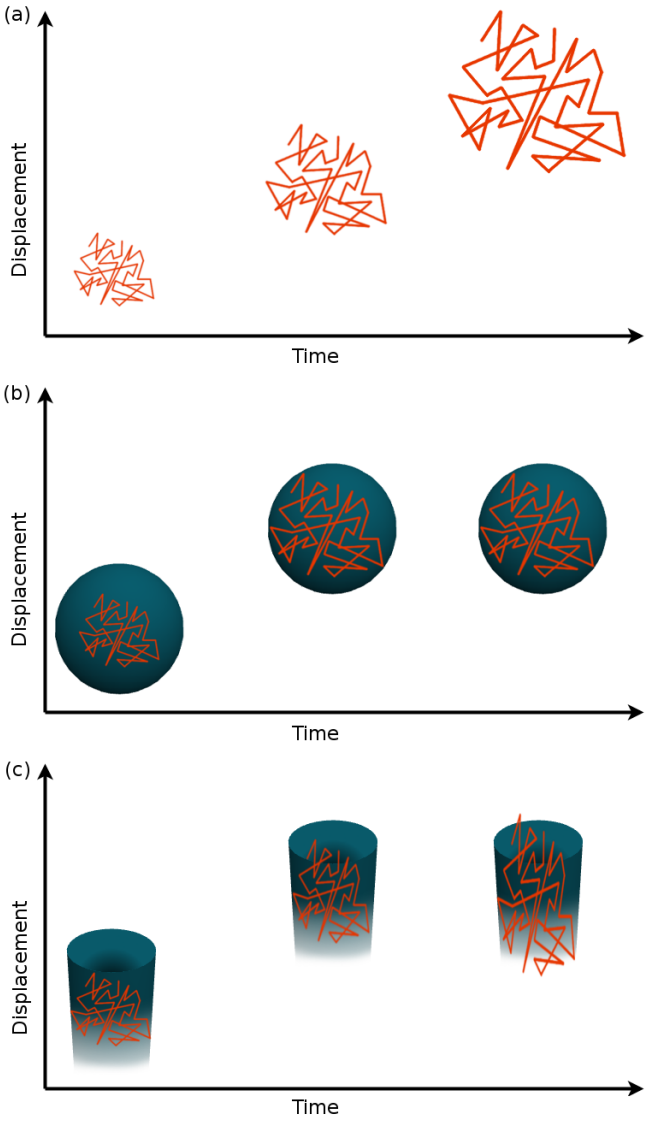


Fig. 2.6: (a) The displacements of the water molecules increase equally in all directions in an unhindered environment when more time is permitted for diffusion. If sufficient time is allowed for the diffusing molecules to be impeded by barriers such as the edges of the sphere (b) and the cylinder (c), then the molecular displacements give an indication of the shape (and orientation) of the structure. At some point, the molecular displacements no longer increase with diffusion time because they are physically restricted from going any further.

2.3.2 Diffusion Weighted Imaging

When the probability density function of displacement of water molecules is Gaussian, we have from the Stejskal and Tanner equation (Eq. 2.9) the following expression

$$\frac{S_g}{S_0} = \exp(-b \cdot ADC) \tag{2.16}$$

where b is a parameter proportional to the intensity and the duration of the gradients applied (Eq. 2.10), S_0 and S_0 are the signal intensities measured with the lower and higher b-values, respectively, and ADC is the apparent diffusion coefficient. In other words, there is a single exponential relationship between the signal and the apparent diffusion coefficient, the coupling factor being the b-factor. It is important to point out at this stage that this is only true for Gaussian diffusion.

When this technique, first developed in NMR spectroscopy, is combined with imaging gradients, the effect is to quantify the apparent diffusion coefficient within each voxel of the imaged volume. As such, it is possible to obtain quantitative maps of the apparent diffusion coefficient. In Fig. 2.7 the results of a diffusion weighted experiments is shown, with the acquisition of the baseline b_0 image (a) and a diffusion weighted image with $b = 1000s/mm^2$ (b). In Fig. 2.7 (c) it is represented the ADC map obtained by solving, for each pixel, Equation 2.16. Areas of restricted diffusion in highly cellular areas show low ADC values compared with less cellular areas that return higher ADC values. It is important to recall that although areas of restricted diffusion will appear to be higher in signal intensity on the directional or index DW images, these areas will appear as low-signal intensity areas (opposite to DW images) on the ADC map.

DWI has greatly improved the evaluation of patients with acute neurologic deficits. It provides unique information on the physiologic state of the brain. To date, it has been extremely valuable in detecting acute ischemic infarction at very early time points and in differentiating acute stroke from a wide variety of disease processes that resemble acute stroke. It has greatly affected patient management and may assist in selecting patients for thrombolysis and in evaluating new neuroprotective agents. It may also prove valuable in a wide variety of other disease processes [28, 66, 114, 120, 162].

Besides its extensive use in neurological applications, from oncology to strokes and hemorrhages, in the last few years diffusion weighted images are increasingly acquired not only for the brain but also for other anatomical districts, such as the breast and the prostate.

2.3.3 Diffusion Tensor Imaging

Microstructure fundamentally affects the apparent diffusion properties of water and so non-invasive quantification of diffusion acts as a sensitive probe to any changes in cellular structures that alter the displacement per unit time. Thus, the introduction of diffusion-weighted imaging was met with great enthusiasm as a non-invasive method of gaining new contrast within the brain. The most useful clinical

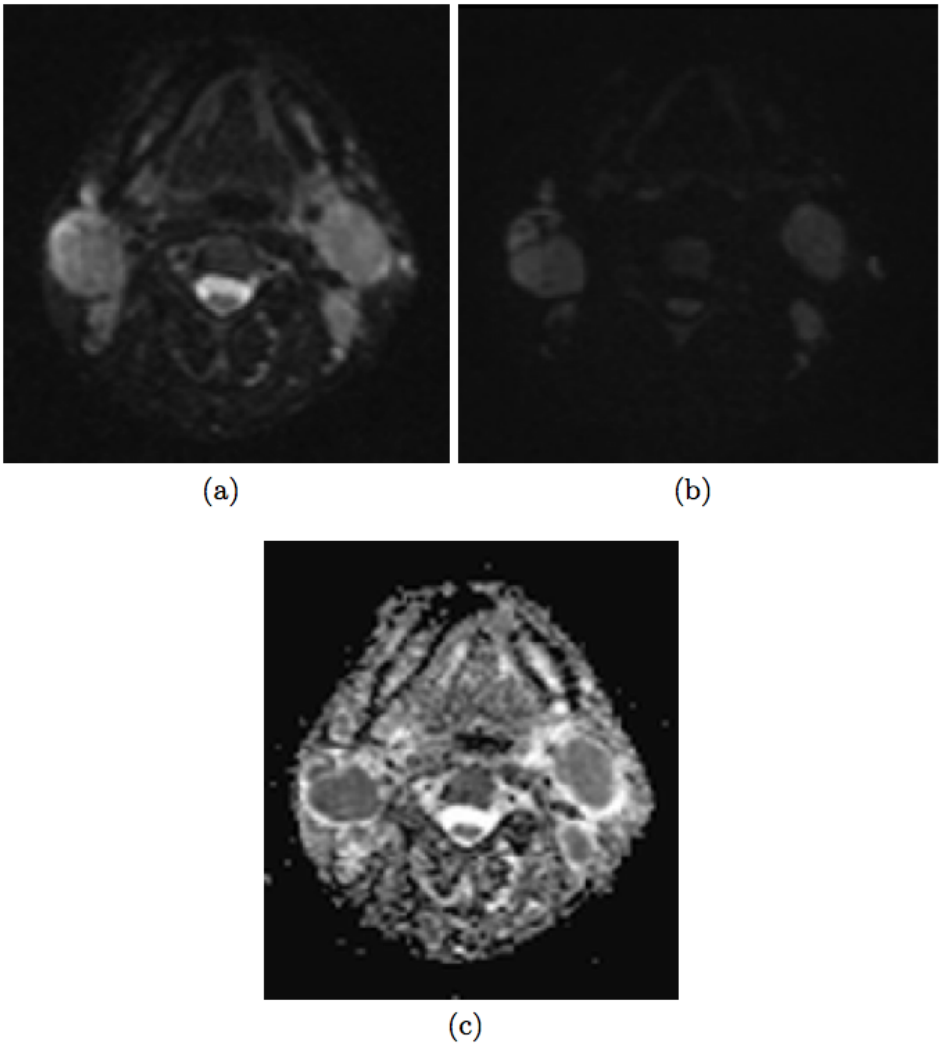


Fig. 2.7: ADC map reconstruction of diffusion weighted MRI of the head: (a) $b = 0$, (b) $b = 1000 s/mm^2$, and (c) ADC map calculated as in 2.16

application to date is the use of the diffusion-weighted scan in acute ischemia in which there is a reduction in the voxel-averaged displacement of water molecules per unit time, hence a reduction in the apparent diffusion coefficient, therefore less signal attenuation - and the lesion appears hyperintense [128], even when "conventional" scans (T1-weighted, T2-weighted, FLAIR) are normal. About the same time as the finding that the ADC was reduced in ischemia, it was observed that, in certain parts of the cat brain, the ADC that was measured depended strongly on the direction in which it was measured (i.e. the direction of the applied diffusion encoding gradient) [129]. These findings confirmed previous ex vivo measurements in muscle and brain tissue made almost two decades earlier by Hansen [70]. Shortly

after Moseleys observation in the cat brain, the directional dependence of the ADC was reported in human white matter [37]. As these observations were initially made within white matter of the mature adult brain, it was understandably first concluded that diffusion anisotropy is the result of myelin, acting as a hydrophobic barrier to diffusion [181]. However things are not so straightforward and in fact anisotropy can be observed when myelin is absent [16].

When the shape of the tissue cells under investigation is not isotropic in our imaged volume, we can no longer characterize the behavior of the water molecules adequately with a single apparent diffusion coefficient. The ADC we measure will depend on the direction in which we measure it. The more constrained along a preferred direction are the tissue components within the sample, the more the ADC will depend on the measurement direction. Therefore, we have to look to a more complex model to characterize diffusion. In this context we can assume the diffusion tensor to characterize Gaussian diffusion in which the displacements per unit time are not the same in all directions.

In the white matter, diffusion MRI has already shown its potential in diseases such as multiple sclerosis [134]. However, DTI offers more through the separation of mean diffusivity indices, such as the trace of the diffusion tensor, which reflects overall water content, and anisotropy indices, which point toward fiber integrity. It has been shown that the degree of diffusion anisotropy in white matter increases during the myelination process [8], and diffusion MRI could be used to assess brain maturation in children [207], newborns, or premature babies [76]. Abnormal connectivity in white matter based on DTI MRI data has also been reported in frontal regions in schizophrenic patients [167] and in left temporo-parietal regions in dyslexic patients [24]. The potential of diffusion MRI has also been studied in brain tumor grading [81], trauma [164] and AIDS [180].

In what follows we overview the methodologies that have been developed in order to make the most of the information provided by DTI, from scalar indexes and diffusion tensor fields to fiber tracking and brain connectivity.

Scalar Indexes

Prior to the introduction of the tensor model into diffusion MRI, several indices for anisotropy of diffusivity were proposed, such as the ratio of ADCs obtained in two orthogonal directions.

The limitation of such indices can be understood by referring to Fig. 2.8. For the fibers oriented at 45 degrees to the $x-$ and $y-$axes, the ratio $\mathrm{ADC}_y/\mathrm{ADC}_x$ is equal to unity, for the fibers oriented along the $y-$axis, the ratio $\mathrm{ADC}_y/\mathrm{ADC}_x$ takes its maximal value, and for the fibers oriented along the $x-$axis, the ratio takes its minimal value. This is, therefore, another example of a measure that is rotationally variant. Anisotropy indices formed from the eigenvalues of the tensor will, by definition, be rotationally invariant. The simplest anisotropy index, analogous to the ratio $\mathrm{ADC}_y/\mathrm{ADC}_x$ would be the ratio of the largest to the smallest eigenvalue (i.e. λ_1/λ_3). However, it has been shown that sorting the eigenvalues according to their magnitude introduces a bias in the measurements at low SNRs [146]. To circumvent this problem, indices that do not require sorting [145] have been proposed and have been shown to be less sensitive to the SNR. A natural choice

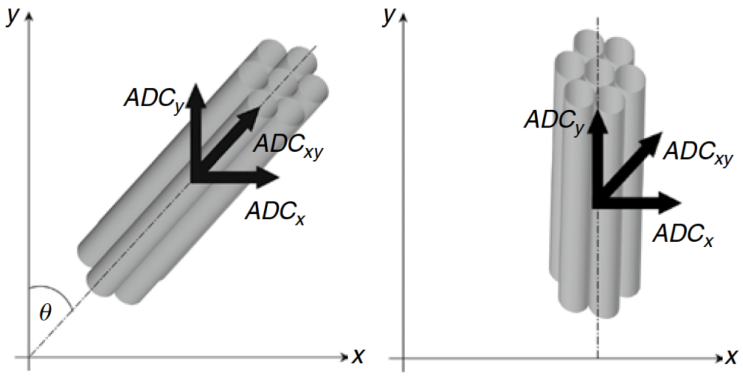


Fig. 2.8: The difference between D_{xy}, the off-diagonal element of the diffusion tensor and the ADC_{xy}, the ADC in the xy direction. In (a), the anisotropic medium is oriented at 45 degrees to both the x- and y-axes. The diffusivity in the x direction is equal to the diffusivity in the y direction, and displacements along the two axes are perfectly correlated (reflected by D_{xy} taking its maximal value). In (b), with the anisotropic medium aligned with the y-axis, displacements along the x- and y-axes are no longer correlated and D_{xy} equals zero. However, the ADC in the direction $[x, y]$ is not zero. Further, while D_{xy} can take negative values, the ADC in the direction $[x, y]$ can, by definition, never take negative values. [89].

is the variance of the three eigenvalues about their mean. As mentioned in the previous sections, this embodies information from all three eigenvalues, yet does not require any to be labeled as the largest or smallest. The variance, however, needs to be normalized to account for regional differences in the overall magnitude of diffusivity.

The two most popular indices based on this concept are the fractional anisotropy (FA) and relative anisotropy (RA) [12]:

$$FA = \frac{3}{2}\frac{\sqrt{(\lambda_1 - <\lambda>)^2 + (\lambda_2 - <\lambda>)^2 + (\lambda_3 - <\lambda>)^2}}{\sqrt{\lambda_1^2 + \lambda_2^2 + \lambda_3^2}} \tag{2.17}$$

and

$$RA = \frac{1}{3}\frac{\sqrt{(\lambda_1 - <\lambda>)^2 + (\lambda_2 - <\lambda>)^2 + (\lambda_3 - <\lambda>)^2}}{<\lambda>} \tag{2.18}$$

where $<\lambda>$ is one third of the trace of the tensor, which is equivalent to another important measure, the Mean Diffusivity (MD). Conceptually MD is equivalent to the Apparent Diffusion Coefficient (ADC) in DWI-MR. MD is low within the white matter, whereas, for example in the ventricles, it is high due to the unrestricted diffusion of the water molecules.

$$MD = <\lambda> = \frac{\lambda_1 + \lambda_2 + \lambda_3}{3} = \frac{Trace(D)}{3} \tag{2.19}$$

The FA index normalizes the variance by the magnitude of the tensor as a whole. The FA index is appropriately normalized so that it takes values from zero (when diffusion is isotropic) to one (when diffusion is constrained along one axis only).

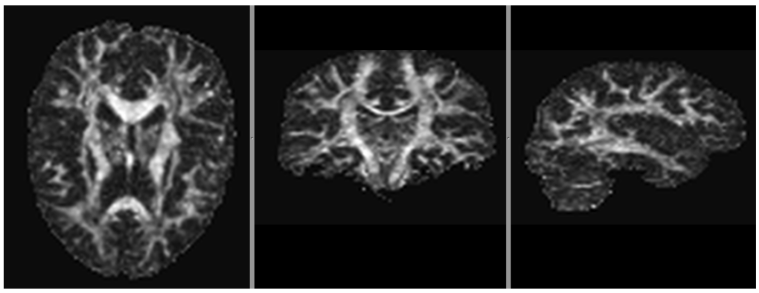


Fig. 2.9: Brain fractional anisotropy data collected in axial, coronal, and sagittal section. The intensity of the image is directly proportional to anisotropy. The CSF filled regions (sulci and ventricles) and gray matter have low intensity as the self-diffusion of water is isotropic at the voxel resolution. In the white matter, where diffusion is more anisotropic, the image appears bright.

The denominator of the RA index is the mean diffusivity. This index is mathematically identical to a coefficient of variation, i.e. standard deviation divided by the mean. To ensure that this index scales from zero to one, an additional scaling factor of $\sqrt{1/2}$ is needed in front of the expression given above for RA. The most commonly used anisotropy index in the literature is the FA. Example images showing FA for the brain in axial, coronal, and sagittal planes are presented in Fig. 2.9. However, it's worth mentioning that even though measures such as FA and RA are less sensitive to noise than measures such as λ_1/λ_3, they are nevertheless sensitive to noise. As the SNR is lowered, the anisotropy indices become increasingly overestimated [145]. In consequence, comparisons of anisotropy indices obtained from different studies in which different imaging parameters have been used should be treated with caution.

This measure of overall diffusion rate can be used to delineate the area affected by a stroke, as demonstrated by Van Gelderen [134]

Diffusion Tensor Fields

In each voxel within an imaging volume we can consider making maps or images of all of the quantitative parameters described earlier in this chapter. In general, with this new spatial information, we are now able to characterize fields of diffusion ellipsoid and other tensor-derived quantities within an imaging volume. Some features involve the spatial rate of change of tensor-derived quantities within the imaging volume, such as the gradient of $< D >$. Imaging methods that apply this idea include direction field mapping, in which the local fiber direction is displayed as a vector in each voxel, and fiber tract color mapping, in which a color,

assigned to a voxel containing anisotropic tissue, is used to signify the local fiber tract direction [20]. An important achievement of color mapping has been the ability to identify unambiguously all major commissural, association and projection pathways in the human brain.

Tractography

Diffusion MR data contains a high quantity of information that has to be processed in order to provide maps of fiber tracts. This essential step towards maps of brain connectivity is called tractography. DTI fiber tractography is a natural extension of diffusion ellipsoid imaging. In brain regions where the ellipsoid are prolate, indicative of well ordered white matter fascicles in which the purported fiber direction is slowly changing from voxel to voxel, we can imagine to connect the ends of these discrete ellipsoids into an extended object, like a link-sausage or natural-pearl necklace, depicting the trajectory of a large fiber tract. This process results in lines or trajectories capturing coherent orientations of maximal diffusion that are likely to represent real axonal trajectories. The main bottleneck of this method is that by comparison with invasive techniques, tractography measurements are indirect, difficult to interpret quantitatively, and error prone. In fact we have to keep in mind that there are several orders of magnitude between the resolution of the MR acquisitions and the diameter of the axons. Therefore, tractography is only able to map large axonal bundles, and a single fiber produced by any algorithm is in fact representative of a huge coherent set of real anatomical trajectories. However, their non-invasive nature and ease of measurement mean that tractography studies can address scientific and clinical questions that cannot be answered by any other means.

When going through specialized literature, it is striking to see the huge quantity of tractography algorithms that have been proposed, demonstrating (i) the great interest it has raised in the scientific community and (ii) the variety of strategies proposed to optimally extract information from the diffusion MR images.

All diffusion tractography techniques rely on one fundamental assumption: when a number of axons align themselves along a common axis, the diffusion of water molecules will be hindered to a greater extent across his axis than along it. Since its introduction, several different schemes have been proposed to follow fiber tracts. The earliest methods proposed for fiber-tract following were deterministic. In this so-called *streamline* approaches, fiber tract trajectories are generated from the local fiber-tract direction field much in the same way fluid streamlines are generated from a fluid velocity field. Starting from a "seed point", fibers are launched in both directions until some stopping or "termination" criteria are satisfied, such as the FA dropping below its level in background noise. A *fiber tract* is the name given to all points along such a continuous trajectory. One disadvantage of this computational strategy is that errors can accumulate during the tracking process. In general, it cannot be assured that the trajectories represent actual or even probable pathways of nerve fibers. This is an important issue that limits the validation of the whole diffusion MRI analysis and therefore prevents the wide exploitability in the clinical practice.

The second strategy is the probabilistic one [31]. A seed point is assumed to be connected to all points within the imaging volume, but the most probable con-

nections are those that minimize some cost function. This approach holds great promise in that many paths are explored, but on,y those that are frequently traversed are assigned high likelihood. Of course, a disadvantage of this approach is that we do not know what physical constraints nature uses to construct nerve pathways. The development of probabilistic approaches that do not rely on functional minimization appear to be quite promising, although further validation studies still should be performed.

It is important to note that a number of well- documented artifacts in DTI fiber tractography arise as discrete, coarsely sampled, noisy, voxel-averaged direction field data [108]. These artifacts can produce "phantom" connections between different brain regions that do not exist anatomically (a "false positive"), or they can result in missing anatomical connections that do exist (a "false negative"). There are usually a number of thresholds and free parameters that can be set in existing tractography codes whose adjustment can alter one's findings. Therefore great care must be exercised in obtaining "anatomical" connectivity with fiber tractography [106]. Especially clinical users show skepticism when interpreting such data - less in the coherent primary pathways, but increasingly in finer and more complex white matter structures.

Structural Connectivity

Structural connectivity refers to the a set of physical or structural (anatomical) connections linking neural elements. The physical pattern of these anatomical connections may be thought as relatively static at shorter time scales (seconds to minutes) but may be plastic or dynamic at longer time scales (hours to days) - for example during development or in the course of learning and synaptic remodeling. Even if the function of the brain cannot be reduced to its wiring diagram, brain networks shape patterns of spontaneous and evoked neural activity. Therefore, in order to better understand how brain networks function, one must first know how their elements are connected. A principle goal of a comprehensive description of the structural network of the human brain, the so-called *connectome* [171], is the representation of structural brain networks in the form of graphs, collections of nodes and edges, which allow the quantitative analysis of brain connectivity with the mathematical tools of network science. The connectome is an approach to reveal structural principles of brain networks that illuminate brain function, not merely a database of "what connects to what". In this direction, the signal generated by diffusion imaging can provide information about the direction of fiber tracts within individual voxels of the brain. The spatial resolution of the signal is limited by the voxel size and could be improved by imaging at higher field strength. However there are more challenging issues that limit the validation of connectivity results and limit their use in the clinical practice:

- A more fundamental limitation encountered in DTI is that the diffusion tensor captures only a single diffusion direction per voxel, which does not account for crossing fibers.
- Another issue comes from the fact that diffusion MRI and tractography data are often difficult to validate against more "classical" invasive anatomical tech-

niques, such as tract tracing [138], and the comparison of the different methods is still limited to few studies on animals, leaving the validation problem still open.

- Moreover, in the field of diffusion-based connectivity analysis, a direct comparison of the results obtained in the different studies is made difficult because of the adoption of different grey-matter parcellation schemes (which define the network nodes and hence their connections as well) and acquisition methods.

To conclude DTI has brought new possibilities and methods to better understand the structure of the brain. However proceeding in the analysis of brain anatomy from the DTI perspective is also showing the limits of this methodologies, while facing more complex problems and scenarios. In order to face this issues, especially connected to the regions of the brain characterized by fiber crossings, several methods have been recently developed. In what follows we briefly describe the problem of fiber crossing and then we overview the state-of-the-art solutions, by mainly focusing on Diffusion Spectrum Imaging.

2.4 Dynamic Contrast-Enhanced MRI

Pathological conditions frequently involve the microvasculature and associated blood flow of the affected tissues. Angiogenesis is a prominent example from cancer models, for without the development of new blood vessels, tumors are incapable of growth beyond the diffusion range of oxygen (1), which is a few tenths of a millimeter (11). The growth and decay of the tumor vasculature can therefore provide information about the response of tumors to therapy, and the use of non-invasive imaging modalities to probe the microvascular characteristics of tissue is critical to biomedical research (2). This may be achieved by exploiting the fact that vessels produced by tumors are often malformed and leaky, resulting in hyperpermeability to small molecules, such as contrast agents (12), producing signal changes which may be observed using advanced imaging techniques [151].

Dynamic contrast enhanced magnetic resonance imaging (DCE-MRI) is a MRI methodology using 2D or 3D T1-weighted MR sequences to record the uptake of Gd-based contrast agents(CA) during and after intravenous administration. It is achieved by exploiting the T1-shortening effects of contrast agents, while a focus on the T2*-shortening effects is the basis for Dynamical Susceptibility Contrast (DSC-MRI). DCE-MRI highlights the dynamic response of the tissue to the inflow of the agent, which are reflected in a change of contrast in the resulting imaging volumes. This feature can be missed with a conventional contrast enhanced (CE) MR image, which is acquired after most of the distribution has been accomplished and some of the contrast has already been washed out [105].

DCE-MRI with its capability of revealing microvasculature and perfusion in soft tissues, has been well established as an important non-invasive tool in various tumor entities [30, 78, 98, 151]. various methods have been proposed to describe and evaluate the time-intensity curves of contrast agent enhancement, including qualitative descriptions [98, 104] and pharmacokinetic compartment model-based quantitative parameters [22, 183, 184].

Because of signal intensity enhancing characteristics in the tumor regions and the capability to noninvasively assess the tumor vascular characteristics, it has been considered as a potential candidate for an imaging biomarker in the diagnosis and detection of the tumors. The use of DCE-MRI has been advocated by an increasing number of investigators studying physiological processes that involve neoangiogenesis, such as cancer and inflammatory processes. DCE-MRI has been considered for monitoring the therapy response of the tumors during a variety of treatment planning including chemotherapy, radiotherapy or anti-angiogenesis inhibitors treatment [78].

2.4.1 Contrast agents in MRI

MR contrast agents are the chemical compounds introduced to the anatomical or functional region being imaged in MR, towards the aim of increasing the signal intensity differences between the different tissues [17, 18]. The most important requirements of MRI contrast agents include: relaxivity, tissue specificity, excretability and lack of toxicity [19].

MRI contrast agents are not directly visible. The modification of contrast is due to their effects of shortening the relaxation time T1 or T2 of the hydrogen nuclei located in their vicinity. The enhanced relaxivities by the metal complex-based contrast agent must be sufficient to increase the relaxivity of the target tissue. The dose of complex-based contrast agent should be nontoxic to the patient, and the contrast agent should increase the relaxation rate at least $10 - 20\%$ to be detectable by MRI.

Contrast in the MR images depends mainly on the proton spin density and the T1 (longitudinal) and T2 or T*2 (transverse) relaxation times. These characteristics contribute to the weighting components that generate the T1-weighted and T2-weighted images in MR. If the contrast agent reduces time T1 , we observe an increase in the T1 signal. On the other hand, if it shortens T2 , there will be a reduction in the T 2 and T*2 signal. The effectiveness of the contrast agent mainly depends on its capacity to modify relaxation times. Due to these basic principles of modification of T1 and T2 relaxation times, two main classes of contrast agents can be distinguished.

MRI contrast agents can be defined of as biotracers or magnetic dyes and must be biocompatible pharmaceutical and magnetization relaxation probes. The complexbased contrast agent, if possible, should only localize the targeted tissue and its compartments in preference to nontargeted regions for better delineation and diagnosis of the tissues.

Metal-based MR contrast agents should be used in a chelated form to prevent the accumulation of toxic metals in the body, and they must be safely excreted within hours of administration. Biodistribution of the contrast agents should be initially intra-vascular with a rapid passage to the interstitial sector and should not pass the healthy blood-brain barrier [110].

2.4.2 Analysis of DCE-MRI

2.4.3 Pharmacokinetic Modeling

Dynamic contrast enhancement patterns can be affected by a wide range of physiological factors which include vessel density, blood flow, endothelial permeability and the size of the extravascular extracellular space in which contrast is distributed [72, 103, 178]. A quantitative analysis aims to directly measure physiological parameters such as tissue blood flow, blood volume, interstitial volume or permeability-surface area. The goal of such kind of model is to approach the *true* (absolute) values underlying the patophysiologic processes that are being measured and they are specially useful in therapy assessment.

Extraction of hemodynamic parameters from DCE-MRI data requires the calculation of contrast medium concentration as a function of time $C(t)$ either in each image voxel or in regions of interest *(ROI)*. $C(t)$ is analyzed based on various pharmacokinetic models from which hemodynamic parameters, such as perfusion rate, blood volume and capillary permeability, are extracted. However, the accuracy of such parameters depends on an appropriate theoretical model and related assumptions used to interpret data [49].

The main perfusion parameters are the *tissue plasma flow* (ml/100g/min), which measures the volume of plasma flowing through the capillaries of a given amount of tissue per unit of time; and the *tissue plasma volume* (ml/100g), which measures the volume of plasma in the capillary bed of 100g of tissue. The main permeability parameters are the *extravascular, extracellular volume* (ml/100 g), and the *permeability-surface area product* PS (ml/100 g/min), which measures the volume of plasma flowing across the capillary wall of 100g of tissue per unit of time [170].

These hemodynamic parameters are linked to the measured DCE-MRI signal through *tracer-kinetic theory*, which relates the hemodynamic parameters to the time-concentration curves in the tissue and through *MRI signal theory*, which relates those concentrations to changes in MR signals. [170]

The form of the tracer concentration-time curves $C(t)$ in the tissue are determined by the hemodynamic parameters, the concentration $C_A(t)$ in the blood plasma of an arterial vessel feeding the tissue which form the *arterial input function* (AIF). The theory of linear and stationary systems is valid for tracer-kinetic analysis as long as the response of the tissue to an injection of tracer at any given time is proportional to the injected dose and independent of the time of injection, that is, complies with linearity and stationarity. $C(t)$ and $C_A(t)$ are related by convolution with a residue function $R(t)$:

$$C = F_P C_A \otimes R \tag{2.20}$$

Here F_P is the tissue plasma flow, and $R(t)$ is the fraction of contrast agent concentration left in the tissue at time t for the case of an ideal instantaneous dose injected at time $t = 0$. The residue function is a tissue characteristic that fully defines the kinetics of a particular tracer. It is always a positive, decreasing function which satisfies $R(0) = 1$. Eq.(2.20) is valid only if AIF is measured at the inlet to the tissue. From a measurement perspective, the main complication for a

quantification is the need to accurately measure the concentration in the lumen of a major feeding artery, usually $C_A(t)$ is measured at a more upstream location. Also, additional post-processing steps are required: MRI signal analysis to calculate or approximate the tracer concentrations from the MRI signals, and tracer-kinetic analysis to determine quantitative parameters from the concentrations [170, 183].

Tracer-kinetic models provide a representation of the residue function in terms of the hemodynamic parameters. In most tissue types, tracers distribute over two different spaces: the blood plasma P and the extravascular, extracellular space E. The two compartment model is defined by the assumption that P and E are compartments, E does not exchange tracer directly with the environment, the clearance for the outlets connecting P and E are equal and the clearance for the outlet of P to the environment equals the plasma flow [21,170]. A two compartment model can be reduced to a single compartment in the following situations: when one of the spaces has a negligible volume, if the tracer extravasates slowly, so that the concentration in the extravascular space is negligible within the acquisition time, and if the tracer extravasates rapidly, so the system behaves as a single well mixed space. In these cases the residue function becomes monoexponential, but the precise interpretation of the parameters differs [170, 183].

Tofts et al. [183, 184] defined the parameters which now serve as the golden standard for quantitative parameter extraction in the frame of a pharmacokinetic model. The *Tofts* model is a one-compartment representation of the physiologic quantities that determine the dynamic behavior of contrast agent. In the Tofts model the transfer constant K^{trans} is a combined measure of blood flow and capillary permeability and v_e the extravascular extracellular space (EES) fractional volume. Eq.(2.21) represents the standard Tofts model whereas the modified model includes the term $v_p C_A(t)$ to account for the tracer in the vasculature Eq.(2.22).

$$C(t) = K^{\text{trans}} \int_0^t e^{-\frac{K^{\text{trans}}}{v_e}(t)} C_A(t) \tag{2.21}$$

$$C(t) = K^{\text{trans}} \int_0^t e^{-\frac{K^{\text{trans}}}{v_e}(t)} C_A(t) + v_p C_A(t) \tag{2.22}$$

Comparisons between different pharmacokinetic models can be found in the literature [58,113,172]. The choice of a model is not only determined by the state of the tissue, but also by the quality of the data [73, 116]. In particular, the injection protocol, temporal resolution, acquisition time and noise level play an important role. For instance, if the injection rate is too slow, intra and extravascular spaces are in constant equilibrium, and the monoexponential Tofts model must be applied (2.21). Conversely, a more rapid injection rate may create strong intra and extravascular concentration differences in the first pass, so that only a full biexponential model provides a good fit. This implies that ambiguities in the choice of a model may be resolved by appropriate optimization of the measurement protocol [170, 183].

From a statistical point of view, quantitative methods are based on the theory of nonlinear regression. Nonlinear models are typically difficult to estimate due to convergence issues and consistency problems by specifying starting values for the optimization [159].

Wang et al. [190] proposed a quantitative method for estimating the input function and the kinetic parameters. Their model is on the pixel domain, whose parameters are initialized using a sub-space based algorithm and refined by an iterative maximum likelihood estimation procedure.

Schmid et al. [159] proposed a statistical method for estimating parameters in pharmacokinetic models using nonlinear regression not only within a traditional (likelihood) framework, but also with a Bayesian inference. To obtain an improved distribution of kinetic parameter estimates across the image, contextual information from neighboring voxels are combined by incorporating a Gaussian Markov random field (GMRF) prior on the kinetic parameters into the model. Kinetic parameter estimations using neighboring voxels reduce the observed variability in local tumor regions while preserving sharp transitions between heterogeneous tissue boundaries. The proposed model is described in three stages: the data model, the process model, and the prior parameters. For the process model it is assumed that functions or, more precisely the pharmacokinetic parameters in neighboring voxels are similar. As voxel borders are arbitrary and do not correspond the borders between tissue types, this is a reasonable assumption. Each voxel is assumed to be a mixture of different tissue types and an unique parameter is assigned for each neighboring voxel-to-voxel combination.

In 2009, the same group of Schmid et al. [160] proposed another model for quantitative analysis; a semi-parametric penalized spline smoothing approach, where the arterial input function (AIF), which is approximated mathematically, is convolved with a set of B-splines to produce a design matrix using locally adaptive smoothing parameters based on Bayesian penalized spline models (P-splines). It has been shown that kinetic parameter estimation can be obtained from the resulting deconvolved response function. Their model captures the upslope of the time series accurately, which is important for the accurate fit of the concentration time curves and proper calculation of K^{trans}. The proposed technique also intrinsicly allows for quantification of estimation errors both in fitting the observed data and in estimating kinetic parameters.

These semi-parametric, statistical and analytic approaches provide more flexibility and consistency since they can incorporate contextual information to reduce estimation errors and bias [160].

2.4.4 Qualitative and Semi-Quantitative Analysis

Generally speaking, data can be analysed by visual assessment, descriptive parameters, or quantitative parameters. Many quantitative or semi-quantitative approaches for the classification of enhancement curve shapes have been described and are now in relatively common use in clinical settings. Descriptive parameters are indices that characterize the shape and structure of the curves, such as the time to peak enhancement, bolus arrival time, maximum upslope, maximum downslope, area under the curve, or maximum enhancement. Deriving descriptive parameters is straightforward, but the link to physiology is not always clear, and they are only reproducible when an identical measurement protocol is used [170].

Although the quantification of DCE MRI data by means of pharmacokinetic models is well suited to longitudinal studies as well as to comparison between

studies, it also suffers from large output variability, which is a consequence of the large variety of models used [23].

As described by Lavini [104], an alternative halfway analysis methods between the pharmacologic quantitative analysis and the mainly qualitative methods has been targeted by some authors who have investigated uptake curve shape (or pattern) and who tried to relate them to pathological findings. This approach is based on the observation of the timeintensity curve (TIC) generated from an ROI chosen in the lesion by the radiologist.

Although not quantitative, this approach is less sensitive to variations in the MRI protocol (although still dependent on the duration of the scan and on the injection procedure), making it more suitable for a comparative (meta-) analysis. Furthermore, as shown in the simulations in Tofts et al. [184], curve shapes represent a mirror of those physiological parameters (e.g., the capillary permeability) that can be extracted by means of the abovementioned analysis using compartmental models. In fact, increased tumor angiogenesis has often been associated with a specific TIC pattern with rapid wash-in and washout [72], which is described by a large ktrans in these compartmental models. TIC analysis has, therefore, been often used as a surrogate for quantitative analysis [100].

Quantitative or descriptive parameters can be calculated on the level of voxels, or of regions of interest (ROI). The classic evaluation of the enhancement kinetics is carried out by the analysis of time-intensity curves using ROI analysis. A region is outlined manually and the time-intensity curves of all voxels in the ROI are averaged to produce one single curve. The post-processing protocol is then applied to this curve. For qualitative evaluation, the curve's shape is classified into different morphological categories such as: steady, plateau and washout. For the evaluation of washout curves, the ROI should be manually selected based on the (postcontrast) anatomical scan. It is placed in the area of a tumor with the most rapid and intense enhancement, an area that can, in principle, contain pixels with different enhancement patterns. In many cases, the suspicious tumor is very heterogeneous and the most appropriate part of the tumor cannot be determined easily. To avoid the distortion of the average time-intensity curve, the ROI should be very small and must not include dead tumor cells (necrosis) or surrounding tissue.

The use of ROIs include several disadvantages; the heterogeneity of tumor vascularization, the close neighborhood of necrotic and vital tumor tissue, and the subjectiveness of ROI placement. The combination of these factors harden the interpretation of the kinetics and may even result in unrevealed malignant tissue. However, the main disadvantage is an intrinsic part of the technique; the averaging of information. This is specially detrimental when a ROI covers malignant and benign tissue, with an average curve shape indicating a false benignity. In Fig.(2.10), an example for ROI placement is provided. If the radiologist has detected a possibly malignant tumor, core needle biopsy must be carried out to confirm or reject his hypothesis due to the moderate specificity of DCE-MRI. Only if the histopathologic report of the removed tissue confirms the radiologist's report, further treatment, e.g. surgery, will be carried out. Since the application of core needle biopsy to a benign part of a malignant tumor results in a false report, it is not only important to determine the most malignant part of the tumor, but also the extent and the localization of this part. Although it is important to

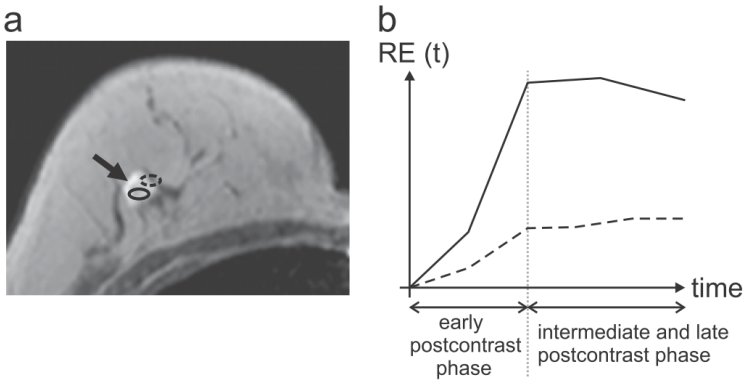


Fig. 2.10: Evaluation of the enhancement kinetics in a tumor. In (a), a slice of a breast DCE-MRI dataset with a breast tumor (see brighter area and arrow) is depicted at the first time point after the early postcontrast phase. For diagnosis of the enhancement kinetics, the radiologist places regions of interest (ROI) (see black and dashed black ellipses) and evaluates the relative intensity increase, depicted in(b). The black ROI is characterized by a suspicious curve with a fast washin and a moderate washout. In contrast, the dashed black ROI covers benign tumor tissue as well as surrounding tissue and the corresponding curve is not suspicious [62].

preserve sensitivity,for clinical practice it is as vital to prevent an unduly high rate of false-positive biopsy calls. Accordingly, specificity is a major issue in breast MRI [62, 99, 100, 130].

For a voxel-based analysis, a curve is extracted for each voxel. The post-processing protocol is applied to each voxel-curve individually [42]. A common technique is to produce an image or map for each calculated parameter from pharmacokinetic models; such as K^{trans}, maximum enhancement or area under the time-intensity curve (not to confuse with the maps obtained through classification after feature extraction [104]). In contrast with the ROI based approach, the main advantage of a voxel-based analysis is that it produces information on the heterogeneity of perfusion or permeability within the organ or tissue [35].

From a signal processing point of view, being n_s the number of acquired volumes, the n_s signal values at voxel $\mathbf{p} = (\mathbf{x}, \mathbf{y}, \mathbf{z})$ are interpreted as a point $\mathbf{x}$ in a n_s-dimensional space, that is also often referred to as a feature vector $\mathbf{x} \in \mathbf{R^{n_s}}$.

Moate et al. [125] created a logistic model with a modified logistic equation that describes the signal enhancement in DCE-MRI, defined in Eq.(2.23).

$$SI = \frac{P_1 + (P_5 \cdot t)}{\{1 + exp(-P_4 \cdot (t - P_3))\}} + P_1 \tag{2.23}$$

Where $SI(t)$ is the signal intensity at time t, P_1 approximates the baseline signal intensity, P_2 is the amplitude of the plateau above the baseline, $P_3(s)$ is the time at which the maximum slope occurs and $P_4(s^{-1})$ is the maximum slope. The inclusion of $(P_5 \cdot t)$provides the flexibility to describe signal intensity curves with either increasing or decreasing terminal slope.

3

Relevant elements of Unsupervised Classification

3.1 Overview

In this chapter the relevant background of unsupervised classification is described. We start focusin on the important issue of data representation and the specific paradigm on which our medical imaging methodology relies, that is, dissimilarity representations. Later the basic clustering problem is described, followed by cluster ensembles. The chapter finalizes with the issue of validation in unsupervised classification.

3.2 Introduction

We are living in a world full of data. Every day, people encounter a large amount of information and store or represent it as data for further analysis and management. Pattern recognition is an intrinsic human ability that starts in infancy. It takes however a long development time before we can accurately describe how we do this, and perhaps sometimes we are not able to outline it.

As one of the most primitive activities of human beings, classification plays an important and indispensable role in the long history of human development. In order to learn a new object or understand a new phenomenon, people always try to seek the features that can describe it, and further compare it with other known objects or phenomena, based on the similarity or dissimilarity, generalized as proximity, according to some certain standards or rules. Basically, classification systems are either supervised or unsupervised, depending on whether they assign new inputs to one of a finite number of discrete supervised classes or unsupervised categories, respectively [194].

In supervised classification, the mapping from a set of input data vectors ($x \in \mathbb{R}^d$, where d is the input space dimensionality), to a finite set of discrete class labels ($y \in 1, \ldots, C$, where C is the total number of class types), is modeled in terms of some mathematical function $y = y(x, w)$, where w is a vector of adjustable parameters. The values of these parameters are determined (optimized) by an inductive learning algorithm (also termed inducer), whose aim is to minimize an empirical risk functional (related to an inductive principle) on

a finite data set of input - output examples, $(x_i, y_i), i = 1, \ldots, N$, where N is the finite cardinality of the available representative data set.

In this chapter we will focus on the second category, unsupervised classification, the process of grouping objects into clusters such that objects from the same cluster are similar and objects from different clusters are dissimilar.

The ability to form meaningful associations between objects starts with the important question of how the data should be formally represented to create a discriminating methodology.

3.3 Data Representation

In theories of information processing, the importance of choosing the right representation for a given computational problem is widely acknowledged.

Following Cummins [33], an important question arises in the problem of representation: how, in principle, can an internal state of a system refer to anything at all in the external world. The philosopher John Locke suggested that an idea represents a thing in the world if it is naturally and predictably evoked by that thing, and not necessarily, as the Aristotelians would have it, if the idea resembles the thing in any sense [44]. This undertaking, putting Representation on a principled basis that does not presuppose reconstruction, is a challenging philosophical and computational problem.

Let us consider a situation in which an observer recalls a previously encountered scene that contained a cat. If the representational story is at all true, the observer harbors an internal representation of the cat (or, as it may be, of the class of cat-like objects). The first question that suggests itself in this context is, what can it be about the internal state of the observer recalling a cat that makes it refer to the shape of the cat? The question is abstract, as it deals with a possible mode of representation, and not with representational means actually used by any particular system. Very few people these days believe that a representation of a cat in an observer's brain is cat-shaped, striped, or fluffy. Instead of little pictures in the head, a representation is seen as a set of measurements which collectively encode the geometry and other visual qualities of its target. Typically, it is assumed that structural or metric information stored in the brain reflects corresponding properties of shapes in the world, *on a one to one* basis. Edelman refers to this approach as "first-order structural isomorphism" between the representation and its target object [44].

The main challenge in devising a representation suitable for supporting categorization is the need for abstraction. Theories of representation that treat categorization as a *sui generis* problem and not as an appendix to identification, usually start from the notion of prototype: the most typical member of a class of previously encountered stimuli, or perhaps an abstraction that serves as a surrogate member and is charged with representing a specific class [44].

As a relevant example, even experts such as cardiologists have difficulties in defining accurately how a particular ECG signal is consciously recognized as evidence of a heart disease. In this as with many other cases, even though the human expert is aware of different pattern classes, he finds himself in trouble when asked to put forward a description of the perceived class in terms of explicit observations.

A common approach is to rely on descriptions based on the structure of the objects, that is, the relations between internal parts when complicated elements need to be described. This is common in cases where there is an inability to depend on a straightforward set of directly measurable observations such as color, weight or size. In this way, characteristics are not only expressed in measurable quantities, but also in a discursive description of the structure. If we want to incorporate this approach to represent data in way that is useful for a classification or clustering system, both, a set of sensors as well as a structural model may be necessary [41].

It is difficult for an expert to define exactly how sensor outputs have to be combined to accurately describe a pattern class or category. Generallyt this works well for objects represented in vector spaces by measurements or by features derived from measurements. In supervised classification, the lack of structural knowledge or the lack of its representation may hereby be partially compensated by statistical properties derived from a (large) set of examples. On the contrary, it is much more difficult to apply such procedures in order to optimize decision functions based on structural models. In order to build automatic machines that mimic human recognition, the expert is forced to become gradually more and more aware of his own decision making, while he tries to make his recognition process explicit. In this process he becomes more conscious of his own internal recognition procedure. The result is a description in terms of both observations and models.

The fact that human experts, when tasked to explicitly decscribe their discriminating faculties, experience a consciuous division of their knowledge into observations and structural models may lead to clear and computerizable representations, but has also severe drawbacks. Observations originated from structural relations may be represented by vectors related to sensors or sensor samples. This representation is poor as dependencies are not included. They may be partially reconstructed from a statistical analysis of a large set of observations, such as it is done when learning from examples in supervised classification. Structural models, on the other hand, may preserve dependencies and relations, but it is difficult to enrich such a knowledge-based description by new observations [41].

This dichotomy of methodological approaches led decades ago to a separation between research areas in machine learning and pattern recognition: structural and statistical [41].

Both approaches use features to describe objects, but these features are defined differently (Table 3.1). The statistical, decision-theoretical approach is usually metric and quantitative, while the structural approach is qualitative [131, 142]. This means that in the statistical approach, features are encoded as purely numerical variables, in which an object is represented by the results of measurement of its various properties. A measurement result is called a *feature* in pattern recognition or a *variable* in statistics. The concatenation of all the features of a single object forms the *feature vector*. By arranging the feature vectors of different objects in different rows, we get a pattern matrix (also called data matrix) of size n by d, where n is the total number of objects and d is the number of features. This representation is very popular because it converts different kinds of objects into a standard representation. This constitute a feature vector space, usually Euclidean, in which each object is represented as a point of feature values. In this case classification is then inherently restricted to the mathematical methods that one can

apply in a vector space, equipped with additional algebraic structures of an inner product, norm and the distance.

In any structural decomposition model, an object is described in terms of relatively few primitives, fundamental structural elements, like strokes, corners or other morphological elements. The primitives are encoded as syntactic units from which objects are constructed and joined by relationships that are chosen from an equally small fixed set. As a result, objects are represented by a set of primitives with specified syntactic operations. For instance, if the operation of concatenation is used, objects are described by strings of (concatenated) primitives. A crucial characteristic of the structural approach is the standardization of the primitives (the parts and their relationships), which allows novel objects to be treated on par with familiar ones, as required, for example, in categorization tasks [41, 44].

The strength of the statistical approach relies on well-developed concepts and learning techniques, while in the structural approach, it is much easier to encode existing knowledge on the objects.

As we have mentioned, representations in Euclidean vector spaces, defined by a set of features, are well suited for generalization. These should ideally characterize the patterns well and also be relevant for class differences at the same time. Such features have to be defined by experts exploiting their knowledge of the application. A drawback of the use of features is that different objects may have the same representation as they differ by properties that were not expressed in the chosen feature set. This results in class overlap: in some areas in the feature space objects of different classes are represented by the same feature vectors. Consequently, they cannot be distinguished, which leads to an intrinsic classification error, usually called the Bayes error. In the next Section we make an overview of a different paradigm proposed by Pekalska and Duin which intends to bridge the structural and statistical approaches, avoinding at the same time overlaps in the representation of diverse objects.

3.4 Dissimilarity-Based Representation

Pekalska and Duin, inspired by early work of Goldfarb [64], initiated a research to bridge the structural and statistical approaches to pattern recognition by replacing the traditional feature representation with a distance representation that could be applied to structural models. They called it the *dissimilarity representation* as it allows various non-metric, indefinite or even asymmetric proximity measures.

The notion of similarity plays a pivotal role in class formation, since it might be seen as a natural link between observations on objects on the one hand and a judgment on their shared properties on the other. In essence, similar objects can be grouped together to form a class, and consequently a class is a set of similar objects. However, there is no such thing as a general object similarity that can be universally measured or applied .

A comparison of two objects is always with respect to a frame of reference, i.e. a particular point of view, a context, basic characteristics, a type of domain, or attributes considered [142].

An inspiration for this approach was also the observation that a human observer is primary triggered by object differences and that the description in terms

Table 3.1: Basic differences between statistical and structural Pattern Recognition [131]. Distances are a common factor used for discrimination in both approaches [142].

Properties	Statistical	Structural
Foundation	Well-developed mathematical theory of vector spaces	Intuitively appealing: human cognition or perception
Approach	Quantitative	Qualitative: structural / syntactic
Descriptors	Numerical features: vectors of a fixed length	Morphological primitives of a variable size
Syntax	Element position in a vector	Encoding process of primitives
Noise	Easily encoded	Needs regular structures
Learning	Vector-based methods	Graphs, decisions trees, grammars
Dissimilarity	Metric, often Euclidean	Defined in a matching process
Discrimination	Relies on distances or inner products in a vector space	Grammars recognize valid objects; distances often used
Class overlap	Due to improper features and probabilistic models	Due to improper primitives leading to ambiguity in the description

of features and models comes second. We consciously observe differences, while similarity is a usually assumed context in which comparisons take place.

The emphasis of the renewed interest in dissimilarities in pattern recognition, however, was in the construction of vector spaces that are suitable for classification using the extensive theory and toolboxes available in multivariate statistics, machine learning and pattern recognition.

An alternative to the use of features is the dissimilarity representation based on direct pairwise object comparisons. If the entire objects are taken into account in the comparison, then only identical objects will have a dissimilarity zero. For such a representation class overlap does not exist if the objects can be unambiguously labelled: there are no real world objects in the application that belong to more than one class. Only identical objects have a zero-distance and they should have the same label as they are identical. Another advantage of the dissimilarity representation is that it uses the expert knowledge in a different way. Instead of features, a dissimilarity measure has to be supplied.

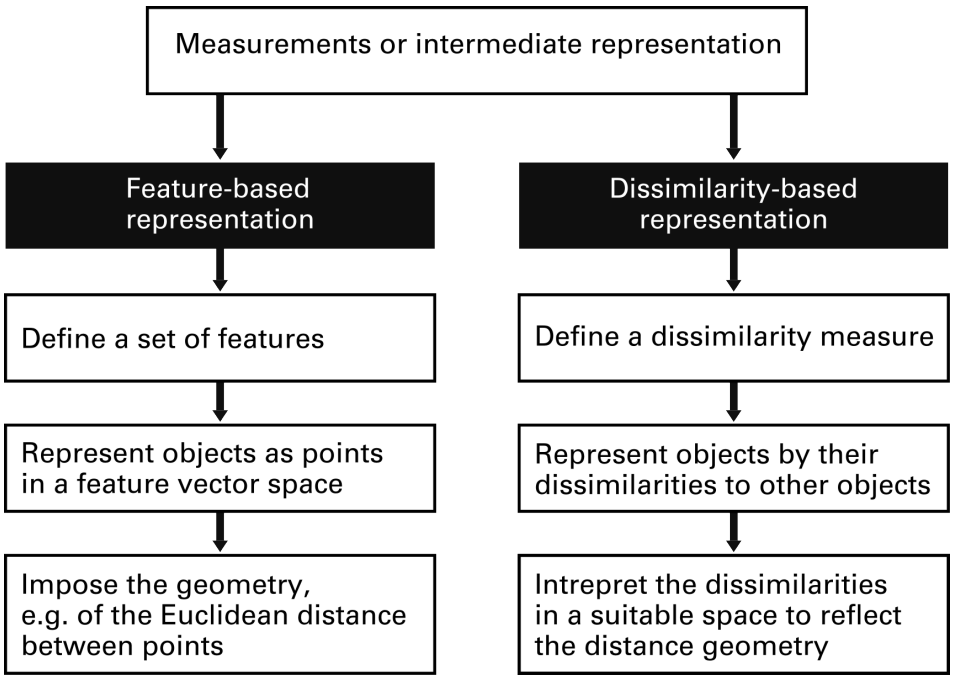


Fig. 3.1: The difference with respect to the geometry between the traditional feature-based (absolute) representations and dissimilarity-based (relative) representations [142].

In general, the suitability of a measure depends on the problem at hand and should rely on additional knowledge one has about this particular problem. Proximity underpins the description of a class as a group of objects possessing similar characteristics. This implies that the notion of proximity is more fundamental than the notion of a feature or of a class. Thereby, it should play a crucial role in class constitution. This proximity should be possibly modeled such that a class has an efficient and compact description [143].

The representation is derived from pairwise object comparisons, where the shared degree of commonality between two objects is captured by a dissimilarity value. There are many ways of comparing two objects, and hence there are many dissimilarity measures. In general, the suitability of a measure depends on the problem at hand and should rely on additional knowledge one has about this particular problem. Which to choose depends on expert knowledge or problem characteristics. If there is no clear preference for one measure over the other, a number of measures can be studied and combined. This may be beneficial, especially when different measures focus on different aspects of patterns.

There are two essential ways of constructing a vector space from a dissimilarity representation [140, 142]: Euclidean or Pseudo-Euclidean embedding and the so-called dissimilarity space.

The first one is based on an extension of linear multi-dimensional scaling, an attempt is made to embed the dissimilarity matrix in a Euclidean vector space such that the distances between the objects in this space are equal to the given dissimilarities.

Especially, an intriguing issue was the topic of embedding the given non-Euclidean dissimilarities into a vector space such that the obtained distances are sufficiently accurate in comparison to the original dissimilarities. This can only be realized error free, of course, if the original set of dissimilarities are Euclidean themselves. If this is not the case, either an approximate procedure has to be followed or the objects should be embedded into a non-Euclidean vector space.

The second way of handling the dissimilarity representation, the postulation of the dissimilarity space, raises less problems and is of high interest for practical applications. It can, without problems, be used for almost any kind of dissimilarity measure. Moreover, it has good asymptotic properties and offers the possibility of an adjustable computational complexity.

The complete dissimilarity representation yields a square matrix with the dissimilarities between all pairs of objects. Formally, a dissimilarity space is constructed as a square matrix. This matrix consists of a set of row vectors, one for each voxel. These vectors represent the voxels in a vector space constructed by the dissimilarities to each other object. Usually, such a space can be safely treated as an Euclidean space equipped with the standard inner product definition.

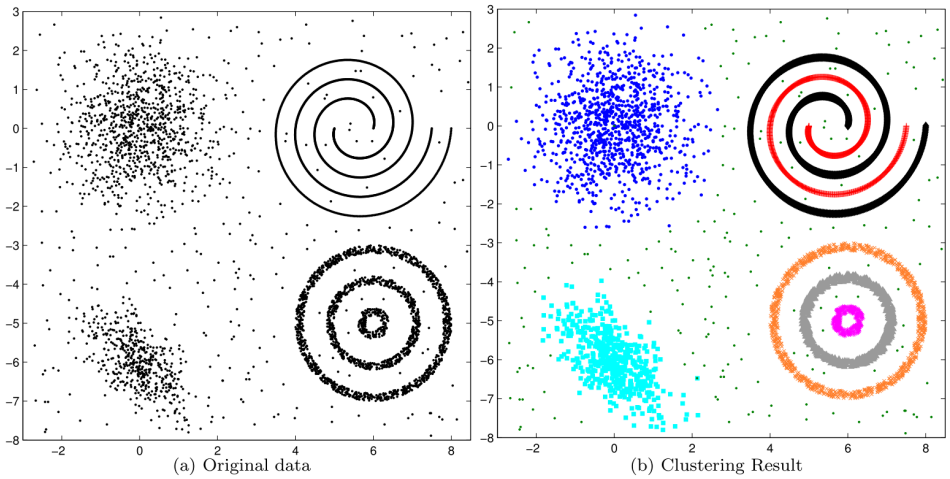
Let $X = \{x_1, \ldots, x_n\}$ be a dataset. Given a dissimilarity function, a data-dependent mapping D is defined as $D(\cdot, X) : X \to \mathbb{D}^n$ linking X to a *dissimilarity space* [142]. The complete dissimilarity representation yields a square matrix consisting of the dissimilarities between all pairs of objects. In this matrix every object is described by an n-dimensional vector of distances between the object x and all the elements of X, such that $D(x, X) = [d(x, x_1) \ldots d(x, x_n)]^T$.

A set of elements representative to the problem may be used instead of the complete dataset X. This set is called the representation or prototype set and it may be a subset of X. Using a k-element set of prototypes $R = \{r_1, r_2, \ldots, r_k\}$, the dissimilarity representation is calculated between X and R, $D(X, R)$, defined as $D(\cdot, R) : X \to \mathbb{R}^k$.

One of the advantages of this representation is that every classifier defined for feature spaces can be used in the dissimilarity space.

The dissimilarity space in fact interprets the dissimilarities as features. Their characteristics of dissimilarities are not used when a general classifier is applied. The dissimilarity space can be used for any dissimilarity representation, including ones that are negative or asymmetric [142].

The performance of this representation depends on the choices of the dissimilarity measure or features and thereby on the ability of the analyst or application expert to express his knowledge on the problem in a particular way. Thereby, the preference for one representation or the other depends on the application as well as on the expert [41].



(a) Original data (b) Clustering Result

Fig. 3.2: Diversity of clusters. The seven clusters in (a) (denoted by seven different colors in (b)) differ in shape, size, and density. Although these clusters are apparent to a data analyst, they are not easily indentificable by any clustering algorithms. Clustering at a coarse level produces four major clusters, while a finer clustering leads to a greater number [85].

3.5 Unsupervised Classification

The ability to form meaningful groups of objects is one of the most fundamental modes of intelligence. Humans perform this task with remarkable ease. In early childhood one learns to distinguish, for example, between cats and dogs or apples and oranges. However, enabling the computer to do this task of grouping automatically is a difficult and often ill-posed problem [174].

According to Jain [84], cluster analysis is the organization of a collection of patterns (usually represented as a vector of measurements, or a point in a multidimensional space) into groups based on similarity . Clustering algorithms partition data objects (patterns, entities, instances, observances, units) into a certain number of clusters (groups, subsets, or categories). However, there is no universally agreed upon and precise definition of the term cluster. Everitt et al. [48] indicate that a formal definition (of cluster) is not only difficult but may even be misplaced. Intuitively, patterns within a valid cluster are more similar to each other than they are to a pattern belonging to a different cluster.

An operational definition of clustering can be stated as follows: Given a representation of n objects, find K groups based on a measure of similarity such that the similarities between objects in the same group are high while the similarities between objects in different groups are low. The presence of noise in the data makes the detection of the clusters even more difficult. An ideal cluster can be defined as a set of points that is compact and isolated. In reality, a cluster is a subjective entity that is in the eye of the beholder and whose significance and interpretation requires domain knowledge. But, while humans are excellent cluster seekers in two and possibly three dimensions, we need automatic algorithms

for high-dimensional data. It is this challenge along with the unknown number of clusters for the given data that has resulted in thousands of clustering algorithms that have been published and that continue to appear [83, 86, 193].

It is important to understand the difference between clustering (unsupervised classification) and discriminant analysis (supervised classification) [84]. In supervised classification, we are provided with a collection of labeled (pre-classified) patterns; the problem is to label a newly encountered, yet unlabeled, pattern. Typically, the given labeled (training) patterns are used to learn the descriptions of classes which in turn are used to label a new pattern. Unlike classification, clustering does not require assumptions about category labels that tag objects with prior identifiers. The problem is to group a given collection of unlabeled patterns into meaningful clusters. In a sense, labels are associated with clusters also, but these category labels are data driven; that is, they are obtained solely from the data.

Clustering is useful in several exploratory pattern-analysis, grouping, decision-making, and machine-learning situations, including data mining, document retrieval, image segmentation, and pattern classification. However, in many such problems, there is little prior information (e.g., statistical models) available about the data, and the decision-maker must make as few assumptions about the data as possible. It is under these restrictions that clustering methodology is particularly appropriate for the exploration of interrelationships among the data points to make an assessment (perhaps preliminary) of their structure. The term clustering is diversely used in several research communities to describe methods for grouping of unlabeled data [84].

The importance of data representation in clustering cannot be understated. A good pattern representation can often yield a simple and easily understood clustering; a poor pattern representation may yield a complex clustering whose true structure is difficult or impossible to discern [83].

3.5.1 Clustering Definition

A cluster in these definitions is described in terms of internal homogeneity and external separation, i.e., data objects in the same cluster should be similar to each other, while data objects in different clusters should be dissimilar from one another. Both the similarity and the dissimilarity should be elucidated in a clear and meaningful way. Here, we give some simple mathematical descriptions of two types of clustering, known as partitional and hierarchical clustering [193].

Given a set of input patterns $X = \{x_1, \ldots, x_j, \ldots, x_N\}$, where $x_j = (x_{j1}, x_{j2}, \ldots, x_{jd}) \in \mathbb{R}^d$, where each measure x_{ji} is a feature (attribute, dimension, or variable):

Hard partitional clustering attempts to seek a K-partition of X, $C = \{C_1, \ldots, C_K\}$ $(K \leq N)$, such that

- $C_i \neq \phi,\ i = 1, \ldots, K$;
- $\cup_{i=1}^{K} C_i = X$;
- $C_i \cap C_j = \phi,\ i, j = 1, \ldots, K$ and $i \neq j$.

Hierarchical clustering algorithms produce a nested series of partitions based on a criterion for merging or splitting clusters based on similarity. They attempt to construct a tree-like, nested structure partition of $X, H = \{H_1, \ldots, H_Q\}$ $(Q \leq N)$, such that $C_i \in H_m$, $C_j \in H_l$, and $m > l$ imply $C_i \in C_j$ or $C_i \cap C_j = \phi$ for all $i, j \neq i$, $m, l = 1, \ldots, Q$.

For hard partitional clustering, each data object is exclusively associated with a single cluster. It may also be possible that an object is allowed to belong to all K clusters with a degree of membership, $u_{ij} \in [0, 1]$, which represents the membership coefficient of the j^{th} object in the i^{th} cluster as introduced in fuzzy set theory by Zadeh [205].

3.5.2 Clustering Procedure

Figure 3.3 depicts the procedure of cluster analysis with the following four basic steps:

1. **Feature selection or extraction**. In statistical pattern recognition, feature selection chooses distinguishing features from a set of candidates, while feature extraction utilizes some transformations to generate useful and novel features from the original ones. Clearly, feature extraction is potentially capable of producing features that could be of better use in uncovering the data structure. However, feature extraction may generate features that are not physically interpretable, while feature selection assures the retention of the original physical meaning of the selected features. In the literature, these two terms sometimes are used interchangeably without further identifying the difference [193]. Both feature selection and feature extraction are very important to the effectiveness of clustering applications. Elegant selection or generation of salient features can greatly decrease the storage requirement and measurement cost, simplify the subsequent design process, and facilitate the understanding of the data. Generally, ideal features should be of use in distinguishing patterns belonging to different clusters, immune to noise, and easy to obtain and interpret [40,84].
2. **Clustering algorithm design or selection**. This step usually consists of determining an appropriate proximity measure and constructing a criterion function. Intuitively, data objects are grouped into different clusters according to whether they resemble one another or not. Almost all clustering algorithms are explicitly or implicitly connected to some particular definition of proximity measure. The subjectivity of cluster analysis is thus inescapable. Clustering is ubiquitous, and a wealth of clustering algorithms has been developed to solve different problems from a wide variety of fields. However, there is no universal clustering algorithm to solve all problems. Clustering algorithms that are developed to solve a particular problem in a specialized field usually make assumptions in favor of the application of interest.
3. **Cluster validation**. Given a data set, each clustering algorithm can always produce a partition whether or not there really exists a relevant structure in the data. Moreover, different clustering approaches usually lead to different clusters of data, and even for the same algorithm, the selection of a parameter or the presentation order of input patterns may affect the final results. Therefore, effective evaluation standards and criteria are critically important

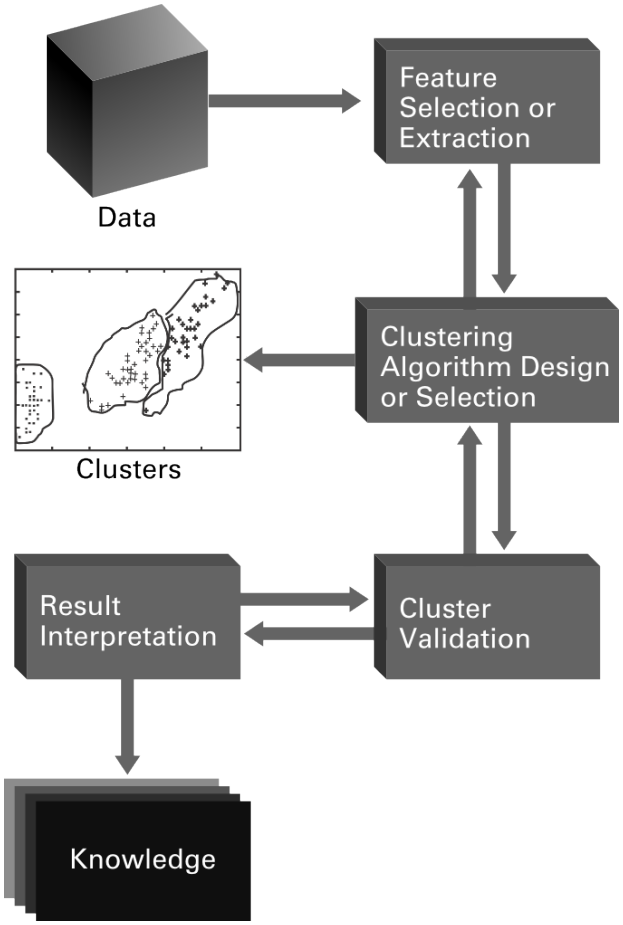


Fig. 3.3: Clustering procedure. The basic process of cluster analysis consists of four steps with a feedback pathway. These steps are closely related to each other and determine the derived clusters [193].

to provide users with a degree of confidence for the clustering results. Little in the way of *gold standards* exist in clustering except in some specific subdomains where well-known contextual information define the validation procedure and benchmarks [83]. The validation assessment should be objective and have no preferences to any algorithm. Generally, there are three categories of testing criteria: external indices, internal indices, and relative indices. External indices are based on some prespecified structure, which is the reflection of prior information on the data and is used as a standard to validate the clustering solutions. Internal tests are not dependent on external information, instead, they examine the clustering structure directly from the original data. Relative criteria emphasize the comparison of different clustering structures in order to

provide a reference to decide which one may best reveal the characteristics of the objects [86].

4. **Result interpretation**. The ultimate goal of clustering is to provide users with meaningful insights from the original data so that they can develop a clear understanding of the data and therefore effectively solve the problems encountered [193]. Experts in the relevant fields are encouraged to interpret the data partition, integrating other experimental evidence and domain information without restricting their observations and analyses to any specific clustering result. Consequently, further analyses and experiments may be required.

It is interesting to observe that the flow chart in Fig. 3.3 also includes a feedback pathway. Cluster analysis is not a one-shot process. In many circumstances, clustering requires a series of trials and repetitions. Moreover, there are no universally effective criteria to guide the selection of features and clustering schemes. Validation criteria provide some insights into the quality of clustering solutions, but even choosing an appropriate criterion is a demanding problem.

3.6 Relevant clustering algorithms

In this section we present three different clustering algorithms, chosen because of their properties and diverse origin. These algorithms are: the classic K-means algorithm, Affinity Propagation and Support Vector Clustering. Each one of these three clustering methods involve different assumptions and by consequence they are diversely biased, a desirable characteristic that is exploited when we create an ensemble of cluster solutions, explained in Chapter 6.

3.6.1 The K-means algorithm

Let $X = \{x_i\}, i = 1, \ldots, n$ be the set of n-dimensional points to be clustered into a set of K clusters $C = \{c_k, k = 1, \ldots, K\}$. The K-means algorithm finds a partition such that the squared error between the empirical mean of a cluster and the points in the cluster in minimized. Let μ_k be the mean of cluster c_k. The squared error between μ_k and the points in cluster c_k is defined as

$$J(c_k) = \sum_{x_i \in c_k} \|x_i - \mu_k\|^2 \tag{3.1}$$

The goal of K-means is to minimize the sum of the squared error over all K clusters,

$$J(C) = \sum_{k=1}^{K} \sum_{x_i \in c_k} \|x_i - \mu_k\|^2 \tag{3.2}$$

Minimizing this objective function is known to be an NP-hard problem (even for K=2). Thus K-means, which is a greedy algorithm, can only converge to a local minimum, even though recent study has shown with a large probability K-means could converge to the global optimum when clusters are well separated [85].

K-means starts with an initial partition with K clusters and assign patterns to clusters so as to reduce the squared error. Since the squared error always decreases with an increase in the number of clusters K (with $J(C) = 0$ when $K = n$), it can be minimized only for a fixed number of clusters. The main steps of K-means algorithm are as follows [86]:

1. Select an initial partition with K clusters; repeat steps 2 and 3 until cluster membership stabilizes.
2. Generate a new partition by assigning each pattern to its closest cluster center.
3. Compute new cluster centers.

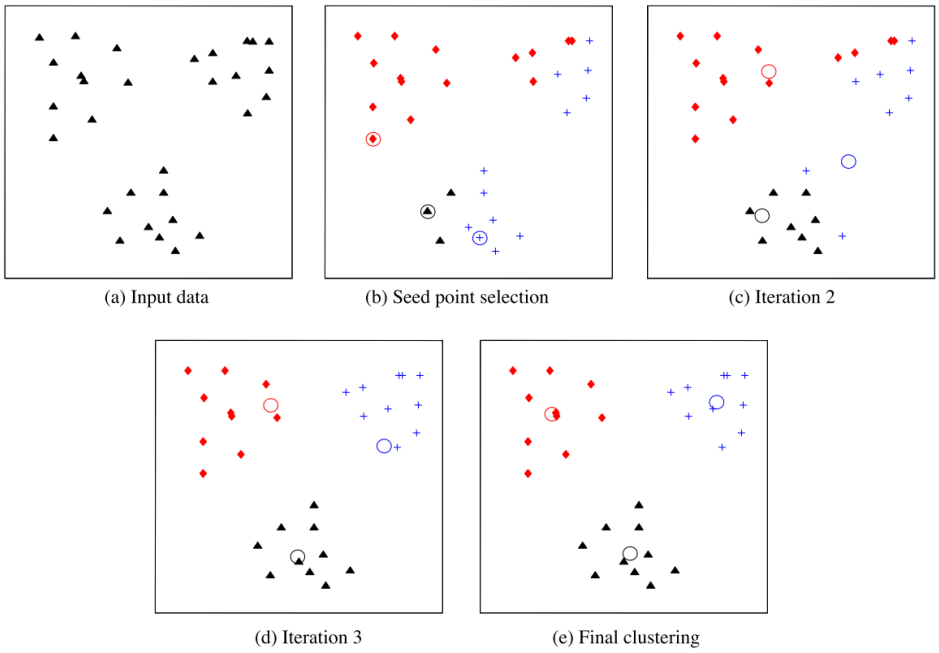


Fig. 3.4: Illustration of K-means algorithm. (a) Two-dimensional input data with three clusters; (b) three seed points selected as cluster centers and initial assignment of the data points to clusters; (c) and (d) intermediate iterations updating cluster labels and their centers; (e) final clustering obtained by K-means algorithm at convergence [83].

The K-means algorithm requires three user-specified parameters: number of clusters K, cluster initialization, and distance metric. The most critical choice is K. While no perfect mathematical criterion exists, a number of heuristics are available for choosing K. Typically, K-means is run independently for different values of K and the partition that appears the most meaningful to the domain expert is selected. Different initializations can lead to different final clustering because K-means only converges to local minima. One way to overcome the local minima is to run the K-means algorithm, for a given K, with multiple different

initial partitions and choose the partition with the smallest squared error. The basic K-means algorithm has been extended in many different ways. Some of these extensions deal with additional heuristics involving the minimum cluster size and merging and splitting clusters [40, 179]. A further description of the variants can be found in the following references [54, 83, 132, 144].

3.6.2 Clustering by Affinity Propagation

Affinity Propagation (AP) is a recent clustering methodology first proposed by Frey and Dueck in [56]. Fundamental to clustering by AP is the concept of *exemplars*. In AP an exemplar is a data point that represents a cluster. The exemplars are analogous to the centers in commnon algorithms such as k-centers in which the actual data set is used to learn a set of centers such that the sum of squared errors between data points and their nearest centers are small.

AP simultaneously considers all data points as potential exemplars. Instead of requiring that the number of clusters k be prespecified, AP requires a set of similarities (also denoted as affinities) between pairs of data points, s_{ij}, and partitions the data set into clusters such that each partition is associated with an exemplar point that best describes that cluster. In AP each data point is associated with a single exemplar. Thus, the objective of AP is to maximize the overall sum of similarities between data points and their exemplars [56]. The self similarities $s(i,i)$ are called *preferences*, and describe to what extent a point is suitable for being an exemplar. The number of identified exemplars, which in turn denotes the number of partitions or clusters, is influenced by the values of the input preferences, but also emerges from the message-passing procedure.

In order for the assignments of data points to exemplars to give sensible clustering solutions, an exemplar must never be assigned to another exemplar, an exemplar must always be assigned to itself. This is referred as the *exemplar consisteny constrain*.

At a high level, the AP algorithm can be viewed as iteratively sending messages between points. The messages are scalar values, and there are two types of messages which are sent. First, each point sends to all other points a message indicating to what degree each of the other points is suitable to be its exemplar. These messages, denoted by ρ_{ij}, are referred to as *responsibilities*.

$$\rho_{ij} = s_{ij} - \max_{k \neq j}(s_{ik} + \alpha_{ik}) \tag{3.3}$$

Then, each point sends to all other points a message indicating to what degree the point itself is suitable to serve as an exemplar. These messages, denoted by α_{ij}, are referred to as availabilities (Eq. 3.4). The messages are sent iteratively until the messages no longer change, at which point the algorithm is said to have reached a fixed point, or converged.

$$\alpha_{ij} = \begin{cases} \sum_{k \neq j} \max(0, \rho_{kj}) & i \neq j, \\ \min\left[0, \rho_{ij} + \sum_{k \notin \{i,j\}} \max(0, \rho_{kj})\right] & i \neq j \end{cases} \tag{3.4}$$

At convergence the set $\mathcal{K} = \{k|\alpha_{kk} + \rho_{kk} > 0\}$ is chosen as the set of exemplars. Each non-exemplar point i is assigned to its most similar exemplar, $k = \arg\max_{k \in \mathcal{K}} s_{ik}$.

3.6.3 Support Vector Clustering

Conceptually Support Vector Clustering (SVC), just as KPCA and other kernel based methods, maps the data points from data space to a high dimensional feature space using a Gaussian kernel. In feature space SVC looks for the smallest sphere that encloses the image of the data. This sphere is mapped back to data space, where it forms a set of disjoint contours which enclose the data points. These contours are interpreted as cluster boundaries. Points enclosed by each separate contour are associated with the same cluster. As the width parameter, σ, of the Gaussian kernel is decreased, the number of disconnected contours in data space increases, leading to an increasing number of clusters. Since the contours can be interpreted as delineating the support of the underlying probability distribution, SVC can be viewed as one identifying valleys in this probability distribution [18].

In its original formulation SVC can deal with outliers by employing a soft margin constant that allows the sphere in feature space to leave outliers out, not enclosing all points. For large values of this parameter, we can also deal with overlapping clusters.

SVC starts with a classic formulation of a data set as a support vector description [161]. For a data set $x_i \in \mathbb{R}^d$, where $i = 1, \ldots, n$, a non-linear mapping Φ is used to transform the *input space* $\mathbb{R}^d$ to a high dimensional *feature space* $\mathcal{H}$ (Eq. 6.10). Subsequently SVC finds the smallest hypersphere in $\mathcal{H}$ that encloses the projected image of the data set:

$$\|\Phi(x_i) - a\|^2 \leq R^2 \quad \forall i, \tag{3.5}$$

where a is the center of the hypersphere and R the radius. Slack variables are added to allow for soft boundaries in which some data points can be allowed to lie outside the borders.

$$\|\Phi(x_i) - a\|^2 \leq R^2 + \xi_i \quad \forall i, \tag{3.6}$$

where $\xi_j > 0$. Ben-Hur [18] solves the problem from Eq. 3.6 with the introduction of the Lagrangian and a regularization constant C in the penalty term,

$$M(R, a, \alpha_i, \xi_i) = R^2 - \sum\nolimits_i (R^2 + \xi_i - \|\Phi(x_i) - a\|^2)\alpha_i - \sum \xi_i \mu_i + C \sum \xi_i \tag{3.7}$$

where $\alpha_i \geq 0$ and $\mu_i \geq 0$ are Lagrangian multipliers and $C \sum \xi_i$ is a penalty term. The Karush-Kuhn-Tucker conditions [53] allows the problem to be rewritten as

$$\begin{aligned} \text{Maximize} \quad & M = \sum\nolimits_i \alpha_i \Phi(x_i)^2 - \sum_{i,j} \alpha_i \alpha_j \Phi(x_i)\Phi(x_j) \\ \text{subject to} \quad & 0 \leq \alpha_i \leq C, \ \sum \alpha_i = 1, \ i = 1, \ldots, n \end{aligned} \tag{3.8}$$

Following the Support Vector Machine (SVM) method [161], a kernel representation is used in which $k(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$. Using a kernel function, the Equation (3.8) can be rewritten as

$$M = \sum\nolimits_i \alpha_i k(x_i, x_i) - \sum_{i,j} \alpha_i \alpha_j k(x_i, x_j) \tag{3.9}$$

The Lagrangian multipliers α_i are obtained optimizing Equation (3.9). Only the points with non-zero α_i lie outside or on the boundary of the hypersphere. These points are called *Support Vectors* (SVs). Points with $\alpha_i = C$ have hit the upper bound for the radius and lie outside the sphere. These points are called *Bounded Support Vectors* (BSVs) and are treated as noise.

In conceptual terms, the SVs are those data points that lie closest to the hyperspheric decision surface. SVs are the most difficult points to assign in clustering and novelty detection. s such, they have a direct bearing on the optimum location of the decision surface, and they play a prominent role in the operation of this class of learning machines [202].

For every point x, the distance to its image in feature space, $\Phi(x)$, from the center, a, of the hypersphere is given by

$$R^2(x) = \|\Phi(x) - a\|^2 \tag{3.10}$$

which, using a kernel function, can be equally expressed as

$$R^2(x) = k(x, x) - 2 \sum\nolimits_i \alpha_i k(x, x_i) + \sum\nolimits_{i,j} \alpha_i \alpha_j k(x_i, x_j) \tag{3.11}$$

The radius of the hypersphere is

$$R = \{R(x_i) \mid x_i \text{ is a support vector}\} \tag{3.12}$$

As a basis for clustering the previous representation describes the data distribution in terms of support information and it provides a straightforward modeling prototype for SVC.

The main idea behind SVC is that the support vectors extracted for novelty detection can be used to construct the boundaries of clusters in a data set. From its SVM nature, the SVC is able to detect clusters with arbitrary shapes and different density distributions. Besides being able to work with high dimensional data it provides a way to deal with outliers. SVC is carried out in two main phases, namely SVM training and cluster labeling.

The SVM training phase determines the general cluster structure of the data and the boundaries that enclose the partitions. The SVC is a boundary-based clustering method. When the hypersphere is mapped back to the original data space, this method produces a set of disjoint contours that enclose the data points. These contours can be interpreted as cluster boundaries, and linkages between each pair of data items can be estimated [202]. An illustration of the Support Vector Clustering method is shown in Figure 3.5.

Recalling Equations (3.11) and (3.12), cluster boundaries can be constructed by a set of contours$\{x|R(x) = R\}$ which, when mapped back to data space, enclose

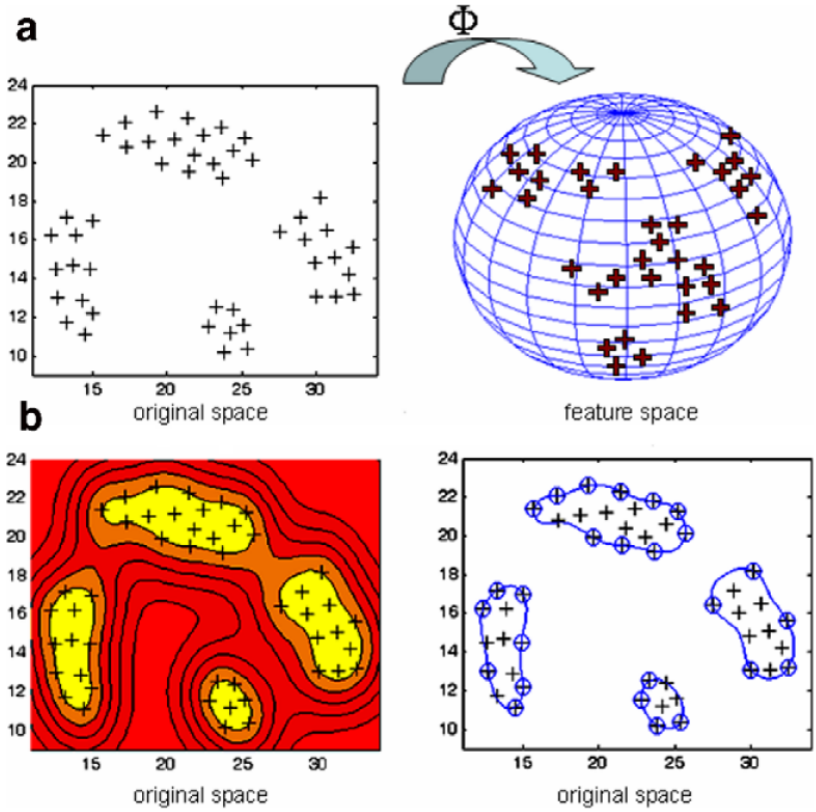


Fig. 3.5: The main steps of SVC can be appreciated in this illustration (published by Hao et al. [71]). The data is projected from data space to a feature space using kernel methods and the smallest hypersphere enclosing the data is found (a). The hypersphere is mapped back to data space where it forms a set of disjoint contours enclosing the data (b).

the data points. In this construction, SVs lie on the cluster boundaries, BSVs lie outside, and all other points lie inside the clusters.

The number of SVs and BSVs (n_{SV} and n_{BSV}, respectively) affects the overall cluster structure, which therefore can be controlled by the SVM training parameters q (Eq. 6.20) and C. As the width q of the Gaussian kernel increases, the number of SVs increases. A higher number of SVs result in a boundary whose shape is rougher and more complex, defined with more precision by the greater number of SVs. This leads to a splitting tendency in the contours that define the partitions.

On the other hand, the number of BSVs can be controlled by the parameter C, more precisely by $n_{BSV} < 1/C$. That is, if C = 1, there are no BSVs. To allow for BSVs, one should set $C < 1$ Instead of using C, it is more natural to work with the parameter $p = 1/(nC)$, which represents an upper bound for the fraction

of BSVs. The parameter q determines the scale or resolution at which the data is probed, and p decides the softness of the boundaries [18].

The SVM training phase determines the structure of the data and the cluster boundaries, however it does not differentiate between points that belong to different clusters. The cluster labeling part checks the connectivity for each pair of points based on the cut-off criteria obtained from the trained SVMs, typically the radius R of the hypersphere. To do so Ben-Hur [18] defines an adjacency matrix A_{ij} based on a geometric approach. The approach is based on the observation that every path that connects any given pair of data points that belong to different partitions must intersect the boundaries of the hypersphere in data space.

The approach is based on the observation that, for every pair of data points that belong to different partitions, any path that connects them in data space has, as a corresponding projection in the feature space, a path that intersects with the boundaries of the hypersphere. Therefore that path contains at lest a segment of points y such that $R(y) > R$.

For a pair of points x_i and x_j whose image lie in or on the sphere, the binary adjancency matrix A_{ij} is defined as:

$$A_{ij} = \begin{cases} 1, & \text{if, for all } y \text{ on the line segment connecting } x_i \text{ and } x_j,\ R(y) \leq R \\ 0, & \text{otherwise.} \end{cases} \tag{3.13}$$

Clusters are now defined as the connected components of the graph induced by A. Calculating A_{ij} for points x_i and x_j is implemented by sampling a number points on the line segment between these two points. In the original procedure by Ben-Hur BSVs are unclassified by this procedure since their feature space images lie outside the enclosing sphere. After all the points inside the hypersphere are assigned to a cluster the BSVs may remain unclassified, therefore considered as noise, or they may be assigned to the closest cluster. The time complexity of the cluster labeling phase is $\mathcal{O}(n^2 m)$, where n is the number of data points and m is the number of points sampled on the line segment joining two data points.

Yang proposed the use of proximity graphs to address improve the high computational costs of the cluster labeling phase in SVC [200, 202]. The use of proximity graphs start with the concept of the connectivity matrix M. Any particular clustering solution of a data set X can be presented in the form of an $n \times n$ cluster connectivity matrix defined by

$$M_{ij} = \begin{cases} 1, & \text{if, points } x_i \text{ and } x_j \text{ belong to the same cluster} \\ 0, & \text{otherwise.} \end{cases} \tag{3.14}$$

Clusterings expressed in this matrix model can be mapped to a subgraph of a proximity graph. In proximity graphs, vertices represent data points and edges connect pairs of points to model their proximity and adjacency. The main principle of proximity graph modeling is encoding proximity and topology between data points. With a proximity graph, points are connected by edges if they are close to each other according to some proximity measure. Points that are closer to each other are naturally more likely to be in the same cluster than distant points. Thus,

Yang argues, cluster labeling with a proximity graph strategy is a good strategy to reduce the time of testing linkages [200].

The use of proximity graphs in SVC is originally based on the boundary-based methods proposed in [46, 47].

After the data is modeled in the SVM training phase, an appropriate proximity graph model is implemented. It should reflect the data distribution and incorporates proximity and topology information. Among the choices for neighboring graphs are included Delaunay Diagrams (DD), Minimum Spanning Trees (MST) and k-Nearest Neighbors (k-NN). They can be derived by considering different aspects of proximity and topology. DD represents a "is-neighbo" relationship. The MST is based on the local closeness of data points. It is a subgraph of DD, and encodes less proximity information. k-NN is based on distance concepts.

The idea is to calculate coefficients of the adjacency matrix A_{ij} only for pairs of x_i and x_j, where x_i and x_j are linked by an edge e_{ij} in a proximity graph. While estimating the edges of a proximity graph with a cut-off criteria (i.e., R), the sampling strategy for computation of A_{ij} is performed as Ben-Hur [18]. For simplicity, all edges in the current proximity graph are called *candidate edges*. We refer to an edge e_{ij} as active edge if $A_{ij} = 1$, and as passive edge if $A_{ij} = 0$. An *active path* in the current proximity graph will be formed if every edge in the path is an active edge. A connected component is equivalent to an active path.

This strategy avoids redundant checks in a complete graph and also avoids the loss of neighborhood information as it can occur when only estimating adjacencies to support vectors. The time complexity of this labeling strategy is $\mathcal{O}(mn \log n)$. Where n is the size of data set and m is the number of sampling points on the edge.

Being of no interest, passive edges are removed from the proximity graph. Clusters correspond to connected components of edges, which are active paths. After the removal of the passive edges, the cluster assignment task lies in recognizing all active paths formed. Once active edges have been determined the connected components should be grouped. In this task is done with a classical algorithm such as Depth-First-Search (DFS). BSVs can eventually be included as a members of their respective closest cluster, or regarded as noise [202].

3.7 Cluster Ensembles

CE Definition

Although, a large number of clustering algorithms have been developed for several application areas [84, 85], the famous *no free lunch* theorem by Wolpert and Macready suggests there is no single clustering algorithm that performs best for all datasets [69, 102], i.e., unable to discover all types of cluster shapes and structures presented in data (Duda et al. 2000; Fred and Jain 2005; Xue et al.2009). It is known that the current clustering methods may suggest very different structures in the same data, which are the result of the different clustering criteria being optimized. There are no clear guidelines to choosing a clustering method for a given data set and so the risk of picking an inappropriate clustering method

is high. Choosing a single clustering algorithm for the problem at hand requires both expertise and insight, and this choice might be crucial for the success of the whole study. Selecting a clustering algorithm is more difficult than selecting a classifier [102]. One of the common difficulties is the typical lack of ground truth against which the result can be matched, therefore the benchmarks for validation remain in a subjective realm.

Often, different clusterings of the same data can be obtained either from different experimental sources or from multiple runs of non-deterministic clustering algorithms [63]. Indeed, apparent structural differences may occur within the same algorithm, given different parameters. Each clustering algorithm has its own strengths and weaknesses. For any given dataset, it is usual for different algorithms to provide distinct solutions. As a result, it is extremely difficult for users to decide a priori which algorithm would be the the most appropriate for a given set of data [79].

In the last decade the Cluster Ensemble approach has emerged as an effective solution that is able to overcome these problems. Cluster ensembles (CE) address the problem of combining multiple *base clusterings* of the same set of objects into a single consolidated clustering. CE formalizes the idea that combining different base clusterings into a single representative, or consensus, would emphasize the common organization in the different data sets and reveal the significant differences between them. The goal of CE is to find a consensus which would be representative of the given clusterings of the same data set [63]. It has been found that such a practice can improve robustness, as well as the quality of clustering results. It is widely recognized that combining multiple classification or regression models typically provides superior results compared to using a single, well-tuned model [174]. Thus, the main objective of CE is to combine different decisions of various clustering algorithms in such a way as to achieve a superior accuracy to those of individual clusterings or to be more informative in regards to the structure of the data [79]. Furthermore, CE provide a more universal solution in that various structures and shapes of clusters present in data may be discovered by the same ensemble method, and the solution is less dependent upon the chosen ensemble type. In the classic definition of CE, each base clustering refers to a grouping of the same set of objects or its transformed version using a suitable clustering algorithm. The consolidated clustering is often referred to as the *consensus solution* [59].

Multiple clusterings of the same data arise in many situations. Goder [63] mentions two classes of instances. The first class is when different attributes or features of large data sets yield different clusterings of the entities, such with the many experiments performed on gene expression data. In addition to the individual value of each experiment, combining the data across multiple experiments could potentially reveal different aspects of the genomic system and its properties. In this case one useful way of combining the data from different experiments is to aggregate their clusterings into a consensus or representative clustering, using CE, which may both increase the confidence in the common features in all the datasets and also reveal the important differences among them [63]. The second class of instances results from situations where multiple runs of the same non-deterministic clustering or data mining algorithms yield multiple clusterings of the same entities. Non-deterministic clustering algorithms, e.g. K-means, are sensitive to the choice

of the initial seed clusters; running K-means with different seeds may yield very different results. This is in fact desirable when the data is non-linearly separable, as the multiple weak clusterings could then be combined into a stronger one. To address this, it is becoming more and more common to analyze jointly the resulting clusterings from a number of K-means runs, seeded with different initial centers. One way to aggregate all those clusterings is to compute a consensus among them, which would be more robust to the initial conditions [63].

At first glance, the CE problem sounds similar to the widely prevalent use of combining multiple classifiers to solve difficult classification problems, using techniques such as bagging, boosting, and output combining [7, 59, 95]. However, combining multiple clusterings poses additional challenges. First, the number of clusters produced may differ across the different base solutions. The appropriate number of clusters in the CE consensus is also not known in advance and may depend on the scale at which the data is inspected. Moreover, cluster labels are symbolic and thus aligning cluster labels across different solutions requires solving a potentially difficult correspondence problem. Also, the original data used to yield the base solutions are not available to the consensus mechanism, which has only access to the sets of cluster labels.

Advantages of Using CE

There are many reasons for using a cluster ensemble. In fact, the potential motivations and benefits are much broader than those for combining supervised classifiers. As enumerated by Ghosh and Acharya [59], some of these reasons include:

1. **Improved quality of solution**. Just as ensemble learning has been proved to be more useful compared to single model solutions for classification and regression problems, one may expect that cluster ensembles will improve the quality of results as compared to a single clustering solution. It has been shown that using cluster ensembles leads to more accurate results on average as the ensemble approach takes into account the biases of individual solutions [69, 102].
2. **Robust clustering**. It is well known that the popular clustering algorithms often fail for certain datasets that do not match well with the modeling assumptions. A cluster ensemble approach can provide a meta clustering model that is much more robust in the sense of being able to provide good results across a very wide range of datasets. As an example, by using an ensemble that includes algorithms known to perform better on low-dimensional spaces as well as clusterers designed for high-dimensional sparse spaces, one can perform well across a wide range of data dimensionality [174].
3. **Model selection**. Cluster ensembles provide a novel approach to the model selection problem by considering the match across the base solutions to determine the final number of clusters to be obtained.
4. **Knowledge reuse**. In certain applications, domain knowledge in the form of a variety of clusterings of the objects under consideration may already exist due to past projects. A consensus solution can integrate such information, reusing it to obtain a more consolidated clustering. Strehl and Ghosh provide

several examples in [174], where such scenarios formed the main motivation for developing a consensus clustering methodology.

5. **Multi-view clustering**. Often the objects to be clustered have multiple aspects or views, and base clusterings may be built on distinct views that involve nonidentical sets of features or subsets of data points. In marketing applications, for example, customers may be segmented based on their needs, psychographic or demographic profiles, attitudes, etc. Different views can also be obtained by considering qualitatively different distance measures, an aspect that has been exploited in clustering multifaceted proteins to multiple functional groups. Consensus clustering can be effectively used to combine all such clusterings into a single consolidated partition. Strehl and Ghosh [174] illustrated empirically the utility of cluster ensembles in two orthogonal scenarios: (a) Feature distributed clustering (FDC): where different base clusterings are built by selecting different subsets of the features but utilizing all the data points. (b) Object distributed clustering (ODC): base clusterings are constructed by selecting different subsets of the data points but utilizing all the features.
   We should remember that these cases apply to the use of nonidentical features or subsets of data points coming from a single dataset. In Chapter 6 we extend the 'Multi-View' notion to create a Consensus Clustering derived from two spatially-alligned MRI volumes, both of different MRI modalities.
6. **Distributed computing**. In certain situations, data is inherently distributed and it is not possible to first collect the entire data at a central site due to privacy/ownership issues or computational, bandwidth and storage costs. An ensemble can be used in situations where each clusterer has access only to a subset of the features of each object, as well as where each clusterer has access only to a subset of the objects.

It is important to remark that all these reasons described by Ghosh [59] refer mainly to the classic data representation by feature descriptors in a vectorial space. As we present in Chapter 6, the use of dissimilarity representations allow us to formulate a different 'multi-view' approach for multi-modality MRI acquisitions, either by the use of established metrics or by the appropriate definition of new distance functions.

3.7.1 The Cluster Ensemble Problem

To formulate the Cluster Ensemble problem we have decided to use the notation presented by Iam-On et al. [79].

For a set $X = x_1, x_2, \ldots, x_N$ of N data points let $\Pi = \pi_1, \pi_2, \ldots, \pi_M$ be a set of M base clustering results, forming what is called a *cluster ensemble*. Each base clustering result is referred to as an *ensemble member*. It returns a set of clusters $\pi_i = C_1^i, C_2^i, \ldots, C_{k_i}^i$, such that $\bigcup_{j=1}^{k_i} C_j^i = X$, where k_i is the number of clusters in the i-th clustering. Each clustering is denoted by a collection of subsets of the original dataset. For each $x \in X$, $C(x)$ denotes the cluster label to which the data point x belongs. In the i-th clustering, $C(x) = j$ if $x \in C_j^i$.

The Cluster Ensemble problem is to find a new partition π^* of a data set X that summarizes the information from the cluster ensemble Π.

The general methodology or framework of cluster ensembles is presented in Fig. 3.6. In this typicalformulation, multiple base clusterings are obtained through the independent application of diverse partitioning algorithms or procedures on the dataset X. These base clusterings form the ensemble which is the result from the first main stage of the procedure. The second stage has at its center a *consensus function* which considers all the labels in the base elements of Π to produce a unified solution. A consensus function Γ maps an ensemble $\Pi = \{\pi_1, \pi_2, \ldots, \pi_M\}$ composed of M *base clusterings* to a final unified partition π^*, $\Gamma : \Pi \mapsto \pi^*$.

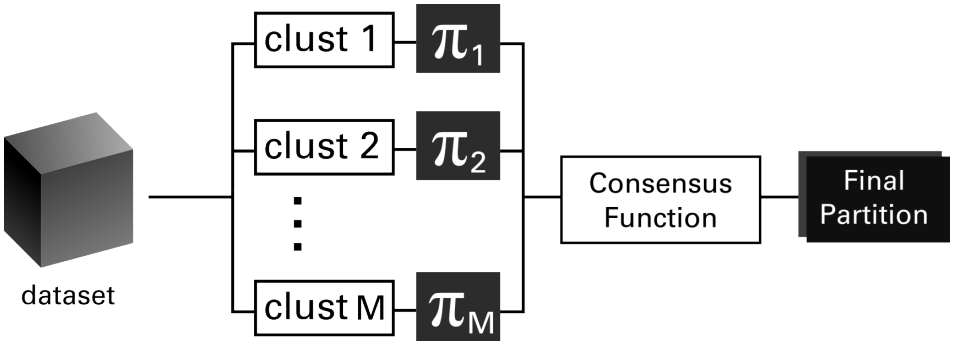


Fig. 3.6: Classic Cluster Ensemble methodology. Multiple base clusterings are performed on a dataset X to obtain diverse partitions or *base clusterings* $\pi_1, \pi_2, \ldots, \pi_M$, which together form the set known as cluster ensemble Π. These base partitions are combined methodically by a consensus function which takes into consideration all the base labelings in Π to determine the final clustering result π^*.

3.7.2 Sources of Variation for Cluster Ensembles Generation

It has been shown, by Kittler et al. [95], that ensembles are most effective when constructed from a set of clusterers whose errors are distinct [80]. This appears to be analogous to the Central Limit Theorem in which multiple samples that contain errors/randomness, when combined, reveal the true underlying distribution [80]. This observation leads to the accepted strategy of of ensuring diversity via a suitable selection of algorithms or parameters to enhance the final result of the ensemble procedure.

Diversity within an ensemble is of vital importance for its success. In such a circumstance where all ensemble members agree on how a dataset should be partitioned, a consensus solution formed from the aggregation of these base clusterings will show no improvement over any of the constituent members [69].

Several approaches have been proposed to introduce artificial instabilities in clustering algorithms, hence the diversity within a cluster ensemble. The following ensemble generation methods yield different clusterings of the same data, by exploiting different cluster models and different data partitions.

Homogeneous ensembles: Base clusterings are created using the repeated runs of a single clustering algorithm, each with a unique set of parameters. Following this, the k-means agorithm has often been employed with a random initialization of the seeding cluster centers. An ensemble of k-means is computational efficient as its time complexity is $\mathcal{O}(kNM)$, where k, N and M denote the number of clusters, the number of data points and the number of base clusterings, respectively. Other non-deterministic clustering techniques, whose results obtained from multiple runs are dissimilar, can also be used to form homogeneous ensembles.

Selection of K: For almost every clustering algorithm, the output is dependent on the initial choice of the number of clusters k. The exception to this rule are the algorithms that reach an optimal number of clusters solution according to their particular optimization criteria, such as Support Vector Clustering. However, even the result of those algorithms depends on a variety of input parameters or initializations beyond the simple dataset. To generate the ensemble diversity, base clusterings are commonly created using randomly selected values of k from a prespecified interval or even the complete interval . Although there is not a clear rule or consensus found in literature, some authors set k greater than the expected number of clusters, using also as a common rule-of-thumb $k = \sqrt{N}$ [55, 69]. This generation scheme allows a large number of clustering algorithms, both partitioning and hierarchical, to be used as base clusterings. However, k-means is still often employed for the efficiency reason [79].

Data subspacing/sampling: Cluster ensembles can also be created by applying multiple subsets of initial data to base clusterings. It is assumed that each clustering algorithm can provide different levels of performance for different partitions of a dataset. Fern and Brodley [51] used random projections of high dimensional data into subspaces to construct the base clusterings in a CE approach. The most common approach is, however, to obtain the base clusterings choosing different subsets of features [174], or a variety of data sampling schemes.

Heterogeneous ensembles: As an alternative to the homogeneous method, heterogeneous ensembles may be exploited, where the diversity is induced by allowing each base clustering to be generated using a different clustering algorithms [6, 69].

3.7.3 Consensus Methods

Having obtained the cluster ensemble, a variety of consensus functions have been developed and made available for generating the ultimate data partition. In general, consensus methods found in the literature can be categorized into: (i) pairwise similarity, (ii) graph-based and (iii) feature-based approaches, respectively.

Pairwise Similarity

This category of cluster ensemble methods is based principally on the pairwise similarity among data points. In particular, given a dataset $X = \{x_1, x_2, \ldots, x_N\}$, it first generates a cluster ensemble $\Pi = \{\pi_1, \pi_2, \ldots, \pi_M\}$ by applying M base clusterings to the dataset X.

Following that, the results of any hard clustering can be represented as a binary, symmetric, $N \times N$ similarity matrix, constructed for each ensemble member and

denoted as $S_m, m = 1 \ldots M$. Each entry in this matrix represents the relationship between two data points. If they are assigned to the same cluster, the entry will be 1, otherwise a similarity value of 0 will be assigned. Commonly this matrix is called the *coassociation matrix*. Formally, the similarity between two data points $x_i, x_j \in X$ from the m-th ensemble member can be computed as follows:

$$S_m(x_i, x_j) = \begin{cases} 1 & \text{if } C(x_i) = C(x_j), \\ 0 & \text{otherwise} \end{cases} \tag{3.15}$$

An M number of similarity matrices are merged to form an ensemble coassociation matrix, denoted henceforth as CO matrix [55]. It is also known as consensus matrix [126], similarity matrix [174], ensemble coassociation matrix [59], or agreement matrix [177]. Each element in the CO matrix represents the similarity degree between any two data points, which is a ratio of a number of ensemble members in which these data points are assigned to the same cluster to the total number of ensemble members. Formally, this similarity between $x_i, x_j \in X$ is defined as

$$CO(x_i, x_j) = \frac{1}{M} \sum_{m=1}^{M} S_m(x_i, x_j) \tag{3.16}$$

Another variant is to use a weighted version of the CO matrix [59]:

$$wCO(x_i, x_j) = \frac{1}{M} \sum_{m=1}^{M} w_m S_m(x_i, x_j) \tag{3.17}$$

where w_m specifies the weight assigned to the mth base clustering.

As Ghosh mentions [59], the CO matrix size is itself quadratic in the data size n, which thus forms a lower bound on computational complexity as well as memory requirements, inherently handicapping such a technique for applications to very large datasets. However, it is independent of the dimensionality of the data.

Since the CO matrix is a similarity matrix, any similarity-based clustering algorithm can be applied to this matrix to yield the final partition π^*. Among several existing similarity-based methods, the most well-known technique to obtain the final partition is agglomerative hierarchical clustering [80].

Graph-Based

A second type of methodology makes use of graphs representations to solve the partitioning problem in cluster ensembles. The earliest and most well known graph based ensemble methods were introduced by Strehl and Ghosh [174], with the CSPA, HGPA and MCLA algorithms, and Fern and Brodley [50], with HBGF.

Regarding the graph-based proposal by Strehl and Ghosh, the cluster-based similarity partitioning algorithm (CSPA) creates a similarity graph, where vertices represent data points and the weight of the edges is determined by the similarity scores obtained from the CO matrix (Eq. 3.16). Afterwards, a graph partitioning algorithm called METIS [92] is used to partition the similarity graph into k clusters. METIS was chosen for its scalability and because it tries to enforce

comparable sized clusters. Although this characteristic is desired in many applications, if the data is actually labeled with imbalanced classes, then it can lower the match between cluster and class labels. In a similar fashion, the hyper-graph partitioning algorithm (HGPA) poses the CE problem as a partitioning problem of a suitably defined hypergraph where hyperedges represent clusters and objects are represented by the vertices. The hypergraph partitioning algorithm HMETIS [91] was applied to partition the underlying hypergraph into k clusters. As with CSPA, employing a graph clustering algorithm adds a constraint that favors clusterings of comparable size [59].

Finally, the meta-clustering algorithm (MCLA) forms a meta-graph with a vertex for each base cluster. The edge weights of this graph are proportional to the similarity between vertices. METIS is also employed to partition this meta-level graph into k meta-clusters, where each data point has a specific association degree to each meta-cluster. The final clustering is produced by assigning each data point to the meta-cluster with which it has the highest association degree.

Feature-Based

The approach transforms the problem of cluster ensembles to clustering categorical data. Specifically, each base clustering provides a cluster label as a new feature describing each data point, which is utilized to formulate the ultimate solution. The iterative voting consensus algorithm presented by Nguyen and Caruana [133] It aims to obtain the consensus partition π^* of data points X from the categorical data induced by a cluster ensemble $\Pi = \{\pi_1, \pi_2, \ldots, \pi_M\}$.

This method utilizes the feature vector $Y = \{y_1, y_2, \ldots, y_N\}$, with N denoting the number of data points and $y_i, i = 1 \ldots N$ composed as $y_i = \{\pi_1(x_i), \ldots, \pi_M(x_i)\}$, where $\pi_g(x_i)$ represents a specific cluster label in a given clustering $\pi_g, g = 1, \ldots, M$. Each cluster in the target consensus clustering has a cluster center which is also a M-dimensional vector. Each iteration of the algorithm involves two steps: computing the cluster center of each cluster in the target consensus clustering, and a second step of reassigning each data point to its closest cluster center [133].

Probabilistic methods

In a typical mixture model approach to clustering, such as fitting the data using a mixture of Gaussians, there are k mixture components, one for each cluster. A component-specific parametric distribution is used to model the distribution of data attributed to a specific component. Such an approach can be applied to form the consensus decision if the number of consensus clusters is specified. Topchy [2] derives the consensus clustering from a solution of the maximum likelihood problem for a finite mixture model of the ensemble of partitions. Ensemble is modeled as a mixture of multivariate multinomial distributions in the space of cluster labels. These probabilistic assumptions give rise to a simple maximum log-likelihood problem that can be solved using the expectation maximization algorithm. This model also takes care of the missing labels in a natural way. Bayesian version of the multinomial mixture model described above can also be formulated, one variant was proposed by Wang et al. [188].

3.8 Validation in Unsupervised Classification

Evaluating the quality of a clustering is a nontrivial and ill-posed task [133]. In supervised learning, model performance is assessed by comparing model predictions to targets. In clustering we do not have targets and usually do not know a priori what groupings of the data are best. This hinders discerning when one clustering is better than another, or when one clustering algorithm outperforms another. In general, there are two main approaches to evaluate consensus clusterings: consensus criteria measure how the target consensus clustering is in agreement with all the base clusterings that form the ensemble, and clustering criteria measure how well the final partition obtained through the consensus clustering methodology relates to the underlying features extracted from the dataset. Referring to clustering in general, *internal* validity indices evaluate the goodness of a data partition using only quantities and features inherited from the dataset, the trie to determine if the obtained structure is intrinsically appropriate for the data [84]. They are usually employed in problems where true cluster labels are unknown. In contrast, *external* validity measures exploit *a priori* information of the true data partition, expressed by known labels of the data. This is similar to the cross-validation process in supervised classification. Given a dataset whose correct clusters are known, it is possible to assess how accurately a clustering method clusters the data relative to this correct clustering [80]. These validity criteria assess the degree of agreement between two data partitions, where one of the partitions is obtained from a clustering algorithm and the other is the known partition. Importantly, the clustering algorithms at no time has access to the correct data labels or true clustering structure; when it exists, this ground truth is only used to assess the clustering performance.

3.8.1 Internal Validity Measures

Silhouette analysis

The silhouette analysis measures how close each point in one cluster is to points in the same cluster and how far away it is to points in the neighboring clusters. This is performed by quantitatively comparing the clusters by their tightness and separation and its average width provides an evaluation of cluster validity [155].

For each element $x_i \in X$, assigned to the cluster C_k, let $a(x_i)$ be the average distance $d(x_i, C_k)$ from x_i to all the other elements within the same cluster C_k. $a(x_i)$ can be interpreted as a matching measure that quantifies how well suited x_i is to the cluster C_k, where a smaller average distance denotes a better matching. The procedure is repeated with all the clusters to which x_i is not assigned, $d(x_i, C)$, where $b(x_i) = \min_{C \neq C_k} d(x_i, C)$, that is, the lowest average dissimilarity from x_i to every cluster. The cluster with this lowest average, $b(x_i)$, is called the *neighboring cluster* of x_i. Rousseuw [155] defines the related element $s(x_i)$ as

$$s(x_i) = \frac{b(x_i) - a(x_i)}{\max\{a(x_i), b(x_i)\}} \tag{3.18}$$

Which can be rewritten as:

$$s(x_i) = \begin{cases} 1 - a(x_i)/b(x_i), & \text{if } a(x_i) < b(x_i) \\ 0, & \text{if } a(x_i) = b(x_i) \\ b(x_i)/a(x_i) - 1, & \text{if } a(x_i) > b(x_i) \end{cases} \tag{3.19}$$

where it can clearly be seen that $-1 \leq s(x_i) \leq 1$. A value of $s(x_i)$ close to 1 requires $a(x_i) \ll b(x_i)$. As $a(x_i)$ is a measure that denotes how dissimilar is x_i to the cluster it was assigned, a small value means the datapoint is well matched or close to the rest of the elements of the cluster. Furthermore, a large value of $b(x_i)$ would imply that x_i is poorly matched to its neighboring cluster. Consequently a value of $s(x_i)$ close to one means that the data point was well clustered, an inference that arises from the fact that X_i would need to be close to the other elements of the same clusters, denoting compactness, and well separated from the remaining partitions. If $s(x_i)$ is close to -1, by the same logic can be concluded that x_i is improperly clustered and by this criteria it would be better suited to the neighboring cluster, and a value of $s(x_i)$ around zero would denote a point in the border of two partitions.

The average $s(x)$ of a cluster C_k is a measure of how tightly grouped all the data in the cluster are, compactness, and how well separated they are from the neighboring partitions. Thus the average $s(x)$ of the entire dataset is a measure of how appropriately the data has been clustered [155].

Compactness

Compactness measures the average pairwise distances between points in the same cluster [133], it uses only the information inherent to the dataset. Conmpactess is defined as

$$CP(\pi^*) = \frac{1}{N} \sum_{k=1}^{K} n_k \left(\frac{\sum_{x_i, x_j \in C_k} d(x_i, x_j)}{n_k(n_k - 1)/2} \right) \tag{3.20}$$

where K denotes the number of clusters in the clustering result, n_k is the number of data points belonging to the k-th cluster, $d(x_i, x_j)$ is the distance between data points x_i and x_j, and N is the total number of data points in the dataset. Ideally, the members of each cluster should be as close to each other as possible. Thus, a lower value in the Compactness index denotes a better, more condensed, clustering result.

Davies Bouldin index

The Davies Bouldin index (DB) makes use of similarity measure R_{ij} between the clusters C_i and C_j, which is defined upon a measure of dispersion s_i of a cluster C_i and a dissimilarity measure between two clusters d_{ij}. According to Davies and Bouldin [34], R_{ij} is formulated as

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \tag{3.21}$$

where d_{ij} and s_i can be estimated by the following equations. Note that v_x denotes the center of cluster C_x and $|C_x|$ is the number of data points assigned to cluster C_x

$$d_{ij} = d(v_i, v_j) \tag{3.22}$$

$$s_i = \frac{1}{|C_i|} \sum_{\forall x \in C_i} d(x, v_i) \tag{3.23}$$

Following that, the DB index is defined as

$$DB(\pi^*) = \frac{1}{k} \sum_{i=1}^{k} R_i \tag{3.24}$$

where $R_i = \max_{j=1...k, i \neq j} R_{ij}$. The DB index measures the average of similarity between each cluster and its most similar one. As the clusters have to be compact and separated, a lower DB index indicates better goodness of a data partition.

Dunn index

This validity index is introduced by Dunn [43]. Its purpose is to identify compact and well-separated clusters. For a given number of clusters K, the definition of the Dunn index is given by .

$$Dunn(\pi^*) = \min_{i=1...K} \left(\min_{j=i+1...K} \left(\frac{d(C_i, C_j)}{\max_{k=1...K}(diam(C_k))} \right) \right) \tag{3.25}$$

where $d(C_i, C_j)$ is the distance between two clusters C_i and C_j, which can be defined as

$$d(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y) \tag{3.26}$$

In addition, $diam(C_i)$ is the diameter of a cluster C_i, which is defined as follows:

$$diam(C_i) = \max_{x, y \in C_i} d(x, y) \tag{3.27}$$

In a dataset containing compact and well-separated clusters, the distances between the clusters are expected to be large and the diameters of the clusters are expected to be small. Therefore, a large value of the Dunn index signifies compact and well-separated clusters.

3.8.2 External Validity Measures

Classification Accuracy

It measures the number of correctly classified data points of a clustering solution compared with known class labels. To compute the CA, the elements of a given cluster are given the *majority* label, which corresponds to the known cluster label

to which most of the data points in that specific obtained cluster belong. Then the accuracy of the new labels is measured by counting the number of correctly labeled data points in comparison to their known class labels, and dividing by the total number of data in the dataset. Let m_i be the number of data points with the majority cluster label in cluster i, the CA can be regarded as the ratio of the number of correctly classified data points to the total number of data points in the dataset [133]. The CA is defined as

$$CA(\pi^*, \Pi') = \frac{\sum_{i=1}^{K}(m_i)}{N} \tag{3.28}$$

where N is the total number of data in the dataset. The CA ranges from 0 to 1. A value of CA close to 1 denotes a high correspondence between the ground truth or true labels and the results obtained by the clustering procedure.

Rand Index

This validity measure takes into account the number of object pairs that exist in the same and different clusters. More formally, the RI [152] is defined as

$$RI(\pi^*, \Pi') = \frac{n_{11} + n_{00}}{n_{11} + n_{10} + n_{01} + n_{00}} \tag{3.29}$$

where n_{11} is the number of pairs of data points that are in the same clusters in both partitions π^* and Π', n_{00} denotes the number of pairs of data points that are placed in the different clusters in both π^* and Π', n_{10} is the number of pairs of data points that belong to the same cluster in π^* but are in the different clusters in Π', and n_{01} indicates the number of pairs of data points that are put in the different clusters in π^* but are in the same cluster in Π'. Intuitively, n_{11} and n_{00} can be interpreted as the quantity of agreements between two partitions, while n_{10} and n_{01} are the number of disagreements. The RI has a value between 0 and 1. It takes the value of 1 when the two clusterings are identical, and 0 when no pair of points appear either in the same cluster or in different clusters in both clusterings, *i.e.* $n_{00} = n_{11} = 0$. This happens only when one clustering consists of a single cluster while the other consists only of clusters containing single points. However this scenario is not so common and lacks practical value. In fact, it is desirable for the similarity index between two random partitions to take values close to zero, or at least a constant value. Despite of its utility, a criticism against the Rand index is that the expected value between two random partitions does not take a constant value.

Adjusted Rand Index

To correct the main criticisms of the Rand index, that is, its expected value is not zero when comparing random partitions, Hubert and Arabie [75] introduced the adjusted Rand index (ARI). by taking the generalized hypergeometric distribution as the model of randomness, *i.e.* the two partitions are picked at random subject to having the original number of classes and objects in each, found the expected value for $n_{00} + n_{11}$. They proposed the following version of the Rand index

$$ARI = \frac{Index - Expected\,Index}{Max\,Index - Expected\,Index} \tag{3.30}$$

According to notation denoting the Rand index, the adjusted Rand index between partition π^* and Π' is defined by Eq. 3.31.

$$ARI(\pi^*, \Pi') = \frac{n_{11} - \frac{(n_{11}+n_{10})(n_{11}+n_{01})}{n00}}{\frac{(n_{11}+n_{10})+(n_{11}+n_{01})}{2} - \frac{(n_{11}+n_{10})+(n_{11}+n_{01})}{n00}} \tag{3.31}$$

The higher the ARI value is, the greater the agreement becomes. The ARI is bounded above by 1 and takes on the value 0 when the index equals its expected value [75].

4

Multi-modal MRI combination

4.1 Overview

In this Chapter we make an overview of the different proposed techniques for the combination of multi-modal MRI images. The generalities of the strategies for combining and representing the MRI data for classification are discussed, followed by a general description of our proposed methods, which will be developed in subsequent chapters.

4.2 Introduction

Computer-aided prognosis (CAP) and computer-aided diagnosis (CAD) involve developing and applying computerized image analysis and multi-modal data fusion algorithms to digitized patient data (e.g. imaging, tissue, genomic) for helping physicians diagnose and predict disease outcome and patient survival. While a number of data channels, ranging from the macro to the nano-scales are now being routinely acquired for disease characterization, one of the challenges in diagnosing and predicting patient outcome and treatment response has been in our inability to quantitatively fuse these disparate, heterogeneous data sources [119].

Most researchers agree that cancer is a complex disease which we do not yet fully understand. Predictive, preventive, and personalized medicine (PPP) has the potential to transform clinical practice by decreasing morbidity due to diseases such as cancer by integrating multi-scale, multi-modal, and heterogeneous data to determine the probability of an individual contracting certain diseases and/or responding to a specific treatment regimen. There is a consensus among clinicians and researchers that a more quantitative approach, using computerized imaging techniques to better understand tumor morphology, combined with the classification of disease into more meaningful subtypes, will lead to better patient care and more effective therapeutics. With the advent of digital pathology , multifunctional imaging, the acquisition of multiple, orthogonal sources of genomic, proteomic, multi-parametric radiological, and histological information for disease characterization is becoming routine at several institutions.

Computerized image analysis and high dimensional data fusion methods will likely constitute an important piece of the prognostic toolset to enable physicians to predict which patients may be susceptible to a particular disease and also for predicting disease outcome and survival.

Tools for automatic image analysis based on multi-modal imaging can provide objective information about the tissue. These tools include supervised methods that require prior knowledge, usually given as a training set in the form of manually labeled tissues, and unsupervised algorithms which are data-driven, providing unidentified clusters that inherently differ, but whose significance must be further defined.

Automatic tools have been previously used for volumetric measurements and brain tissue segmentation in various brain pathologies including glioblastoma (GB), and for creating recurrence probability maps. However, variability in scanning protocols, acquisition parameters and patient movements, which are inherent in clinical settings, result in variable and incomplete data sets (i.e. missing values) that limit the use of such methods [111].

Recent research has shown that radiological evaluation of high-grade glial tumors may be hampered by inaccurate subjective measurement and by limiting treatment response assessment to evaluation of enhancing tissue. Automatic quantitative methods based on multi-modal data improve both efficiency and accuracy of radiologic evaluation, and their role in routine clinical procedure should be developed [111].

4.3 Strategies in fusion of imaging data

If multiple sensors or sources are used in the inference process, in principle, they could be fused at one of 3 levels in the hierarchy [119, 154];

1. Raw data-level fusion
2. Feature-level fusion
3. Decision-level fusion

Most of the methods found in literature deal with supervised classification, and from them the majority only consider information fusion in the domain of classifier outputs, referred also as the interpretation domain or decision-level (3). From the point of view of unsupervised classification, the Cluster Ensemble problem (reviewed in Section 3.7) is a methodology belonging to this category, that works directly on the labels at the decision-level and reaches a final unsupervised partition by a defined combinational methodology.

If the multiple classifiers are generated using different instantiations of their inputs, then we observe that information fusion is possible on the classifier inputs or the data domain (1). Another fusion scheme, related to the last one in the data domain is the fusion at the feature-level (2), in which descriptors are independently obtained from every source and then combined in a suitable fashion. However, these approaches are thwarted by challenges in (a) homogeneous representation of the data channels, (b) fusing the attributes to construct an integrated feature vector,

and (c) the choice of classification strategy in the space (or spaces) in which the data is represented [182].

Typically the data domain (1 and 2) is a continuous space and the interpretation domain (3) is a discrete space, represented by a set of labels [154]. There are many applications where we can choose between a combination of interpretations and a combination of data as there is not a universal consensus and just as the selection of classifiers, the efficacy of a method is always problem-dependent. An illustreative diagram with the general principle of data fusion for classification is illustrated in Fig. 4.1.

Working with a direct combination of imaging modalities, there have been several attempts to combine diverse imaging data sources and modalities with the reductionist approach of simply concatenating the individual image modality attributes at each spatial location to form a single feature vector. This resulting feature vector can be used as an input to a classifier in the same way as any other feature vector, i.e. combine $F_{MRI}(c)$ and $F_{CT}(c)$ to create $[F_{MRI}(c), F_{CT}(c)]$ for every voxel location c. This approach assumes that the co-registration problem has been effectively solved. In spite of the challenges, data fusion at the feature level aims at retrieving the interesting characteristics of the phenomenon being studied. However, when the individual modalities are heterogeneous (image and non-image based) and of different dimensions, a naive concatenation will not provide a meaningful data fusion solution. This simplistic approach also overlooks relationships and constrains that are particular to each modality. A clear example (that is treated in Chapter 6) is that of the geometric constrains of the Diffusion Tensors, which being 3×3 symmetric positive-definite matrices lie on a Riemmanian submanifold in $\mathbb{R}^6$ and whose 6 independent elements cannot be treated as a direct feature vector. A related challenge in the combination of multi-modal data is to weight the relative contributions of the different channels. While one could naively concatenate the original (or meta-space) based representations to construct a fused attribute vector, different learning strategies could be leveraged to optimally weight and then combine the individual data streams.

Several classifier ensemble or multiple classifier schemes have been previously proposed to associate and correlate data at the decision-level [1724]; a much different task compared to data integration at the raw-data or feature level.

Traditional decision fusion based approaches have focused on combining either binary decisions $Y_\alpha(c) \in \{+1, -1\}$, ranks, or probabilistic classifier outputs $P_\alpha(c)$ obtained via classification of each of the k individual data sources $F_\alpha(c), \alpha \in \{1, 2, \ldots, k\}$, via a Bayesian frameworks, fuzzy set theory, or via classical decision ensembles schemes, such as Adaboost, Support Vector Machines (SVM), or Bagging [119].

Another possible solution to overcome the representational differences is to first project the data streams into a space where the scale and dimensionality differences are removed; this space is also known as a meta-space [182]. For example, imaging and non-imaging data can be homogeneously represented in the format of eigenvectors in a PCA reduced meta-space [109, 182].

Liberman et al. [111] proposed an automatic method based on a modification of the k-Nearest-Neighbors (kNN) algorithm, applied to a multi-modal MRI data with the aim of improving accuracy in assessment of therapy response to a spe-

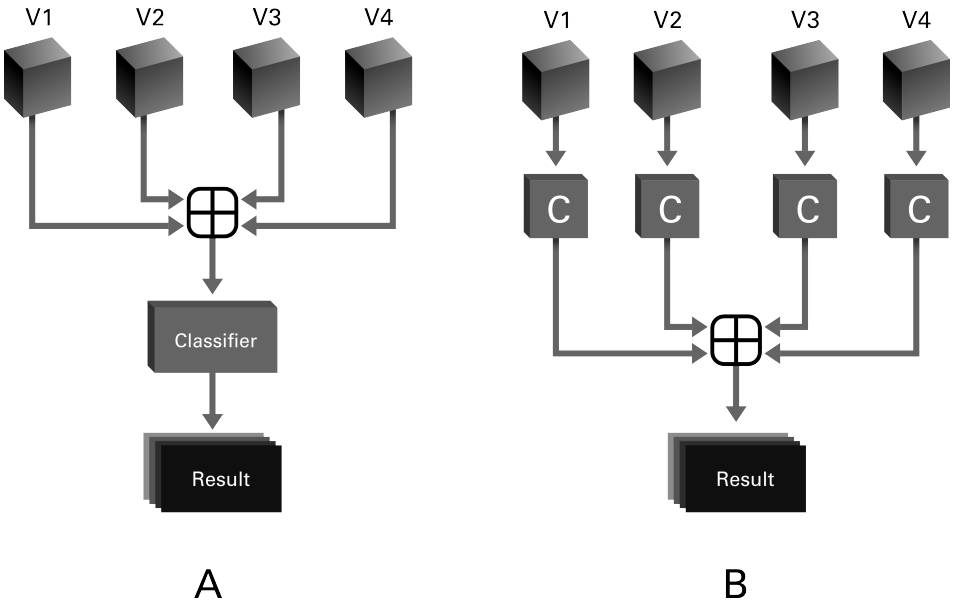


Fig. 4.1: Illustration of principles of a *combination of raw data* approach (a), against a *combination of interpretations* approach (b). In (a) the data from all sources are first combined and then classified by a single classifier. In (b) the data from each source are classified separately by the classifier C, and the outputs are combined into a final interpretation. Note that the combination operators in (a) and (b) typically work on different data types and are, therefore, usually different operators [154].

cific anti-angiogenic therapy in patients with recurrent glioblastoma. This method includes missing values in the kNN algorithm, arising from substandard acquisitions or movements, and performs voxel-based classification based solely on MR characteristics rather than spatial/morphologic properties.

Kernel-based formulations have been used in combining multiple related datasets as well as for heterogeneous data fusion. However the selection and tuning of the kernels used in multi-kernel learning (MKL) play an important role in the performance of the approach. This selection proves to be non-trivial when considering completely heterogeneous, multi-scale data. Additionally these methods typically employ the same kernel or metric, across modalities, for estimating object similarity. Thus while the Euclidean kernel might be appropriate for image intensities, it might not be appropriate for all feature spaces [119]. Lee et al. [109] proposed the Generalized Fusion Framework (GFF) for homogeneous data representation and subsequent fusion in the meta-space using dimensionality reduction techniques.

4.4 Multi-Modal MRI Integration

As it has been explained, one of the major limitation in constructing unsupervised classifying methodologies for diverse imaging data modalities is having to deal with

different data domains, which commonly differ in both scale and dimensionality. The main challenge presents itself when trying to devise an appropiate representation strategy, relevant and informative, which takes into account the differences in data type and their own constrains. Thus, a significant challenge in integrating heterogeneous imaging data has been the lack of a quantifiable knowledge representation framework to reconcile cross-modal, cross-dimensional differences in feature values. While no general theory yet exists for domain data fusion, most researchers agree that heterogeneous data needs be represented in a way that will allow for confrontation of the different channels, an important prerequisite to fusion or classification [119].

As we have seen in Chapter 3, a representation strategy that bridges the structural and statistical approaches has many benefits. We have adoptated the dissimilarity representation paradigm for our methodological proposals regarding the combination of MRI modalities for unsupervised classification.

As an overview, in the first approach (Fig. 4.2)we makes use of a specific dissimilarity function D that takes as input the different voxel-wise information of two co-registered MRI modalities to calculate a vectorial meta-space (dissimilarity space). Once this dissimilarity space is constructed, usupervised classification can be performed as in a regular feature representation. As we have already explained in Chapter 3, the definition of the appropriate dissimilarity function is crutial for a correct representation of the multi-modal data.

In our second approach, taking into consideration the information coming from a high-dimensional and geometrically complex modality such as Diffusion Tensor Imaging, we relied on the theory behind cluster ensembles to create a *multi-view* approach for the combination of MRI modalities (Fig. 4.3). This multi-view approach exploits the use of a diverse set of dissimilarity functions and already established metrics to calculate a series of vectorial spaces for each modality. The dissimilarity functions are chosen taking into account the specific information and constrains in each modality, e.g., metrics in DTI-MR that use the full tensor information and calculate the distance between two DT in their restricted Riemmanian submanifold vs. specifically tuned metrics that calculate the distance between two time-intensity curves derived from a DCE-MRI volume. Each one of these spaces reflects certain aspect of the MRI-derived information, captured through the use of distinct dissimilarity functions. Different clustering procedures are performed independently in each one of these vectorial spaces, after which a cluster ensemble is formed with the obtained labels and processed to reach an unifying solution.

In the following two chapters we describe these specific novel approaches to multi-modal MRI combination for unsupervised classification, specially for heterogeneity assessment in tumoral lesions.

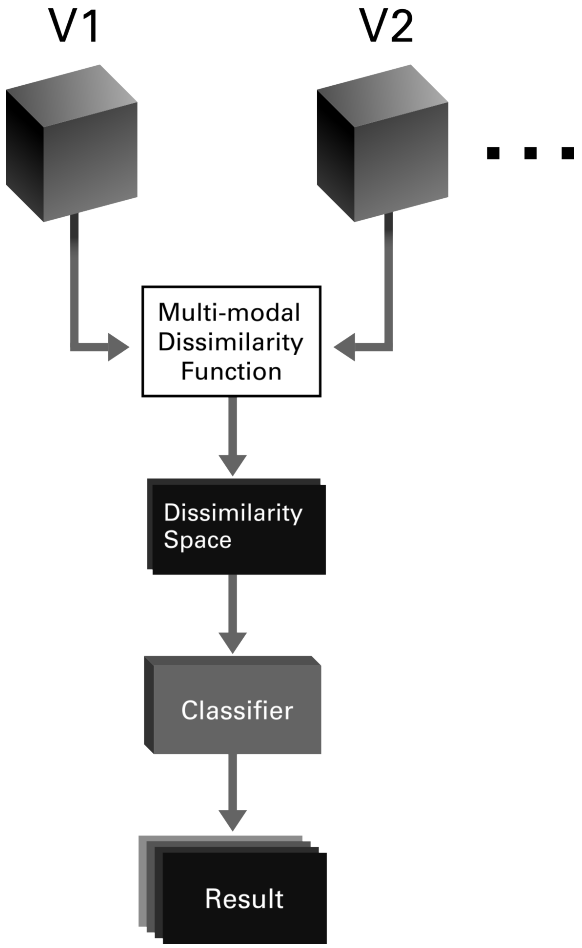


Fig. 4.2: Schematic overview of the principal components forming the first proposed approach for multi-modal MRI combination. A dissimilarity function is defined to combine the information from two co-registered multi-parametric MRI volumes in a voxel-wise manner. After the construction of a dissimilarity space diverse clustering strategies are followed.

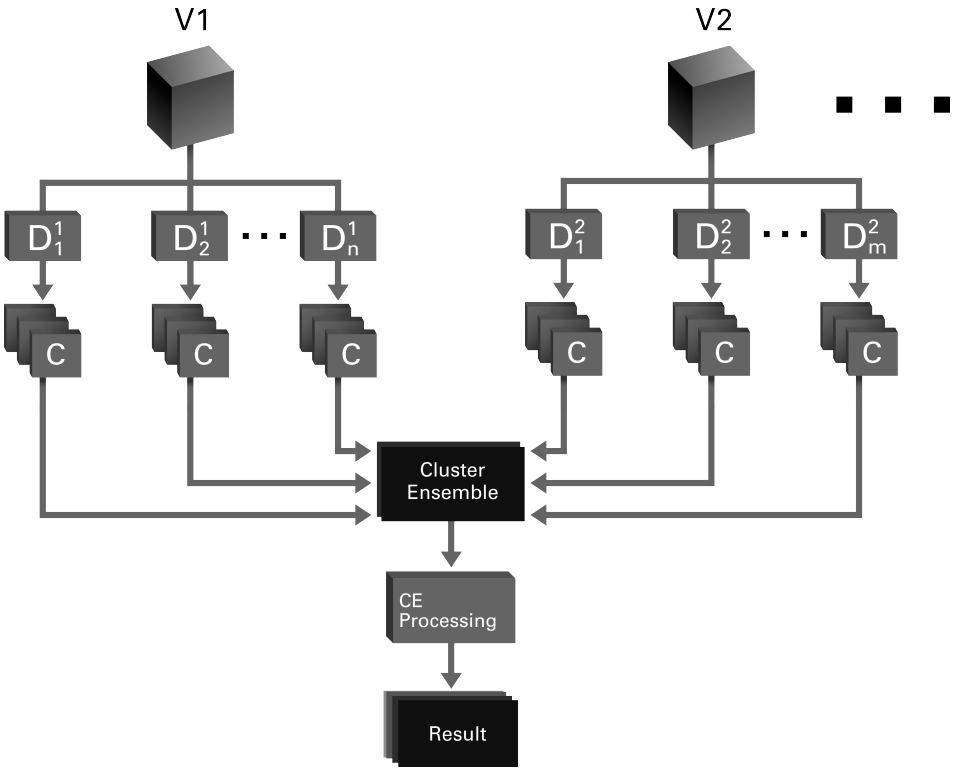


Fig. 4.3: Diagram of the *multi-view* approach for combination of multi-parametric MRI using cluster ensembles. The initial multi-modal volumes are spatially co-registered and pre-processed, for each imaging volume a suitable assortment of distance functions is chosen and used to derive a corresponding set of dissimilarity vectorial spaces (D) using voxel-wise relationships. Afterwards, a set of different clustering algorithms are applied to each dissimilarity space and an ensemble of base clusterings is formed with the obtained labels and processed to reach an unifying solution.

5

Heterogeneity assessment in breast ductal carcinoma

5.1 Overview

This chapter describes two proposed methodologies for the clinical assessment of heterogeneity inside tumoral lesions by the combination of DCE-MRI and DWI-MR acquisitions. The first one depends only on the DCE-MRI information, from which a vectorial space is constructed with the pairwise relations between voxel-wise time-intensity curves, and then clustered and compared with the respective co-registered diffusion volume. The second methodology builds on the results of the former one by actually using the information derived from both DCE-MRI and DWI-MR modalities in the construction of a multi-modal dissimilarity vectorial space which is later clustered and the results analyzed. We present results obtained with real clinical datasets, evaluated both by comparison to a typical feature-based approach as well as by their clinical significance assessed by medical experts.

5.2 Introduction

Responses to cancer treatment are increasingly differentiated not only based on tumor type, but also on genetic and histochemical biomarkers. Exemplifying the progress in this respect is breast cancer. Biopsy-derived histological biomarkers offer high biological specificity and play an important role in determining the choice of chemotherapeutic agent. As different parts of a tumor often show different histological signatures, or have evolved to different stages of tumor progression that may impact on their response to a given therapy, it is important to obtain a complete coverage of the tumor. Biopsies, however, are difficult to localize within the breast, are subject to sampling errors and can seldom be repeated. Thus, there is growing clinical interest in the possible role of imaging to describe anatomical and physiological heterogeneity of tumors [165, 201].

Magnetic resonance imaging (MRI) methods such as dynamic contrast enhanced (DCE) and diffusion weighted (DW) MRI methods are amongst those of interest as they provide non-invasive digital biomarkers with good spatial coverage, and repeatability [115]. DCE-MRI uses serial acquisition of images during

and after the injection of intravenous contrast agent, and has been shown to reflect tumor vascularity [98] [175]. DWI on the other hand,generates images that are sensitized to water displacement at the diffusion scale and can be used to calculate a quantitative index reflecting the apparent freedom of diffusion (ADC, Apparent Diffusion Coefficient). Preclinical and clinical data show that ADC reflects regional cellularity [60] [121] [87].

DCE-MRI has a high sensitivity for breast cancer detection (89-100%), while DWI has shown utility in predicting suitable therapies and monitoring response [135]. A recognized weakness of DCE and DW-MRI is their lack of specificity between tumor types as overlap between the findings of benign and malignant lesions results in variable specificity (37-86%) [135]. This is not entirely surprising given that across cancer types the common features tends to include such processes as cell proliferation, angiogenesis, and necrosis. The ability of DCE- and DW-MRI to provide a spatial depiction of these anatomical and physiological conditions within a tumor makes them natural tools for probing tumor heterogeneity. The reporting of MRI has long relied on visual assessment of several scans having different contrasts, but in relation to breast cancer, few studies have exploited this inherently multiparametric data in a unified manner [203], [88], [196]. Moreover, the most recent works mainly address the problem of comparing and retrospectively integrating the contributions from the different modalities, without exploiting the conjunct information. Nevertheless, these works have highlighted the potential of combining DCE-MRI and DWI to differentiate the core of the tumor from peritumoral tissues and normal tissues, and thus provide an indication of lesion heterogeneity [204].

In this chapter we propose the multi-modal integration of the information provided by DCE-MRI and DWI of breast cancer lesions for evaluating their heterogeneity, that is, to divide the lesion into zones that share certain similarity when using combined information coming from different imaging domains. The ultimate intention of this protocol is to to allow a more extensive, reproducible characterization of heterogeneity in tumors that have been previously identified by a clinician.

In all previous reports on breast lesion segmentation the representation of DCE curves and ADC maps has been that of features in a vector space defined by the image values [35, 57, 100, 104].

In this work a different approach is followed exploiting dissimilarity based representations (DBR) [141]. As it has been explained in Section 3.4, the concept of dissimilarity based representation consists on focusing on the contrast, or distance, between objects and on measuring it by a suitable criterion. The term *object* refers, in the present context, to the information represented by each particular voxel. This information need not be of a single type, and in this case consists of both signal intensities (i.e. the time-intensity enhancement curve for DCE-MRI) and the ADC parameter value (derived from DW-MRI). A key concept in DBR is that of a *proximity relation* between two objects, which does not need to be explicitly represented in a feature space. Objects are characterized through pairwise dissimilarities; instead of using an absolute characterization of the objects by a set of features, problem-centric knowledge is used to define a measure that estimates the dissimilarity between objects. Here, both DCE-MRI and DWI-MR

contribute to such a measure leading to a novel multi-modal approach to tissue characterization.

5.3 Dissimilarity Spaces with Time-Intensity Curves

Dissimilarity based clustering was applied to the perfusion curves in a voxel-wise manner.

From amongst the abundance of features that have been proposed as the basis for classifying lesions from breast DCE-MRI, the raw time series were used in this study.

Lavini *et al.* [104] proposed a general set of kinetic features for DCE-MRI data where time-intensity curves are classified voxel by voxel according to their shape. Kuhl *et al.* [100] showed that the use of curve shape descriptors based on the three-time-points (3TP) method could distinguish malignant from benign lesions. This method, first proposed in [35], is based on using high-spatial-resolution images while scanning the images at the selected three time points (one pre-contrast and two post-contrast time points). However, the 3TP method was criticized for not considering enhancement patterns at full time points. Subsequent pattern recognition proposals for DCE-MRI data have mainly relied on the extraction of kinetic features from the time-intensity curves or their combination with morphometric features. Gal *et al.* [57] made a survey of the available methods in the literature and proposed a set of features which includes morphological, textural and kinetic descriptors, together with variations of existing ones.

In all the methods that can be found in literature, the representation of time-intensity curves is performed by features defining a vectorial space. In this work a different approach is followed exploiting dissimilarity based representations (DBR) [141].

In this framework, following the dissimilarity representation paradigm described in Chapter 3, the first step is to construct a dissimilarity matrix. As it has been explained in Section 3.4, this matrix consists of a set of row vectors, one for each object. These vectors represent the objects in a vector space constructed by the dissimilarities to each other object.

Let $X = \{x_1, \ldots, x_n\}$ be a set of voxel based perfusion curves. Given a dissimilarity function, a data-dependent mapping D is defined as $D(\cdot, X) : X \to \mathbb{D}^n$ linking X to a dissimilarity vectorial space [142].

When using all the elements in the set X the complete dissimilarity representation yields a square matrix consisting of the dissimilarities between all pairs of objects in X, such that every object is described by an n-dimensional dissimilarity vector $D(x, X) = [d(x, x_1) \ldots d(x, x_n)]^T$.

To calculate the pairwise proximity between DCE-MRI perfusion curves we have defined a distance function D_{DCE} which is based on the adaptive dissimilarity index first proposed in [29].

There are two main approaches to quantifiably compare two time-series: the first one makes use of the distances between the absolute values of their elements, such as the classic Euclidean distance or its generalization, the Mikowski distance.

The second main approach focuses on the similarity of the time-series behavior along time, computed in many applications with the Pearson correlation coefficient. Figure 5.1 illustrate the concepts related to both proximities between time-intensity curves.

Unlike conventional time-series distance functions which are value-based, that is, focus only on the closeness of the values observed at corresponding points in time, ignoring the interdependence relationship between elements that characterize the time-series behavior, the proposed distance function takes into account the proximity with respect to values as well as the proximity with respect to their behavior, computed with the temporal correlation.

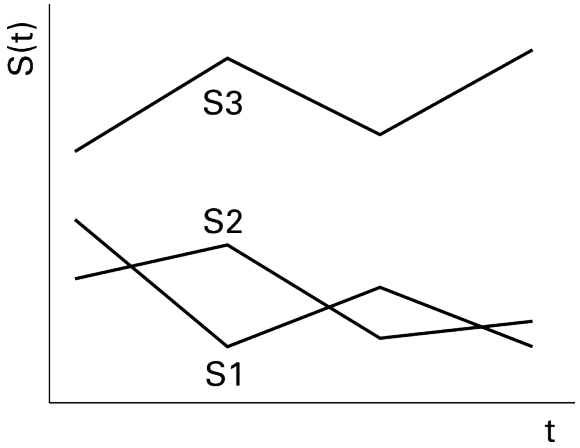


Fig. 5.1: A graphic illustrating the different concepts of simmilarity between time-intensity curves. Curves $S1$ and $S2$ have a high degree of similarity with respect to their absolute intensity values, whereas curves $S2$ and $S3$ share a high similarity with respect to their behavior along time.

For two voxel-derived perfusion curves $S_1 = (u_1, \ldots, u_p)$ and $S_2 = (v_1, \ldots, v_p)$, closeness with respect to behavior is defined as the combination of their monotonicity, that is, if both curves increase or decrease simultaneously, and the closeness of their growth rate over a determined period [29] . Both criteria are quantified by the temporal correlation, present in the first term of the distance function D_{DCE}, (Eq. 5.1). The complete distance function D_{DCE} for DCE-MRI derived perfusion curves is defined as follows:

$$D_{DCE}(S_1, S_2) = \frac{2}{1 + exp(\textsc{Cort}(S_1, S_2))} dH(S_1, S_2) \tag{5.1}$$

where $S_1 = (u_1, \ldots, u_p)$ and $S_2 = (v_1, \ldots, v_p)$ are two voxel-derived perfusion curves sampled at time instants $(t_1, \ldots, t_p)$ [29, 38]. Cort is the temporal correlation (Eq. 5.2) and dH is the Hausdorff distance, defined in Eq. 5.3, which is used to measure the value-based distance between the pair of voxel-wise perfusion curves.

$$\textsc{Cort}(S_1, S_2) = \frac{\sum_{i=1}^{p-1}(u_{(i+1)} - u_i)(v_{(i+1)} - v_i)}{\sqrt{\sum_{i=1}^{p-1}(u_{(i+1)} - u_i)^2}\sqrt{\sum_{i=1}^{p-1}(v_{(i+1)} - v_i)^2}} \tag{5.2}$$

$$dH(S_1, S_2) = \max\{\max_{u \in S_1} \min_{v \in S_2} \|u - v\|, \max_{v \in S_2} \min_{u \in S_1} \|v - u\|\} \tag{5.3}$$

Our decision to use the Hausdorff distance to compute the distance with respect to the absolute values between two time-intensity curves instead of using other options such as the widely know L^2-norm obeys a comparison and analysis of its properties. The Hausdorff distance and its modified varieties have proven to be effective in the field of figure and curve template matching, specifically dealing with a central problem in pattern recognition and computer vision, which is to determine the extent to which one shape differs from another [17, 39, 77].

Following Eq. 5.3, for two generic curves or shapes A and B, the Hausdorff distance can also be expressed as the maximum of $h(A, B)$ and $h(A, B)$, where the function $h(A, B)$ is called the directed Hausdorff distance from A to B, and is expressed as

$$h(A, B) = \max_{a \in A} \min_{b \in B} \|a - b\| \tag{5.4}$$

thus, being the maximum of both directed distances, dH measures the degree of mismatch by measuring the distance of the point of A that is farthest from any point of B and, conversely, the point of B that is farthest from any point of A. Intuitively, if the Hausdorff distance is d, every point of A must be within a distance d of some point of B and vice versa. Furthermore, dH fulfills the properties of a metric over the set of all closed, bounded sets [77].

5.4 First approach: DCE-dependent methodology

In the current study, we propose the multi-modal integration of the information provided by DCE-MRI and DWI of the breast cancer lesions. First, dissimilarity-based clustering is performed on selected DCE-MRI images to identify the different types of enhancement patterns inside the tumor. Then, the resulting regions are mapped onto the corresponding ADC images, which are spatially registered through the use of a multi-resolution elastic registration protocol. Statistical analysis revealed that the PDFs of the subregions corresponding to different clusters are statistically independent, which indicates the self-consistency of the approach and enables the integration of the information gathered from the two modalities for a robust tissue classification.

The DCE-MRI data are first visually inspected to identify the images where the lesion can be detected.

A region of interest (ROI) is delineated around the lesion, leaving space for the segmentation of relevant surrounding tissue. Clustering is performed on a dissimilarity space constructed with the voxel-wise relationships between time-intensity perfusion curves.

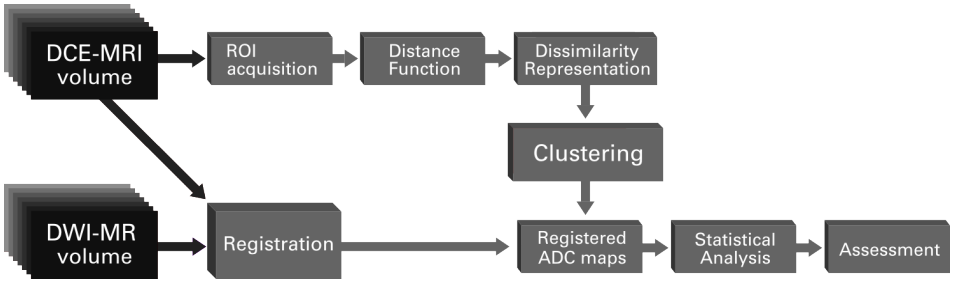


Fig. 5.2: First methodology for analysis and integration of Perfusion/Diffusion MRI information. In this processing methodology the clustering is performed exclusively on the time-intensity curves drived from the DCE-MRi volumes and the obtained regions are projected on the spatially registered ADC maps for analysis.

A multi-modal registration protocol is implemented to spatially allign DWI and DCE-MRI data, allowing a precise spatial mapping of the DCE clustered regions to the DWI. Statistical analysis is performed on the DWI-derived ADC maps to test the consistency of the information provided by this modality with that extracted from the DCE-MRI, which can thus be profitably integrated for an accurate tissue characterization.

5.5 Implementation on Clinical Data

5.5.1 Clinical MRI Data

Data were acquired from 17 patients (age 50±13.2). All the patients were affected by primary ductal carcinoma, 15 having infiltrating and 2 lobular tumors.

DWI was acquired with a single-shot spin-echo (SE) echo planar imaging (EPI) sequence in three orthogonal diffusion encoding directions (x, y and z) using 4 b values (0, 250, 500 and 1000 s/mm^2) with parallel imaging (acceleration factor 2). Subjects were breathing freely, with no gating applied. The dataset consisted of 30 transverse slices (slice thickness 5 mm, no slice gap) and TR/TE 4800/71 ms, matrix 90×150 over the field of view (FOV) 184.5×307.5 mm.

DCE-MRI was performed using a T1-weighted 3D FLASH sequence (TR/TE 7.4/4.7 ms) with a flip angle of 25 ° and an acquisition matrix of $384 \times 384 \times 128$ and field of view (FOV) $340 \times 340 \times 166$ mm. Each 120-slice set was collected in 90 s at 8 time points for approximately 12 min of scanning. A catheter placed within an antecubital vein delivered 0.1 $mmol/kg$ of the contrast agent, gadopentetate dimeglumine, (Magnevist, Wayne, NJ, USA) over 20 s (followed by saline flush) after the acquisition of one baseline dynamic scan.

5.5.2 Multi-Modal Registration

In order to perform voxel-wise dissimilarity based clustering that incorporates both DCE-MRI and DWI data, it is necessary to first spatially align the two datasets.

Given a fixed $I_F(x)$ and a moving image $I_M(x)$ of dimension d and defined on their own spatial domain: $\Omega_F \subset \mathbb{R}^d$ and $\Omega_M \subset \mathbb{R}^d$, the registration problem is the process of finding the optimal transformation $T(x)$ that brings the moving image $I_M(x)$ into spatial alignment with the fixed image $I_F(x)$. The transformation is defined as a mapping from the fixed image to the moving image, $T : \Omega_F \subset \mathbb{R}^d \rightarrow \Omega_M \subset \mathbb{R}^d$, where the alignment is measured according to a reference metric $\mathcal{S}$.

A diagram of the classic general registration methodology and components is shown in Fig. 5.3.

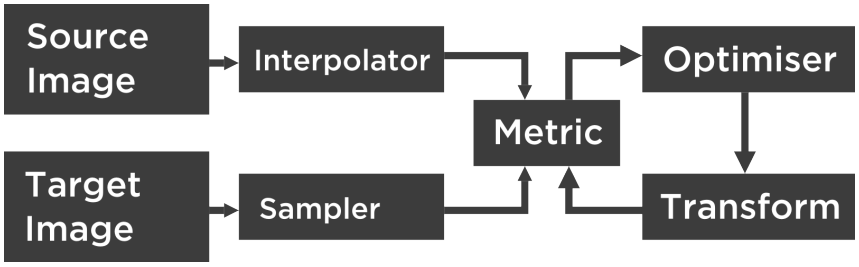


Fig. 5.3: Diagram with the classic components of a registration pipeline.

The problem of registering between DCE-MRI and DWI becomes an increasingly difficult task in a highly compressible and elastic tissues like the breast, with its inhomogeneous anisotropic soft tissue, inherent non-rigid behavior and lack of solid landmarks to guide the registration as fixed references. A standard registration protocol was used. Due to the highly distinct contrast and intensity characteristics of the two modalities, as well as the low resolution of the DWI volumes, the registration process was divided into two steps, each following a standard multi-resolution strategy.

As a registration strategy, the multiresolution methodology is started with fewer degrees of freedom for the transformation model. Specifically, the strategy involves a rigid and affine transformation before the nonrigid registration. In the first step, rigid and affine transformations were performed successively in order to align and match the features of the fixed (DCE-MRI) and moving (DWI) images. Part of this step is the iterative refinement of the segmentation using a multi-level scale-space. It refers to a methodology for handling image structures at different scales, in which the image is represented as a set of smoothed images, the scale-space representation, parametrized by the size of the smoothing kernel used for suppressing fine-scale structures. We use a 5 level Gaussian scale space, allowing at each step the refinement of the segmentation as both the fixed and moving images are iteratively matched by their more relevant features as they go from higher to lower Gaussian smoothing.

In the second step a multi-resolution cubic B-spline transformation with a regularization penalty was performed to elastically refine the alignment. B-splines are used as a a parametrisation :

$$T_\mu(x) = x + \sum_{x_k \in \mathcal{N}_x} p_k \beta^3 \left(\frac{x - x_k}{\sigma} \right) \tag{5.5}$$

with x_k the control points, $\beta^3(x)$ the cubic multidimensional B-spline polynomial, p_k the B-spline coefficient vectors or control point displacements, σ the B-spline control point spacing, and $\mathcal{N}_x$ the set of all control points within the compact support of the B-spline at x [156]. The control points x_k are defined on a regular grid, overlayed on the fixed image.

Lesion specific masks based on regions delineated by clinical experts were used in order to assign a greater weight to the voxels in the lesion area [96].

Normalized mutual information (NMI) was used as registration metric. NMI assumes a relation between the probability distributions of the intensities of the fixed and the moving image and is well suited for multi-modal registration. NMI is given by

$$NMI = \frac{(H(I_F) + H(I_M))}{H(I_F, I_M)} \tag{5.6}$$

where H_F and H_M denotes the entropy of the fixed and moving images respectively, and $H(I_F, I_M)$ represents the joint entropy, which in the case of images is a measure of mutual dispersion [147].

In order to regularize the deformation, we used a bending energy penalty which is based on the spatial derivatives of the transformation [156]. The methodology used for registration was implemented in Elastix [96] and all the steps have been widely validated in literature [68] [156].

For the implemented protocol the DCE-MRI was set as the fixed volume and the DWI as the moving one. The DCE-MRI volume corresponding to the time point where the best contrast could be detected for both the lesion and the subregions was chosen, i.e. the one acquired two minutes after the contrast medium injection, following [35]. From the DWI data, the non-diffusion weighted volume, the so called $b0$, was used as the moving one since it is less subjected to gradient-related artifacts and it is also the one providing the same amount of information on all the structures under investigation.

The registration protocol was applied to the $b0$ images from the DWI dataset and their transformation to the DCE-MRI space validated for each subject through visual assessment by a clinical expert. The resulting transformation was applied to the remaining b-values and the ADC was estimated on the transformed DWI images.

5.5.3 Assessment

In each of the patients, a ROI was delineated by an expert around the lesion in the motion-corrected DCE-MRI volumes. The time-intensity curves from the voxels inside the ROI were treated as independent objects on a voxel by voxel basis.

Using D_{DCE} from Eq. 5.1, a dissimilarity matrix was derived on a slice-wise basis for the pairwise dissimilarities of the elements from the corresponding ROI. In such a space, each element was represented by a row vector whose dimensionality

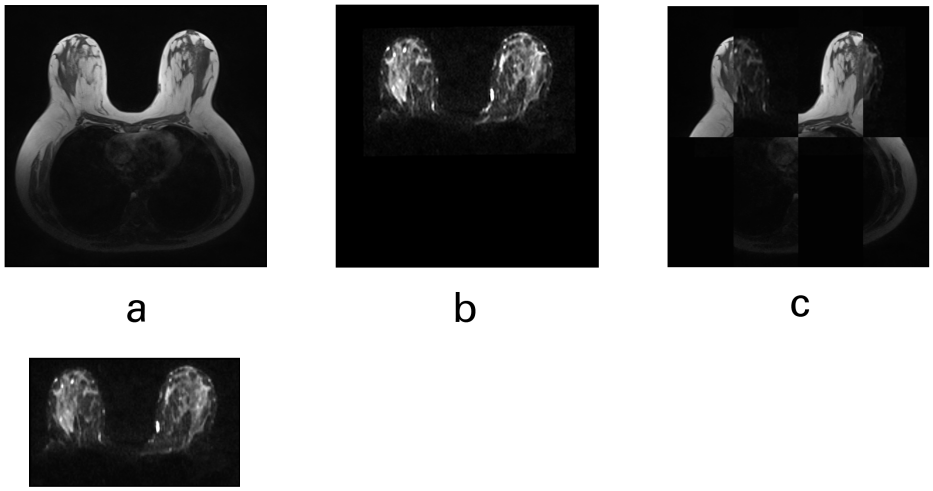


Fig. 5.4: A sample of the volumes involved in the multi-modal registration methodology. The original DCE and DWI volumes (a), the moving DWI volume after the rigid registration stage (b), and the final registration outcome compared with the fixed image in a checked pattern visualization (c).

was defined by the cardinality of the ROI. For an element x, belonging to the n-element ROI set X, the vectorial representation within the dissimilarity space is given by

$$D_{DCE}(x, X) = [D_{DCE}(x, x_1), D_{DCE}(x, x_2), \dots D_{DCE}(x, x_n)]^T \tag{5.7}$$

Once the dissimilarity space was constructed, the K-means algorithm [194] was used to group the voxels with similar perfusion patterns in the ROI into clusters. The initial centroids were calculated automatically with a preliminary clustering stage with a random 10% sample. K-means minimizes the sum over all clusters of the within-cluster sums of point-to-cluster-centroid distances using, in this case, the squared Euclidean distance. The K number of clusters was heuristically set to 5 taking into account the expected perfusion zones of the lesion and the surrounding tissue.

The regions resulting from dissimilarity based clustering were rendered as semi-transparent colored maps overlapping on the morphological images for each slice. The resulting clusters were analyzed by the radiologists of our team who confirmed and validated the segmentation of both the central and surrounding tumoral regions.

Once the registration protocol was applied to the $b0$ images and validated for each subject through visual inspection, the resulting transformation was applied to the remaining b-values in order to calculate the transformed ADC. The clusters obtained on the DCE were projected onto the spatially registered ADC maps to perform the statistical analysis. Normality tests (Jarque-Bera) revealed that the ADC value for the different clusters analyzed was not normally distributed. Accordingly, a non parametric test (Wilcoxon-signed-rank test) was used ($p = 0.05$)

to evaluate whether the tumors subregions clustered on the DCE corresponded to regions in the DWI whose PDFs were statistically different. In particular, separate analysis for single slices (intra-slice) and the tumor as a whole (inter-slice) were performed. In this we found that the distributions of the ADC values in the DCE-MRI defined regions were statistically different, in each one of the two conditions, in 15 out of 17 patients.

This results show that subregions corresponding to different clusters in the DCE volume hold statistically different ADC characteristics, supporting the self-consistency of the method and allowing for the integration of the information obtained from the two modalities. In the following Section these results will justify the extension of the method into an actual integration of the raw information from both modalities through the use of a common dissimilarity space.

5.6 Integration of DCE ad DWI for Heterogeneity Assessment

Following on the first methodology described in 5.4, this second work extends the idea of dissimilarity spaces to an actual integration of the raw information of both DCE-MRI and DWI-MR modalities into a single unfied vectorial space in which it is feasible to follow a unique clustering procedure.

A diagram of the pipeline can be seen in Fig. 5.5. As in the previous methodology (Fig. 5.2), the DCE-MRI data are first visually inspected to identify a time-point where the lesion has the higher contrast with respect to the surrounding tissue. Multi-modal registration is carried out between DW-MRI and DCE-MRI images, allowing a spatial mapping of both volumes. The informatioon of both DCE and DWI modalities is integrated into a single vectorial space using a joint dissimilarity function, after which clustering is performed in this space. Statistical analysis, consisting on non-parametric tests, were applied on the ADC distributions defined by the obtained clusters. An assessment of the results was carried out by clinical experts and an evaluation of the tightness and separation of the clusters is also performed.

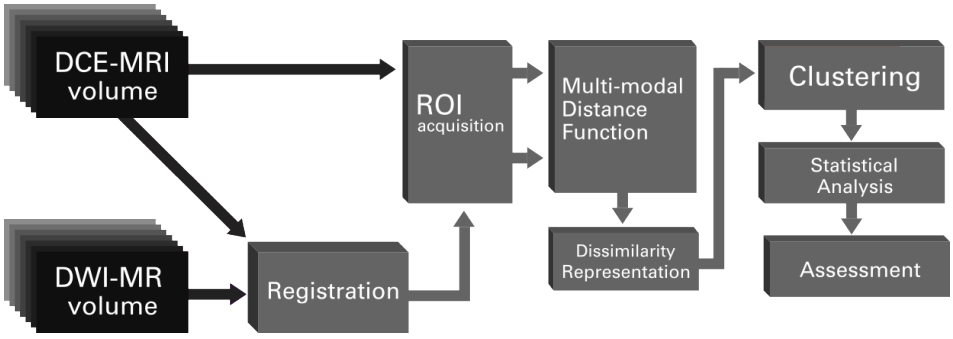


Fig. 5.5: Perfusion/Diffusion analysis and integration pipeline

5.7 Multi-modal Dissimilarity Spaces

The creation of a dissimilarity space entirely dependent of the DCE information was explained in section 5.4, here we extended the concepts behind the DCE-based dissimilarity function D_{DCE} (Eq. 5.1).

The integration of the diffusion information into the dissimilarity function is accomplished through the addition of an ADC dependent term D_{ADC} (Eq. 5.8). This term is defined as a sigmoid function which makes use of the normalized difference between the ADCs (ADC_{S1} and ADC_{S2}) of the two voxels under consideration, which ranges from 0 to 1.

$$D_{ADC}(S_1, S_2) = \frac{1}{1 + exp\left(-k_{ADC}\left(\left\|\frac{ADC_{S1} - ADC_{S2}}{\max\{ADC_{ROI}\}}\right\| - 0.5\right)\right)} \tag{5.8}$$

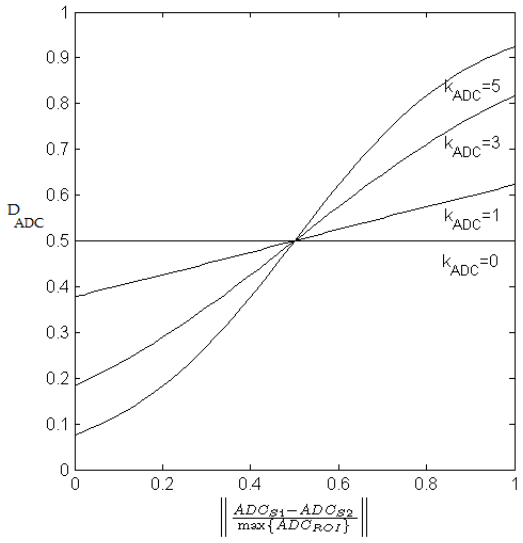


Fig. 5.6: Effects of varying the tuning parameter k_{ADC} from Eq. 5.8.

The tuning parameter k_{ADC} weights the contribution of D_{ADC} to the complete dissimilarity measure D by modulating the shape of the sigmoid function. When the value of the normalized difference between ADCs is low, denoting similar ADC values between voxels, the dissimilarity function D_{ADC} approaches zero. On the contrary, when the value of the normalized difference between ADCs is high, denoting a large dissimilarity between ADC values between voxels, D_{ADC} approaches one, making the overall dissimilarity measure approach the value of D_{DCE}. The impact of the different values of k_{ADC} is illustrated in Fig. 5.6.

The complete dissimilarity function D is then the product of D_{ADC} and D_{DCE} (Eq.5.9).

$$D = D_{ADC} \cdot D_{DCE} \tag{5.9}$$

Using the complete form D is expressed as:

$$D(S_1, S_2) = \frac{2 \cdot dH(S_1, S_2)}{1 + exp(\textsc{Cort}(S_1, S_2))} \times \frac{1}{1 + exp\left(-k_{ADC}\left(\left\|\frac{ADC_{S1} - ADC_{S2}}{\max\{ADC_{ROI}\}}\right\| - 0.5\right)\right)} \tag{5.10}$$

This global measure enables the monitoring of the performance as a function of the relative weight given to the ADC, as well as of different values of k_{ADC}.

5.8 Tests with Clinical Data

5.8.1 Clinical MRI Data

Data were acquired from 21 patients (age 50±13.8 years). All the patients had been diagnosed to have primary ductal carcinoma.

The MRI clinical protocol was the same as in Section 5.5.1.

DWI was acquired with a single-shot spin-echo (SE) echo planar imaging (EPI) sequence in three orthogonal diffusion encoding directions using 4 b values (0, 250, 500 and 1000 s/mm^2) with parallel imaging. Subjects were breathing freely, with no gating applied. The dataset consisted of 30 transverse slices (slice thickness 5 mm, no slice gap) and TR/TE 4800/71 ms, matrix 90×150 over the field of view (FOV) 184.5×307.5 mm.

DCE-MRI was performed using a 3D T1-weighted FLASH sequence (TR/TE 7.4/4.7 ms) with a flip angle of 25 ° and an acquisition matrix of $384 \times 384 \times 128$ and field of view (FOV) $340 \times 340 \times 166$ mm. Each 120-slice set was collected in 90 s at 8 time points for approximately 12 min of scanning. A catheter placed within an antecubital vein delivered 0.1 $mmol/kg$ of the contrast agent, gadopentetate dimeglumine, (Magnevist, Wayne, NJ, USA) over 20 s (followed by saline flush) after the acquisition of one baseline dynamic scan. The DCE-MRI timeseries was motion corrected using the scanner manufacturer's in-line procedure.

5.8.2 Performance Assessment

In each of the patients, a ROI was delineated by an expert around the lesion in the motion-corrected DCE-MRI volumes. Since unsupervised classification is sensitive to the general structure and distribution of the data, the ROI was drawn just exceeding the area of the enhancing lesion, allowing for a clear delineation of the heterogeneity of the lesion inside the ROI. The time-intensity curves normalized to the baseline at $t = 0$ and the corresponding ADC values from the voxels inside the ROI were treated as independent objects on a voxel by voxel basis. Using D from Eq. 5.10, a dissimilarity matrix was derived on a slice-wise basis from the pairwise dissimilarities of the elements in the corresponding ROI. In such a space, each element was represented by a row vector whose dimensionality was defined by the cardinality of the ROI.

Once the dissimilarity space was constructed, the K-means algorithm [194] was used to group the voxels in the ROI into clusters. The initial centroids were calculated automatically following a preliminary clustering step with a random 10% subsample, as a strategy to improve the algorithm initialization avoiding a misplacement of the initial seeds. K-means minimizes the sum over all clusters of the within-cluster sums of point-to-cluster-centroid distances using, in this case, the squared Euclidean distance.

For selecting the K number of clusters the standard clinical assessment protocol has been taken into consideration. It considers only three classes (persistent, plateau and wash-out). An additional has been included for the surrounding tissue considering that the ROI exceeds the estimated limits of the enhancing lesion.

In order to perform a comparison with established methods the clustering procedure was also performed following a morphologic feature-based approach. This method relies on descriptors derived from the voxel-wise time-intensity curves, comprising mainly specific characteristics of the shape of such curve.

The features extracted from the DCE-MRI voxel-wise time-intensity curves are: baseline, maximum signal difference, time to peak, area under curve, maximum enhancement, wash-in rate, maximum slope of increase, wash-out rate and the intercept of the line fitting the tail of the time-intensity curve with the axis $t = 0$. A schematic of the features is shown in Fig. 5.7. More about the use and definition of these morphologic features to describe the contrast agent intake can be found in the related literature [104] [57] [27].

Furthermore, the clustering procedure was repeated incorporating the ADC of each voxel as an additional feature to the morphologic descriptor vectors calculated previously. The ADC and the morphologic features were standardized by subtracting their mean and dividing by their standard deviation. The results of these two procedures were compared with our method in order to assess the clustering and data representation outcome.

5.8.3 Results

The regions resulting from dissimilarity based clustering were rendered as colored overlays on the morphological images on each slice. The results from a representative patient are displayed in Fig. 5.8. After clustering was performed on the normalized curves, the resulting clusters were assessed by the radiologists to validate the segmentation of both the central tumoral and surrounding regions. Fig. 5.8(b) shows examples of the clusters obtained, while Fig. 5.8(c) and (d) represents the plots of the average time-intensity perfusion curves calculated on the raw and normalized data respectively. The plots show the impact that the normalization step has in highlighting the inter-cluster differences. The central region exhibits a characteristic pattern in the DCE-MRI of a high early enhancement followed by a rapid wash-out, indicative of angiogenesis (Fig. 5.8(d), red line). Typically, surrounding this central region lays a cluster featuring a pattern of rapid enhancement followed by a signal plateau (Fig. 5.8(d), orange line). The outermost cluster surrounding these two central regions features a slow enhancement behavior (Fig. 5.8(d), yellow line). The voxels corresponding to the each cluster were extracted from the spatially registered 3D ADC maps in order to perform statistical anal-

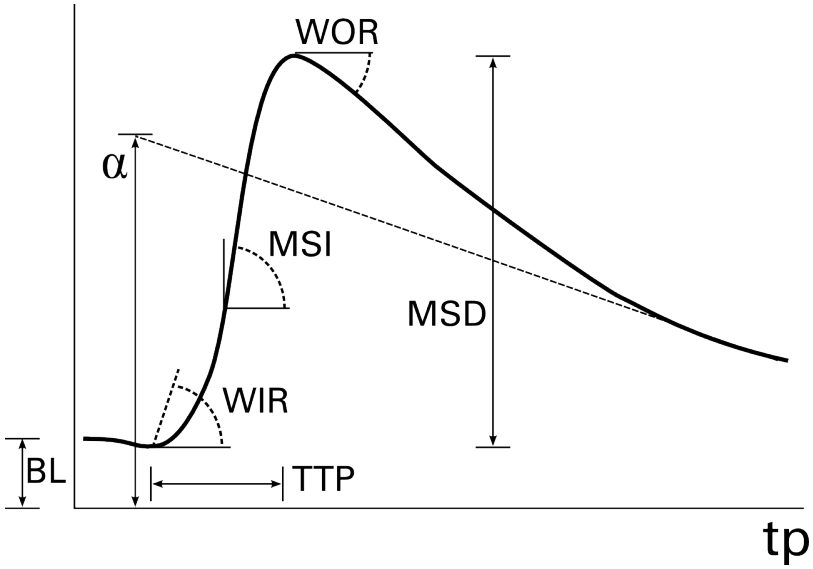


Fig. 5.7: Classic morphologic features derived from the DCE-MRI time-intensity curves, used for comparison purposes: Signal baseline (BL), time to peak (TTP), wash-in rate (WIR), maximum slope of increase (MSI), wash-out rate (WOR), maximum signal difference (MSD), maximum enhancement (MSD/BL), area under the curve, and the intercept of the line fitting the tail of the time-intensity curve with the axis $t = 0$ (α).

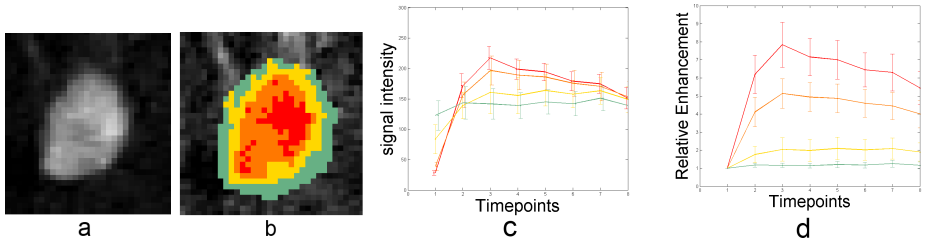


Fig. 5.8: DCE-MRI image (a) and overlayed lesion clustering (b),comparison between the average raw (c) and normalized curves (d) calculated for each cluster.

ysis. The analysis was carried out in the whole 3D ROI, that is, taking into account the ADC values corresponding to all the clustered slices as a single volume. Normality tests (Jarque-Bera) revealed that the ADC values for the different clusters analyzed were not normally distributed. Accordingly, a non parametric test (Wilcoxon-signed-rank test) was used ($p = 0.05$) to evaluate whether the tumor's subregions corresponded to regions in the ADC maps with statistically different PDFs. In this way we found that the distributions of the ADC values in the DCE-MRI defined regions were statistically different, in each one of the two conditions, in 19 out of 21 patients.

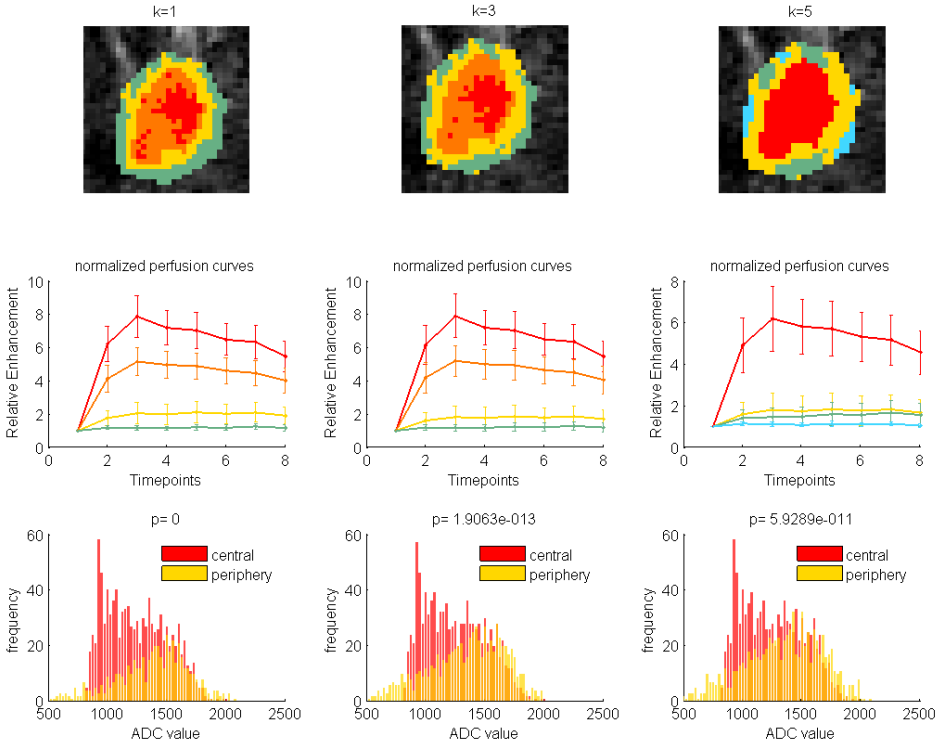


Fig. 5.9: Clustering results using different values for the tuning parameter k_{ADC} (1, 3 and 5).

The radiologist reviewed the overlays in comparison to the DCE seen as a dynamic loop, the DWI images and the ADC maps derived from them, as well as T2 STIR images. Criteria for the review were whether or not any of the subregions obtained by the method corresponded to a zone of necrosis based on the complete set of images, and whether one or more regions that would be classified as either benign or malignant have been subdivided.

Figure 5.9 illustrates a typical case setting k_{ADC} to 1, 3 and 5. From the obtained results it was highlighted by the experts the usefulness of varying the parameter k_{ADC} to emphasize different characteristics of the lesion. A high k_{ADC} allows the discrimination between the core tumor and the surrounding regions by giving a higher weight to the difference between ADCs. This is mainly due to the fact that there is a progressive increase in ADC from the core of the tumor to peritumor tissues to normal tissues, that leads to the possibility to use the ADC for locoregional staging [9]. Lowering k_{ADC} allows the subdivision of the core based on DCE-MRI dissimilarity and the evaluation of the heterogeneity of the tumor thanks to the balanced contribution of DCE and DWI in the distance function D.

For the sake of cluster comparison and validation among different methods, the silhouette analysis was used in all the clustering results. As it is explained in Section 3.8.1 the silhouette analysis measures how close each point in one cluster

is to points in the same cluster and how far away it is to points in the neighboring clusters. This is performed by quantitatively comparing the clusters by their tightness and separation and its average width provides an evaluation of cluster validity [155].

The silhouette analysis of our method highlighted an improved performance of 31% for the clustering performed using $k_{ADC} = 1$ with respect to the established approach that employs morphologic features derived from the DCE-MRI time-intensity curves, including the case where the Apparent Diffusion Coefficient was incorporated as an additional feature (Table 5.1).

Table 5.1: Silhouette analysis scores describing cluster compactness and separation for the whole ROI and for each relevant region for the Kinetic Features and the Multi-Modal Lesion Assessment (MMLA) methods (the higher the better).

Method	Mean	Central 1	Central 2	Periferic
Morphologic Features	**0.51**	0.53	0.51	0.49
Morphologic Features+ADC	**0.47**	0.49	0.48	0.44
MMLA	**0.62**	0.57	0.65	0.61
MMLA+ADC, $k_{ADC} = 1$	**0.62**	0.57	0.64	0.62
MMLA+ADC, $k_{ADC} = 3$	**0.58**	0.54	0.59	0.60
MMLA+ADC, $k_{ADC} = 5$	**0.57**	0.56	0.58	0.58

5.9 Discussion

As a general strategy, we have demonstrated a dissimilarity clustering based on multi-dimensional data derived from diffusion and perfusion MRI. Extension of the algorithm to additional data is straightforward, though the computational demand rises, and the similarity metric will likely need to incorporate further context-specific knowledge. As examination of tumor heterogeneity is carried out on a tumor by tumor basis, the data space can be restricted to areas containing lesions already located, but not necessarily segmented. For the specific use of DCE and DW-MRI, the lower resolution of the DWI data presents an issue of partial volume effects that affects the clustering of small lesions, but this issue is not specific to any one characterization strategy.

The two free parameters of the protocol; number of clusters (K), and relative weighting of the diffusion data (k_{ADC}), warrant discussion as the present work provides only a starting approximation to their choice, and the values may well be pathology dependent. For an unsupervised classification as used herein, the number of clusters should follow the actual structure and separation of the data into natural groups.

For breast tumors such as ductal carcinoma, the reporting of DCE-MRI data is currently based on a three-way division, while DWI is binary between normal

and abnormal. The three DCE curve types (a rise and fall, a rise to a plateau, and a steady rise) have established clinical utility in predicting tumor malignancy [112]. This is not to say however, that only three subgroups are possible, nor that these subgroupings are predictive of treatment response, which is the motive for examining tumor heterogeneity. In fact, works such as [104] have demonstrated that as the temporal resolution increases, a higher number of curve archetypes can be naturally identified and can be used for classification of voxel-wise perfusion curves.

We consider it noteworthy therefore, that when K was reduced to just three or four groups, these were identifiable with the 3 enhancement patterns (or these three and non-enhancement) used in clinical practice for the assessment of the breast cancer. As well, the confines of the groups with DCE-MRI time-course patterns consistent with malignant and benign tumors coincided very closely with the tumor margin drawn by a radiologist. Increasing the K value showed the expected progressive splitting of these groups as K increased, with k_{ADC} providing a distinction in the way this splitting proceeded based on the relative weight given to the diffusion data. The benefits of increasing the number of clusters are evident for understanding the heterogeneity of the lesion and the distribution of voxels that share certain similarities, however the increase of the number of clusters should go hand to hand with cluster and data analysis techniques in order to avoid false or meaningless divisions.

The primary criteria for non-invasive assessment of tumors based on DCE MRI involves three enhancement patterns (four including necrosis / non - enhancement). In the clinical data used for this study this assessment criteria has limited the validation to the visual interpretation of enhancement patterns based on the conventional interpretation of DCE curves, with a reader dependent incorporation of ADC information.

Ultimately, the envisaged application is in anticipating and evaluating treatment response. If tumor heterogeneity in terms of both perfusion and diffusion is to be encompassed, the conventional 3-way categorization may not be adequate or appropriate and indeed for other organs this rating is less common.

We are now looking into robust methods for further validation of the processing pipeline that would enable a clinical exploitation of the multimodal analysis. Access to ground truth beyond radiological and biopsy evaluation is needed and likely requires voxel-wise comparison with histology of resections, a process that requires modifications to the surgical procedure that were not justified for this first demonstration and research of the method. Even in the case of available histologic image data, a significant task remains in the spatially correlation of individual MRI voxels with the histological results in order to get the requisite voxel-scale validation.

In this chapter we presented a general methodology for heterogeneity quantification that integrates information from diffusion (an indicator of cellularity) and perfusion (reflecting blood volume, flow and vascular permeability) MRI images, and illustrated its use in application to ductal carcinoma. The demonstration illustrated multimodal clustering leads to improved selectivity and yields a greater refinement of the segmentation of tissues within the lesion than the separate processing of the two modalities.

By demonstrating that statistically consistent subgroups can be defined within tumors based on a combination of DCE-MRI and DWI-MRI data, we have indicated a means for objectively segmenting tumors that can be used for larger studies to examine clinical impact. Moreover, the appearance of statistically distinct perfusion regions within the tumor at moderate and low ADC weightings that in turn have statistically distinct ADC distributions suggests there is a useable distinction present that is not capitalized upon in present clinical practice.

6

A multi-view approach to multi-modal MRI clustering

6.1 Overview

In this Chapter we present our final methodology for the integration of multi-modal MR images for the unsupervised segmentation tumoral lesions for heterogeneity assessment. This "multi-view" imaging approach calculates multiple vectorial dissimilarity-spaces for each modality and make use of the concepts behind cluster ensembles (CE) to combine a set of base unsupervised segmentations into an unified partition of the voxel data. In the final part of the Chapter we evaluate the method with synthetic MRI datasets.

6.2 Introduction

Generally, it is the case that a discriminating a strategy based on a unique imaging modality is unable to appropriately differentiate normal from cancerous tissue, thus requiring a multi-modal view of the tissue for clinical assessment. Moreover, since many tumors, such as human glioma, are characterized by topographically heterogeneous histopathology or have evolved to different stages of tumor progression that may impact on their response to a given therapy, it is important to obtain a complete coverage of the lesion. This need arises also in response to the limitations of biopsies,in which, besides being difficult to localize and repeat, a sampling error at limited biopsy may mean that the specimen does not reflect the degree of malignancy elsewhere in the tumour and may result in significant mislabeling [198].

A single glioma can display regions of tumor-infiltrated brain tissue, regions containing a high density of tumor cells, and necrotic regions. This heterogeneity within a single tumor calls not only for a simple distinction of normal from pathologic tissue but to the development of methods for the assessment and segmentation of subregions. Preoperative heterogeneity assessment and grading helps in better treatment planning and management [198].

Going beyond the use of a single non fully discriminant MRI modality, numerous studies have shown that the combined information from multiple imaging modalities can yield improved discrimination of diseased tissue [122]. However,

the fusion of dissimilar imaging data for classification and segmentation purposes is not a trivial task. There is an inherent difference in information domains, dimensionality, scale and, due to limitations in certain imaging protocols, imaging resolution.

From the many MRI modalities performed in clinical practice, of particular interest are Dynamic Contrast Enhanced (DCE-MRI) and Diffusion Tensor Imaging (DTI-MR). DCE-MRI uses serial acquisition of images during and after the injection of intravenous contrast agent and has been shown to reflect tumor vascularity [98, 185]. DTI is sensitive to the preferred direction of the microscopic diffusion of water molecules in tissue, such as the one ocurring along the white matter tracts, which is less restricted along the axis of a fiber than along its transverse direction [107].

In DTI a diffusion tensor (DT), a 3×3 positive-definite symmetric matrix, is calculated for each voxel from measurements in several directions of diffusion-sensitized magnetic gradients. Each DT characterizes the directionality and magnitude of the anisotropic diffusion occurring in that particular voxel. Many studies that incorporate DTI process the multidimensional diffusion information contained in the DT by the use of scalar indices such as the Fractional Anisotropy (FA) and Mean Diffusivity (MD). However, such scalar measures do not account for the full information present in the tensor, the spatial relationship between neighboring voxels and commonly require *a priori* knowledge of how pathology affects these measures [93].

Since DT are positive-definite symmetric matrices their mathematic definition restrict them to lie on a manifold of the space $\mathbb{R}^6$, which is known to be a cone embedded in that space. The structure is also determined by the anisotropy property of a particular tissue and the underlying geometry, such as in white matter, which confines the local neighboring tensors to a more restricted submanifold in $\mathbb{R}^6$ [187]. The approach of writing 6 DT components as a feture vector and embed it in a vector space in $\mathbb{R}^6$ is hindered by the non linear nature of DT. To address this problem there have been some attempts to use manifold learning techniques such as ISOMAP [93,187], in which the focus is learning low-dimensional embeddings that parametrize the underlying manifold structure of the tensors, generally through the use of geodesic distances along the Riemmanian DT manifold. Alternatively, kernel methods for manifold learning in DTI were initially used by Khurd [94], in which kernel principal component analysis (KPCA) and kernel Fisher discriminant analysis (kFDA) were used for group-wise statistical analysis and classification of voxel based DTI datasets between normal and diseased groups.

In this Chapter we present a methodology for the integration of multi-modal MR images for the unsupervised segmentation of brain lesions for heterogeneity assessment. The ultimate objective is not tumor detection but the unsupervised segmentation of tumoral lesions into zones of voxels that share certain similarity using available multi-modal images. This multi-view imaging approach calculates multiple vectorial dissimilarity-spaces for each modality and make use of the concepts behind cluster ensembles (CE) to combine a set of base unsupervised segmentations into an unified partition of the voxel data.

Cluster Ensembles (CE) address the problem of combining multiple *base clusterings* of the same set of objects into a single consolidated clustering. Each base

clustering refers to a grouping of the same set of objects or its transformed version using a suitable clustering algorithm or variations of the same algorithm. The consolidated clustering is often referred to as the *consensus* solution. A proven strategy is to create various base clusterings of the same set of objects using diverse clustering algorithms and a different number of clusters. This strategy relies on the requirement that the base algorithms have different biases, i.e. they make different errors on new instances.

The final partition is obtained with a consensus function which maps the set of base clusterings to an integrated final clustering [59]. In unsupervised classification it is often the case that the objects to be clustered have multiple facets or *views*, a clear example of this would be the end result of the different imaging modalities used clinically to assess the same tissue, each one conveys information belonging to a different domain than the rest. In CE different base clusterings may be built on these distinct views. In this work we extend the *multi-view* notion considering the derivation of different vectorial spaces for each modality and then, using a variety of clustering algorithms, calculating a set of base clusterings for each of them which later are combined using a Cluster Ensemble strategy.

To represent each MR modality in diverse spaces we relied on the concepts behind dissimilarity based representations (DBR). The dissimilarity representation is an alternative to the use of features in pattern recognition [141]. Objects are characterized through pairwise dissimilarities; instead of using an absolute characterization of the objects by a set of features, problem-centric knowledge is used to define a measure that estimates the dissimilarity between objects. The term *object* refers, in the present context, to the information represented by each particular voxel such as tensors, scalars and temporal series.

Using a variety of established metrics and distance functions we calculate several dissimilarity spaces for each MR modality. In the case of DTI the used metrics are of different nature; from metrics that use scalar indices to Riemmanian geometry metrics that employ the full tensor information and others based on statistical divergence.

Kernel Principal Component Analysis was used as a non-linear manifold learning technique to address the aforementioned constrains imposed by the manifold in which the DT lie. A key difference to the way KPCA was employed in [94] is that in that work KPCA was used for statistical analysis of groups using as input the 6 DT elements of each voxel or the elements contained in a given neighborhood radius in the form of a vector, whereas in this work KPCA is performed using as inputs the dissimilarity spaces calculated with the DT metrics that make use of the whole tensor information. The use of DBR offers also the inherent advantage of having encoded in the vectorial space the relationship to each other voxel, expressed in the vector correspoding to the representation for each voxel.

6.3 Overview

Figure 6.1 shows the general methodology. Special emphasis should be placed to ensure the correct spatial registration between modalities. This initial registration problem is not covered here as it is the subject of a wide variety of succesful research efforts and frequently the strategy is problem-dependent [32, 147, 157]. The

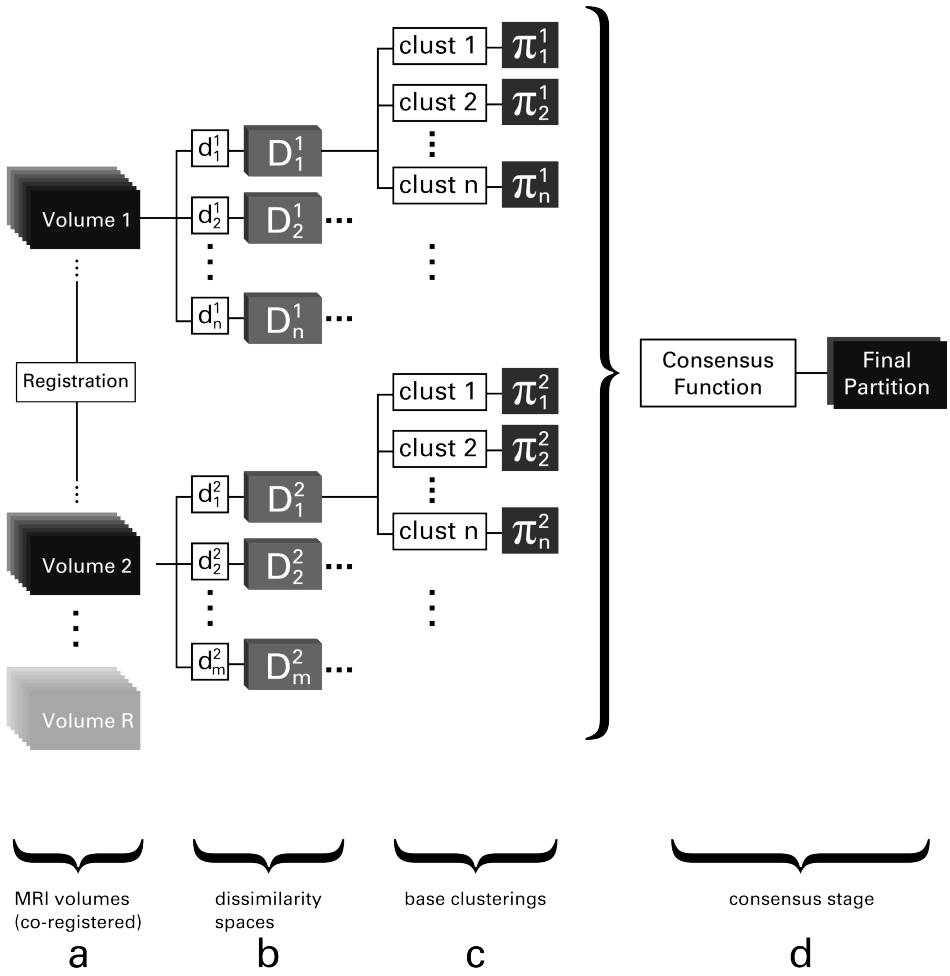


Fig. 6.1: A schematic diagram of the proposed multi-view methodology for cluster ensembles in multi-modal MRI. The initial multi-modal volumes are spatially co-registered and pre-processed (a), from each imaging volume a suitable assortment of distance functions is chosen and used to derive a corresponding set of dissimilarity spaces using voxel-wise relationships (b), a set of different clustering algorithms are applied to each dissimilarity space and a ensemble of base clusterings formed (c), a consensus function combines the many base partitions of the dataset into a final unified clustering (d).

rest of the methodology assumes the registration problem is effectively sorted out. As a second step a *Region of Interest* (ROI) is manually delineated around the detected lesion, larger than the lesion itself in order to include adjacent relevant voxels belonging to relevant zones such as edema and infiltration. The voxels from all the spatially registered multi-modality volumes contained in this ROI are the base for the dissimilarity spaces in which the MRI data are diversely represented. The following step consists in the actual calculation of the pairwise relationships between voxels and the creation of the vectorial spaces, as explained in detail in the following subsection. These diverse and unique vectorial representations of each MRI volume are the base of the methodology, as the set of clusterers operates separatedly in these different measurement spaces. In the case of the DTI volume a variety of established metrics are used, each one leading to a different representation of the data. For the DCE-MRI volume a distance function that measures the distance between two voxel-wise temporal series is used. The DCE-MRI distance function has an adjustable parameter that allows to diversely balance the information contained in the temporal series, this characteristic allows for the distinct representation of the DCE-MRI data in unique spaces. Following the representation of the data a set of diverse clustering algorithms is used to partition each one of the derived spaces for each MRI modality. From a single space and clustering algorithm several base clusterings may be calculated changing the desired number of clusters or the initialization parameters of the algorithm. The labels from all the calculated base clusterings are arranged in an ensemble matrix which serves as input to a *consensus function*. This function evaluates the relationships between all the datapoints belonging to the diverse base clusterings and produces a unified similarity matrix which is later partitioned hierarchically to obtain the final unified result.

6.4 Data Representation in Derived Vectorial Spaces

The first main approach of the proposed unsupervised clustering methodology is the representation of the MRI information into a set of diverse vectorial spaces. These diverse vectorial spaces that emphasize certain *views* or aspects from the lesion are constructed employing a variety of distance functions and metrics.

Unlike the known approach of creating cluster ensembles from the same data representation, in which each clusterer can be considered to produce an estimate of the same a posteriori class probability, in our multi-view approach many representations are created from a single MRI modality and used independetly by a set of clusterers. As it is the case when combining classifiers in supervised classification [95], in this case it is not possible to consider the a posteriori probabilities to be estimates of the same functional value, as the clustering systems operate in different measurement spaces.

Instead of using descriptive features, these vectorial spaces are derived from pairwise object comparisons, where the shared degree of affinity between two objects is captured by a dissimilarity value. There are many ways of comparing two objects, therefore there are many dissimilarity measures. Hand-picking the best set of measures for a specific problem requires additional knowledge of the behavior and the assumptions made by each specific dissimilarity measure, as well as

of the problem itself. Expert knowlege of the problem and the domain to which the specific imaging information belongs plays a crutial role in the definition and selection of the dissimilarities.

Formally, a dissimilarity space is constructed as a square matrix. This matrix consists of a set of row vectors, one for each voxel. These vectors represent the voxels in a vector space constructed by the dissimilarities to each other voxel. Usually, such a space can be safely treated as an Euclidean space equipped with the standard inner product definition (Pekalska). Let $X = \{x_1, \ldots, x_n\}$ be a voxel-based dataset. Given a dissimilarity function, a data-dependent mapping D is defined as $D(\cdot, R) : X \to \mathbb{D}^n$ linking X to a *dissimilarity space* [142]. The complete dissimilarity representation yields a square matrix consisting of the dissimilarities between all pairs of objects. In this matrix every object is described by an n-dimensional vector of distances between the object x and all the elements of X, such that $D(x, X) = [d(x, x_1) \ldots d(x, x_n)]^T$.

One of the advantages of this representation is that every classifier defined for feature spaces can be used in the dissimilarity space.

Choosing different dissimilarity measures allows us to construct a variety of vectorial spaces for each modality under consideration. Each one of them will serve as the space where unsupervised clustering algorithms will produce the base partitions for the cluster ensemble procedure.

6.5 DTI-MR processing

6.5.1 DT metrics

As it was described in the introduction, the succesful analysis of DTI poses a variety of challenges and constrains. The complexity of the diffusion data, belonging to a high-dimensional geometrical manifold structure in which the tensors are by definition restricted, requires a careful selection of methods that guarantee the correct use of the diffusion information fitted into the tensorial model.

Commonly, it is the case that if more than one metric is admissible or available to the problem, selecting among them as well as determining which representation would best characterize the relation between tensors in a relevant way becomes a important issue [136].

Part of this work aims at addressing the issue of choosing among the many measures and representations of complex data by proposing a methodology for the integration of relevant dissimilarity measures acknowledging that the information they convey can be used in a complementary way.

Instead of trying to choose a single measure to derive properties of diffusion tensors, we propose a multi-view approach to combine the results obtained with a set of them.

Many different measures have been proposed and defined to calculate dissimilarities between tensors. The measures present in literautre belong to a wide array of domains and it is often difficult to predict the outcome.

The calculation of dissimilarity spaces with DTI should rely in dissimilarity measures that are both informative and meaningful to the problem, incorporating expert knowledge and common assumptions.

A *measure* is defined as a function m that has two tensors $\mathbf{A,B} \in Sym_3^+$ as input, and returns a non-negative scalar value (Eq. 6.1) [139]. In this work we focus exclusively on dissimilarity measures that return the distance between two tensors.

$$m : Sym_3^+ \times Sym_3^+ \mapsto \mathbb{R}_0^+ \tag{6.1}$$

Although a case can be justly made against the use of scalar indices such as fractional anisotropy (FA) and mean diffusivity (MD) to reduce the tensorial information to a scalar value in order to infer relevant properties of the tissue under consideration, the widespread use of these scalar indexes in literature makes them relevant and well studied in a clinical context. For this reason we choose as a starting point two common scalar measures that employ important derivations of the full tensor, that is, the Fractional Anisotropy (FA) and the Mean Diffusivity (MD). These measures show just one aspect of the diffusion information, fitting in this way into the *multi-view* approach.

Both ds_{FA} and ds_{MD} follow the same structure, which is to calculate the absolute difference of the respective indexes between two tensors (Equations 6.2 and 6.3)

$$ds_{FA}(\mathbf{A}, \mathbf{B}) = |FA(\mathbf{A}) - FA(\mathbf{B})| \tag{6.2}$$

$$ds_{MD}(\mathbf{A}, \mathbf{B}) = |MD(\mathbf{A}) - MD(\mathbf{B})| \tag{6.3}$$

The angular difference (d_{ang_i}) (Eq. 6.4) of the eigevectors ($\mathbf{e}_i^D : i \in 1, 2, 3$) is also considered as a distance between tensors. It measures the changes of orientation and can be calculated with respect to any of the eigenvectors, although commonly it is restricted to the main eigenvector ($i = 1$) that describes the principal direction of diffusion in a particular voxel.

$$d_{ang_i}(\mathbf{A}, \mathbf{B}) = arccos(\mathbf{e}_i^{\mathbf{A}} \cdot \mathbf{e}_i^{\mathbf{B}}) \tag{6.4}$$

Certain measures, such as L^n-norms or the Frobenius distance treat the elements of the diffusion tensor as a vector. Several studies use the Frobenius distance and, although it ignores the actual structure and dependencies among tensors, it is included for evaluation considering its widespread use (Eq. 6.5).

$$d_F(\mathbf{A}, \mathbf{B}) = \sqrt{tr((\mathbf{A} - \mathbf{B})^2)} \tag{6.5}$$

Of more theoretical utility are the measures that use the full tensor information. Measures based on Riemannian geometry take into account the constrain of the diffusion tensors to be positive definite matrices. These measures compute the distances along geodesics in the manifold of symmetric positive definite matrices.

The *geometric distance* (dg) proposed by Batchelor in [14] belongs to this category. It is built on the curved space of positive definite tensors and penalizes small eigenvalues when the tensors approach the set of tensors with negative eigenvalues. The metric also has the additional property of being invariant under any linear change of coordinates. dg is calculated as follows (Eqs. 6.6 and 6.7):

$$d_g(\mathbf{A},\mathbf{B}) = N(\mathbf{A}^{-\frac{1}{2}}\mathbf{B}\mathbf{A}^{-\frac{1}{2}}) \tag{6.6}$$

where $N(\mathbf{D})$ is defined as

$$N(\mathbf{D}) = \sqrt{\sum_{i=1}^{3}(log(\lambda_i^{\mathbf{D}}))^2} \tag{6.7}$$

Also belonging to the Riemmanian category, we employ the *Log-Euclidean* metric d_{LE} proposed by Arsigny in [3,4], equivalent to the d_{L2} metric of the logarithm of the tensors. This metric is part of a general framework developed as a mean to perform fast and straightforward calculations between tensors, as the base for operations such as interpolation and regularization.

To calculate the logarithm an eigendecomposition is performed on the tensor after which the natural logarithm of the eigenvalues is calculated before recomposing again the altered diffusion tensor into a squared matrix.

$$d_{LE}(\mathbf{A},\mathbf{B}) = \sqrt{tr((log(\mathbf{A}) - log(\mathbf{B}))^2)} \tag{6.8}$$

Since another interpretation of a tensor is that of a covariance matrix of a Gaussian distribution that describes the diffusion at a particular voxel, it is natural to define statistical measures based on the overlap of probability density functions.

From this category we use the distance function proposed by Wang and Vemuri [189], based on the square root of the J-divergence (symmetrized Kullback-Leibler) between two Gaussian distributions corresponding to the diffusion tensors being compared (Eq. 6.9).

$$d_{KL}(\mathbf{A},\mathbf{B}) = \frac{1}{2}\sqrt{tr((\mathbf{A}^{-1}\mathbf{B} + \mathbf{B}^{-1}\mathbf{A})) - 2n} \tag{6.9}$$

where n is the size of the square matrices from which the distance is calculated, namely 3 in the case of a diffusion tensor.

6.5.2 Kernel Manifold Learning

In order to account for the diffusion manifold to which the DTI data is constrained we decided for the use of KPCA as a non-linear manifold learning technique. The use of kernel functions makes KPCA more computationally tractable than a general nonlinear feature extraction method.

Principal component analysis is a widely used statistical tool for dimensionality reduction [36, 74]. Let $x_i \in \mathbb{R}^d$, where $i = 1, \ldots, n$ be the training patterns. The principal components are a set of $q < d$ orthonormal vectors and span a subspace in the major directions into which the patterns extend. PCA finds the q major directions of maximal variance within the set of patterns $\{x_i\}$ and which also minimize the least-squares representation error for the samples [94].

Related to linear PCA, in which we find principal directions in the *input space* that maximize the variance of the projections of the samples along those directions, in KPCA similar principal eigendirections in a higher-dimensional Hilbert space are found. Unlike linear PCA which finds the best ellipsoidal fit for the data,

KPCA has the capability to extract non-linear features that are a more natural and compact representation of the data. KPCA can be regarded as a combination of two processes, a first process that implicitly transforms the *input space* into a higher dimensional *feature space*, and a second process that implements PCA in the *feature space* to extract a non-linear representation of the data by projecting it onto the subspace spanned by the eigenvectors of the q largest eigenvalues [90].

Thanks to the kernel trick the algorithm, if defined in terms of dot products, can be implemented in the *input space* by virtue of a kernel function, avoiding the need to perform the explicit computation of the data mapping (Eq. 6.10).

For a given non-linear mapping Φ, the *input data space* $\mathbb{R}^d$ can be mapped into the *feature space* $\mathcal{H}$:

$$\begin{aligned} \Phi : \mathbb{R}^d &\to \mathcal{H} \\ x &\mapsto \Phi(x) \end{aligned} \tag{6.10}$$

As a result, a pattern in the original *input space* $\mathbb{R}^d$ is mapped into a higher dimensional *feature space* $\mathcal{H}$, which is a reproducible kernel Hilbert space.

Assume that the mapping $\Phi(x_i)$ is centered, that is $\Sigma_{i=1}^{N}\Phi(x_i) = 0$.

Being k a suitable kernel function (see REF SCHOLKOPF), where $k(x_i, x_j) = \langle \Phi(x_i) \cdot \Phi(x_j) \rangle$, a kernel matrix K is calculated as

$$K_{ij} = \langle \Phi(x_i) \cdot \Phi(x_j) \rangle = k(x_i, x_j), i = 1, \ldots, N, j = 1, \ldots, N \tag{6.11}$$

Given a set of training samples $x_i \in \mathbb{R}^d$, where $i = 1, \ldots, n$, the covariance operator in the *feature space* $\mathcal{H}$ can be expressed as

$$S_t^{\Phi} = \frac{1}{N} \sum_{j=1}^{N} (\Phi(x_j) - m_0^{\Phi})(\Phi(x_j) - m_0^{\Phi})^T \tag{6.12}$$

where $m_0^{\Phi} = \frac{1}{N} \sum_{j=1}^{N} \Phi(x_j)$. In a finite-dimensional Hilbert space this operator is generally called covariance matrix. The convariance operator satisfies the following properties: it is bounded, compact, positive and it is a self-adjoint (symmetric) operator on the Hilbert space $\mathcal{H}$ [199].

Considering that every eigenvalue of a positive operator is nonnegative in a Hilbert space [W. Rudin, Functional Analysis. McGraw-Hill, 1973.], it follows that all non-zero eigenvalues of S_t^{Φ} are positive.

A principal eigenvector in the higher-dimensional Hilbert space $\mathcal{H}$ lies in the span of the vectors $\Phi(x_i) - m_0^{\Phi}, i = 1, \ldots, N$. Every eigenvector of S_t^{Φ}, β, can be linearly expanded by

$$\beta = \sum_{i=1}^{N} a_i \Phi(x_i) \tag{6.13}$$

where a_i is an N-dimensional vector. To obtain the expansion coefficients, let us denote $Q = [\Phi(X_1), \ldots, \Phi(X_N)]$ and, to form an $N \times N$ Gram matrix $\widetilde{K} = Q^T Q$, we use a suitable kernel function k (see REF SCHOLKOPF) where $k(x_i, x_j) = \langle \Phi(x_i) \cdot \Phi(x_j) \rangle$. The kernel or Gram matrix is calculated as

$$\widetilde{K}_{ij} = \langle \Phi(x_i) \cdot \Phi(x_j) \rangle = k(x_i, x_j), \quad i = 1, \ldots, N, j = 1, \ldots, N \tag{6.14}$$

Since commonly it is the case that the data is not centered in the *feature space*, that is $\Sigma_{i=1}^{N} \Phi(x_i) \neq 0$, $\widetilde{K}_{ij}$ must be centered by

$$K = \widetilde{K} - 1_N \widetilde{K} - \widetilde{K} 1_N + 1_N \widetilde{K} 1_N \tag{6.15}$$

where 1_M is a $N \times N$ matrix defined as $1_N = (1/N)_{N \times N}$. After $\widetilde{K}$ is appropiately centered, solving the eigenvelue problem $K\alpha = \lambda\alpha$ we obtain the corresponding orthogonal eigenvectors $\alpha_1, \alpha_2, \ldots, \alpha_n$ corresponding to the largest eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. The orthonormal eigenvectors of $\beta_1, \beta_2, \ldots, \beta_n$ of S_t^{ϕ} corresponding to the n largest positive eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ are calculated as

$$\beta_j = \frac{1}{\sqrt{\lambda_j}} Q \lambda_j, \quad j = 1, \ldots, N \tag{6.16}$$

After the projection of the mapped sample $\Phi(x)$ onto the eigenvector system $\beta_1, \beta_2, \ldots, \beta_n$, we can obtain the KPCA-transformed feature vector $y = (y_1, y_2, \ldots, y_m)^T$ by

$$y = P^T \Phi(x), \quad \text{where} \quad P = (\beta_1, \beta_2, \ldots, \beta_n) \tag{6.17}$$

specifically, the jth KPCA feature component y_j is calculated by

$$\begin{aligned} y_j = \beta_j^T \Phi(x) = \frac{1}{\sqrt{\lambda_j}} \alpha_j^T Q^T \Phi(x) \\ = \frac{1}{\sqrt{\lambda_j}} \alpha_j^T [k(x_1, x), k(x_2, x), \ldots, k(x_n, x)], \\ j = 1, \ldots, N \end{aligned} \tag{6.18}$$

In addition to finding the orthogonal directions of maximal variance in the higher-dimensional space $\mathcal{H}$, KPCA also provides an estimate of the probability density underlying the samples.

Girolami [61] presented an argument that the nonlinear features extracted using KPCA in conjunction with a Gaussian radial basis function kernel provides features that can be considered as components of an orthogonal series density estimate using Hermite polynomials.

Testing the KPCA

To illustrate the effectiveness of KPCA in calculating informative non-linear features from data laying in a given geometric manifold, we performed an experiment similar to that proposed by Khurd et al. in [94].

For simplification and ease of visualization, this test presents results on a dataset of points with variation in the radial and angular directions. The dataset was composed of points on a 2D plane, forming a semi-circular band. The dataset

was generated using 36 angles in the 0-144 degrees range and 6 radial values $(1.3, 1.4, 1.5, 1.6, 1.7, 1.8)$ which resulted in a dataset of 216 points.

Kernel Principal Component Analysis was applied to this dataset using a Gaussian kernel, setting the width parameter of the kernel, σ^2, to 0.1, which is a suitable function of the average distance between nearest neighbors. An automatic calculation of the optimal kernel width parameter (KPA, described in Section 6.5.3) yields a similar choice for the parameter σ^2.

The principal components obtained through KPCA can be plotted as isocontours representing the hyperplanes having constant projections onto the corresponding eigenvectors in the feature space $\mathcal{H}$. A visualization of the first seven principal components obtained through KPCA can be seen in Figures 6.2 and 6.3. As in the results obtained by Khurd, it can be seen that the first six components represent the angular changes in the data using varying scales. In addition it can also be appreciated in Figure 6.3 that the seventh KPCA component individually captures the radial change in the data and that it smoothly increases from negative to positive values as we move along the arc in the radial direction.

6.5.3 Kernel Parameter Selection

The choice of kernel to use requires some consideration, and indeed much of the research in the field of kernel methods is moving to address this question. From the viewpoint taken by Girolami in [61], the choice of kernel is determined by the desire to model any density function. The gaussian, radial basis function (RBF) kernel has well-known universal approximation properties, and fitting a sufficient number of them to continuous data provides a means of estimating an arbitrary density function which may be useful in certain applications and algorithms [61].

When a radial basis function kernel is used, such as the Gaussian kernel, one simple choice is to set the kernel width to the median distance between points in the aggregate sample. While this is certainly straightforward, it has no guarantees of optimality. The Gaussian kernel is defined by the following function

$$k(x_i, x_j) = exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2}) \tag{6.19}$$

Another common form to express the Gaussian kernel is shown in Eq. 6.20. Sometimes this formulation is preferred to that of Eq. 6.19 in order to work experimentally with q as a scale parameter.

$$\begin{aligned} k(x_i, x_j) &= exp^{(q\|x_i - x_j\|^2)} \\ \text{where} \quad q &= -\frac{1}{2\sigma^2} \end{aligned} \tag{6.20}$$

The smoothing parameter σ, also called *kernel width* plays an important role in the calculation of the Kernel matrix and the subsequent operations. Although a number of heuristics has been suggested to select the parameter σ there is not a universal consensus to reach an optimal value. In this work we use the method proposed by Jørgensen in [90] for kernel scale selection. This method makes use of a kernelized adaptation of Parallel Analysis (PA).

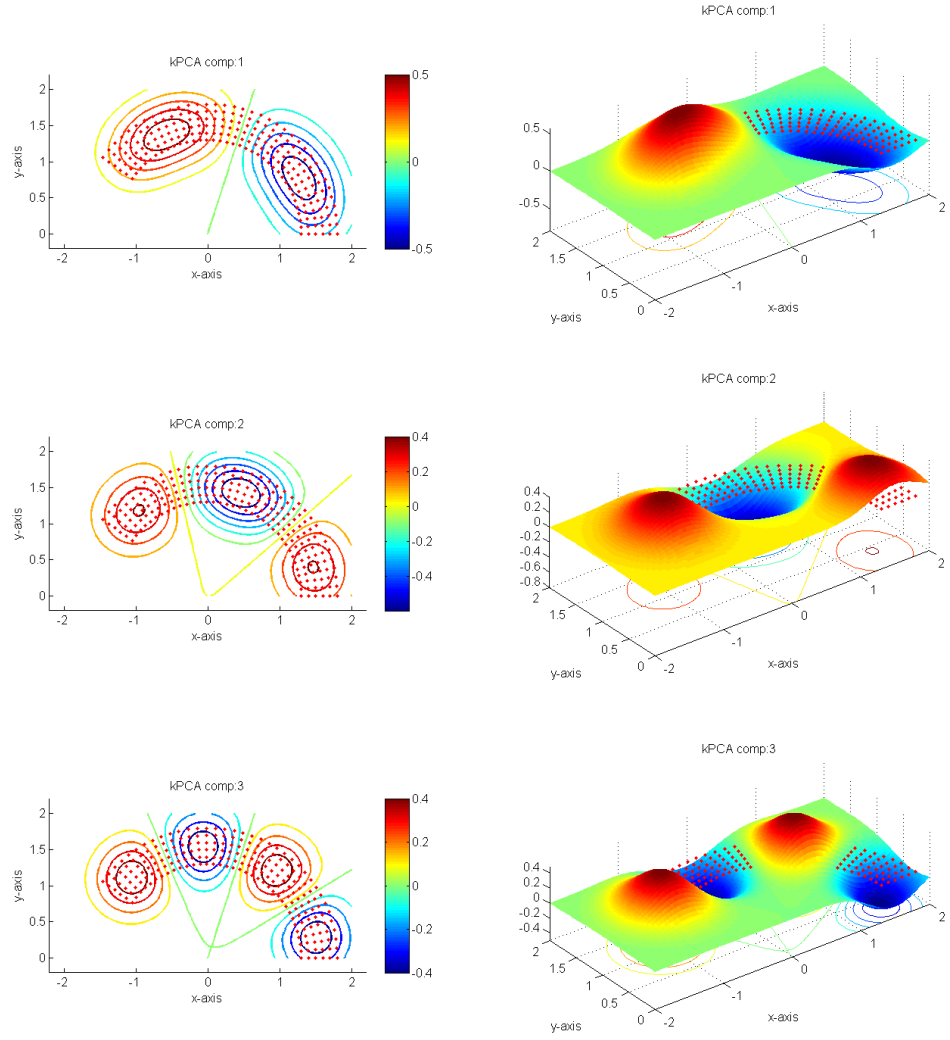


Fig. 6.2: First three KPCA components.

Parallel Analysis (PA) is a resampling based methodology for the estimation of the components or non-trivial factors to retain in linear PCA. PA compares the eigenvalues with the distribution of eigenvalues obtained by PCA on data sets distributed according to a null hypothesis of zero covariance. The PA null distributed data sets are obtained by permuting the measurements among the data points within each feature dimension and the number of factors to be retained is determined as the set of original PCA eigenvalues greater than the 95th percentile of the corresponding null distribution of eigenvalues.

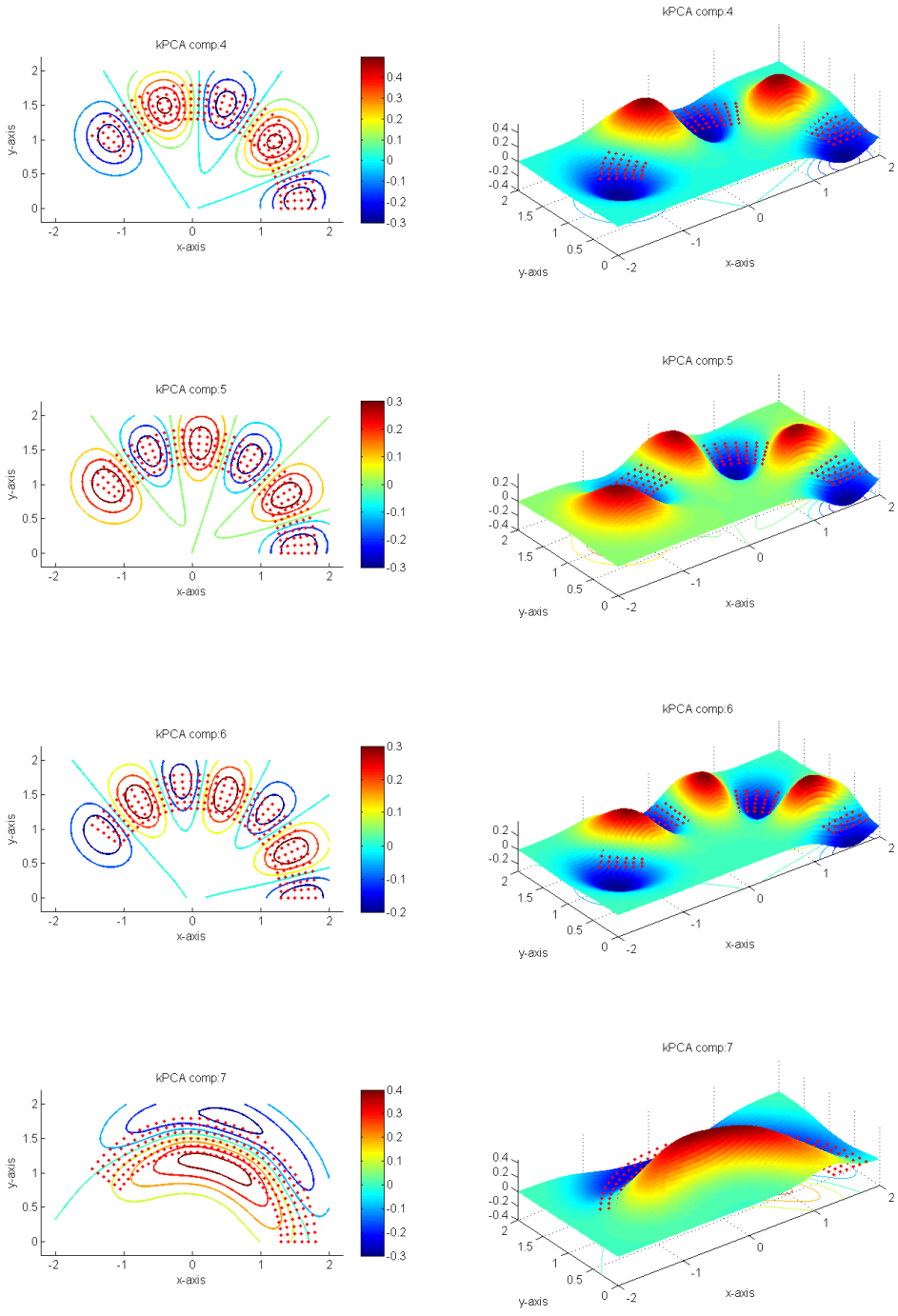


Fig. 6.3: Fourth to seventh KPCA components.

The idea behind PA is extended to KPCA to include the choice of kernel width σ for the Gaussian kernel. The resulting method is presented and referred as kernel Parallel Analysis (kPA) in [90].

In *feature space* the eigenvalue λ_i for component i extracted with of the PCA procedure is compared with the distribution of eigenvalues of null data sets obtained by permuting the data in input space p times.

For component i, a reference threshold T_i is set to the value of the 95th percentile in the distribution of the components of the eigenvalues. The number of components q to retain is chosen such that the original data eigenvalues are larger than threshold for all retained components. It is important to note that the original data eigenvalues, the reference thresholds, and the number of components q will depend upon the Gaussian scale σ

$$q(\sigma) = \max_{\lambda(\sigma)-T_i\sigma>0} i \tag{6.21}$$

A conservative estimate of the signal energy can be obtained as the cumulated difference between the original data eigenvalues and reference threshold levels

$$E(\sigma) = \sum_{i=1}^{q(\sigma)} \lambda_i\sigma - T_i(\sigma) \tag{6.22}$$

kPA chooses the kernel scale σ to maximize $E(\sigma)$. The energy is an estimate of the variance of the retained components in kernel space when accounting for the variance of null data. Thus, maximizing the energy in kernel space will maximize the variance of the true signal.

By column-wise permuting the data between samples for a given input dimension, kPA assures that the null-data is drawn from a distribution which has the same marginal distributions as the original data. Furthermore the input dimensions of the null distribution are statistical independent, i.e., the joint probability density function is fully factorized. This means that all manifold structures in *input space* are destroyed. This is a stronger condition than necessary in PA which only requires a null distribution with no covariance. Hence, the corresponding null distribution in *feature space* is that of a kernel mapped fully factorized distribution in *input space* with the correct input space marginals. The kernel spectrum of permuted data represents this null information. The distribution of the null kernel spectrum, as estimated by repeated permutation, allows us to determine when structure is present, identified in kPA as eigenvalue magnitudes rejected in the distribution of the null spectrum ($p < 0.05$). The details and pseudocode for the kPA procedure can be found in [90].

6.6 DCE-MRI processing

The distance function D_{DCE} (Eq. 5.1), presented in Section 5.3, was used to calculate the dissimilarity spaces derived from the DCE-MRI modality. We have used this function for calculating the pairwise proximity between DCE-MRI perfusion curves in the following works [123, 124].

As it was explained in Sec. 5.3 there are two main approaches to quantifiably compare two time-series: one makes use of the distances between the absolute values of their elements while the other focuses on the similarity of their behavior along time. Both criteria are quantified by D_{DCE}.

A parameter k_{DCE} was incorporated to Eq. 5.1 in order to weight the contribution of both types of similarities in time-intensity curves; the value-based similarity and the similarity with respect to their behavior (Sec. 5.3). The modified equation is then expressed as

$$D_{DCE}(S_1, S_2) = \frac{2}{1 + exp(k_{DCE}\textsc{Cort}(S_1, S_2))} dH(S_1, S_2) \tag{6.23}$$

where $S_1 = (u_1, \ldots, u_p)$ and $S_2 = (v_1, \ldots, v_p)$ are two voxel-derived perfusion curves sampled at time instants $(t_1, \ldots, t_p)$ [29,38]. *Cort* and *dH* stand for *Temporal Correlation* and *Hausdorff Distance*, which are defined respectively in Eq. 5.2 and Eq. 5.3.

The tuning parameter k_{DCE} weights the contribution of both Cort and *dH* to the complete dissimilarity measure D_{DCE} by modulating the shape of the sigmoid function. When $\textsc{Cort}(S_1, S_2)$ is in the negative range of 0 to -1 the total dissimilarity D_{DCE} approaches the value of $dH(S_1, S_2)$, on the contrary when $\textsc{Cort}(S_1, S_2)$ is in the positive range of 0 to 1, D_{DCE} is diminished accordingly.

6.7 Base clustering algorithms

As clustering algorithms we used the three methods described in Chapter 3.6: K-Means, Affinity Propagation and Support Vector Clustering.

6.8 Consensus function

In its most basic definition, consensus function Γ maps an ensemble $\Pi = \{\pi_1, \pi_2, \ldots, \pi_M\}$ composed of M *base clusterings* to a final unified partition π^*, $\Gamma : \pi_q \in \Pi | q \in \{1, 2, \ldots, M\} \mapsto \pi^*$.

When objects are connected according to their relations, it is possible to estimate the similarity of any object pair by using the underlying link information.

Being a normalized similarity matrix built on the average mutual memberships among base clusterings, the *ensemble co-association matrix* (CO matrix) (Eq. 3.16) indicates, for each pair of points, the proportion of times in which they are clustered together. However, despite the advantage of its simplicity and being a widely used method, it fails to account for hidden or unknown relations between data points and partitions.

Its own difinition makes the CO matrix prone to expose only a small proportion of the pairwise similarity between data points, that is, the obvious relationship that exists for a pair of points assigned to the same cluster in any given clustering, expressed originally in the binary similarity matrix (coassociation matrix) Eq. 3.15. This processing presupposes that the relationship between those two points is always zero for any given clustering when they are assigned to different partitions.

More information about the relationships may be better discovered by bringing in additional information regarding the relations or links that exist between clusters considering the complete ensemble.

To this end Iam-On [79] proposes the Connected-Triple approach applied to the cluster ensemble problem. This technique was originally developed to find duplicates of author names in large bibliographical databases [97]. The Connected-Triple approach works on the assumption that if two directly unconnected nodes in a graph share a link to a third node then this relationship is indicative of a certain degree of similarity between those two nodes.

A schematic illustrating this concept can be seen in Figure 6.4. The red circles $\{x_1, x_2\} \in \mathcal{X}$ represent a pair of data points as nodes in a graph with edges connecting each of them to a respective partition in each one of the three *base clusterings* $\{\pi_1, \pi_2, \pi_3\}$ depicted in the figure. There exists an edge between data point x_i and a cluster C_j^m if x_i was assigned to the partition C_j^m in the base clustering π_m.

In contrast to the techniques involving the $N \times N$ co-association matrix S_m, calculated for each one of the base clusterings π_m and in which the information between points not belonging to the same cluster is accounted as zero, the Connected-Triple approach intends to reveal the amount of information that exists between apparently unrelated data points and partitions. In figure 6.4, data points x_1 and x_2 are considered to be similar with respect to the base clusterings π_2 and π_3 on the grounds that they are both grouped together in partitions C_1^2 and C_1^3. On the contrary, when points x_1 and x_2 are evaluated exclusively with respect to the clustering π_1 their similarity is denoted as zero. The Connected Triple method starts with the idea that, despite being assigned to different partitions in clustering π_1, data points x_1 and x_2 may reveal a certain degree of similarity between partitions C_1^1 and C_2^1 when the relations of x_1 and x_2 to other clusterings where they are grouped together are taken into account. According to the Connected Triple technique, clusters C_1^1 and C_2^1 are similar due to the fact that they possess two Connected-Triples in which the clusters C_1^2 and C_1^3 are the centers of such triples.

Given a set Π of M *base clusterings*, a weighted graph $G = (V, W)$ can be constructed where V is the set of vertices representing clusters in the each one of the base clusterings π_m of Π and W is a set of weighted edges between clusters. The weight assigned to the edge w_{ij} connecting clusters C_i and C_j is estimated in accordance with the proportion of overlapping members (Eq. 6.24).

$$w_{ij} = \frac{\left|X_{C_i} \cap X_{C_j}\right|}{\left|X_{C_i} \cup X_{C_j}\right|} \tag{6.24}$$

where $X_{C_i} \subset X$ represents the set of data points that belong to cluster C_i. Originally in [97], the number of triples associated with any pair of objects is summed up as an integer. The approach proposed in [79] postulates that this simple counting, effective for data points or indivisible objects, might be insufficient to evaluate the similarity between clusters. In order to effectively take into account characteristics such as shared members among clusters the Weighted Connected-Triple regards each triple as the minimum weight of the two involving edges (Eq. 6.25).

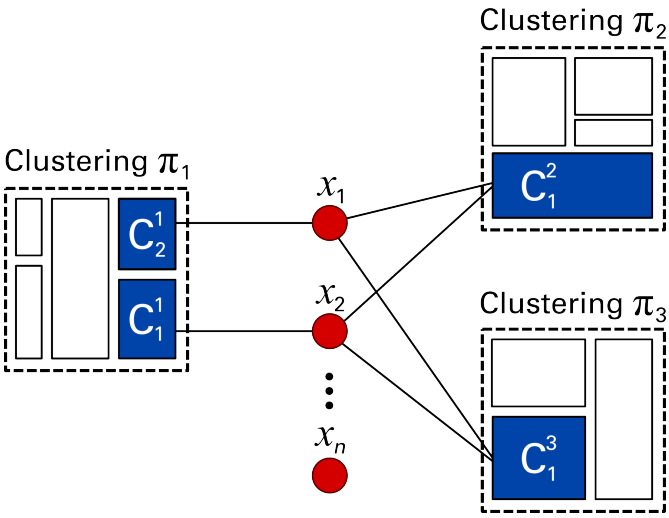


Fig. 6.4: A schematic representation of an ensemble composed of three base clusterings $\{\pi_1, \pi_2, \pi_3\}$.

$$WCT_{ij}^{k} = \min(w_{ik}, w_{jk}) \tag{6.25}$$

where WCT_{ij}^{k} is the count of the connected-triple between clusters C_i, C_j whose common neighbor is cluster C_k. The count of all triples $(1, \ldots, q)$ between cluster C_i and C_j is calculated as:

$$WCT_{ij} = \sum_{k=1}^{q} WCT_{ij}^{k} \tag{6.26}$$

Following Eq. 6.26, the similarity $Sim^{WCT}(i,j)$ between clusters C_i and C_j can be estimated with Eq.6.27, where WCT_{max} is the maximum WCT_{xy} value between any pair of clusters within the cluster ensemble Π.

$$Sim^{WCT}(i,j) = \frac{WCT_{ij}}{WCT_{max}} \tag{6.27}$$

The Connected-Triple approach is applied to the concept behind the co-association matrix in order to enhance the accounting of information between apparently unassociated clusters, which is generally regarded as zero with the classical co-association matrix formulation. More formally, for any base clustering $\pi_m \in \Pi, m = 1, \ldots, M$, a new type of co-association matrix is calculated where the similarity between any pair of datapoints x_i and x_j is estimated according to Eq. 6.28.

$$S_m(x_i, x_j) = \begin{cases} 1 & \text{if } C(x_i) = C(x_j), \\ Sim^{WCT}(C(x_i), C(x_j)) \times DC & \text{otherwise} \end{cases} \tag{6.28}$$

where $DC \in [0, 1]$ is a constant decay factor which denotes the confidence level of accepting two non-identical objects as similar. Following Eq. 6.28, each entry in the final Connected-Triple-based similarity matrix (CTS) is computed as:

$$CTS(x_i, x_j) = \frac{1}{M} \sum_{m=1}^{M} S_m(x_i, x_j) \tag{6.29}$$

6.9 Generation of Synthetic MRI Data

In an ideal experimental setting, methodological validation would be carried out with real anatomical datasets; noiseless and depicting clearly delimited, pathologically homogeneous and labeled regions of relative importance to the problem at hand. However in real settings it results very difficult to have such perfect and real *ground truth* data, with homogeneous delimited zones and all relevant variables controlled in a satisfactory manner. Moreover, in cases where the methodological scope includes statistical tests and quantitative voxel-based analysis among a diverse set of patients or clinical cases, the procurement of suitable ground truth for test and validation purposes becomes difficult or unpractical.

As a validation strategy, the creation of synthetic MRI data, either by software-based simulations or by the use of physical hardware phantoms, occupies a prominent place. The purpose of a software-based simulation is to construct a defined and controlled model of the anatomical environment and to simulate the signal acquisition of the pertinent MRI modalities in the synthetic voxel-based dataset.

In literature, two are the most prominent methods used for the generation of synthetic Diffusion Weighted MR data: the multi-tensor model and the model that calculates the restricted diffusion inside a cylinder of known dimensions.

6.9.1 Multi-Tensor Model

The multi-tensor model starts with the asumption that each fiber inside a single voxel can be described by a second order diffusion tensor. The signal in voxels containing more than one fiber is composed by the superposition of multiple tensors that describe the underlying probability density function. This model assumes that each fiber is independent in the sense that there is no exchange of molecules between the different fiber compartments [137].

For a given set of gradient directions g_i, the corresponding signal will be determined by

$$S(b, g_i) = \sum_{k=1}^{n} p_k \exp^{b g_i^T D_k g_i} \tag{6.30}$$

where b is the b-value and p_k are the different weights for the diffusion tensors D_k that compose the signal in the Stejskal-Tanner equation 2.9.

6.9.2 Restricted Diffusion in Cylindrical Geometry Model

This model, originally proposed by Söderman and Jönsson [169], describes the restricted MRI signal attenuation from molecules inside a cylinder of known dimension. In the presence of an underlying geometry that considers the existence of crossing fibers within a single voxel, the model assumes the presence of more than one cylinder, each representing a fiber component, and the signal attenuation is averaged from the independent signals within each cylinder. Multiple fiber orientations can be modelled with the assumption that the diffusing molecules are constrained within these cylinders with no possibility for exchange between the cylinders.

A pulse field gradient experiment consists of a pair of field gradient pulses of duration δ and magnitude g separated by duration Δ applied to a standard nuclear magnetic resonance spin-echo experiment. In this setup, if the short-gradient-pulse approximation is made, the echo attenuation is given by:

$$S(q,\Delta) = \int\int \beta(r_0) P(r|r_0,\Delta) \exp[i2\pi\mathbf{q}(r-r_0)] \partial r \partial r_0 \tag{6.31}$$

where $\mathbf{q}$ is the reciprocal state vector given by $\mathrm{q} = \gamma\delta g_i/2\pi$, where γ is the gyromagnetic ratio, $P(r|r_0,\Delta)$ gives the probability of a molecule being at r after a time Δ if it started at r_0 and $\beta(r_0)$ is the initial density of molecules.

Using this equation, Söderman and Jönsson [169] presented the equation for restricted water diffusion in a cylindrical geometry of radius ρ and length L, where the signal attenuation is determined by:

$$\begin{aligned} S(\rho,q,\theta,\Delta) = \sum_{n=0}^{\infty}\sum_{k=1}^{\infty}\sum_{m=0}^{\infty} & \frac{2K_{nm}\rho^2(2\pi q\rho)^4 \sin^2(2\theta)\alpha_{km}^2}{[(n\pi\rho/L)^2 - (2\pi q\rho\cos\theta)^2]^2} \\ & \times \frac{[1-(-1)^n\cos(2\pi qL\cos\theta)][J'_m(2\pi q\rho\sin\theta)]^2}{L^2[\alpha_{km}^2 - (2\pi q\rho\sin\theta)^2]^2(\alpha_{km}^2 - m^2)} \\ & \times \exp\left(-\left[\left(\frac{\alpha_{km}}{\rho}\right)^2 + \left(\frac{n\pi}{L}\right)^2\right] D\Delta\right) \end{aligned} \tag{6.32}$$

where J_m is the mth order Bessel function, α_{km} is the kth solution to $J'_m(\alpha) = 0$, with the convention that $\alpha = 0$. Kn_{nm} is a constant value depending on the values of the indexes n and m, θ is dependent on the gradient vector g_i as it represents the angle between the cylinder and the applied diffusion gradient direction, and $q = |\mathbf{q}|$ is the magnitude of the q-space vector.

The simulations from this model employ an exact form of the MR signal attenuation from particles diffusing inside cylindrical boundaries and has become a benchmark for other techniques, as well as the basis for multi-fiber signal reconstruction methodologies [101]. It has the advantage over the multi-tensor model that does not enforce mono-exponential decay.

6.9.3 Synthetic DTI Data Geometry

As an initial dataset to validate the multi-view approach, we created a synthetic dataset composed of four main zones, mutually differentiated by their underlying geometry and simulated diffusion characteristics.

These main four zones were designed to simulate the a cerebral glioma and the regions surrounding it. As in any other synthetic model, this dataset assumes a simplification of the complex characteristics of the tissue such as the soft boundaries that may be present in real cases, abnormalities within a single zone and unknown interactions between tissues not accounted by the model. However the model is clearly delimited in its scope of evaluation of the methodology.

There exists a wide variety of scientific literature describing the MR imaging characteristics of the various types of tissue involved in glioma lesions, and it is the case that some reports of characteristic values are in disagreement among the various different sources.

The four main zones of the synthetic dataset are designed to simulate the diffusion characteristics and geometric relationships of white matter, vasogenic edema, infiltrated fibers and central tumoral region. For the creation of the dataset we relied mainly in the reported measurements by Inoue et al. [82], in which an extensive preoperative evaluation of different grade gliomas is reported, Morita et al. [127] concentrates on the characterization of peritumoral edema, and Wieshmann et al. [192] who investigates the behavior of diffusion in cerebral abnormalities. Other relevant research works that investigate the characterization in diffusion imaging of cerebral tumors, specially gliomas, are [1, 5, 25, 117, 118, 149, 150, 150, 153, 166, 176, 176, 191, 197].

As a further refinement of the synthetic dataset, the zone corresponding to the white matter was designed to be composed of three subzones. These subregions represent two different white matter tracts and a mutual crossing zone. The two tracts follow a very similar pathway and were designed with the express purpose to test the effect of manifold learning on the discrimination methodology between extremely similar zones, such as is the case with normal and displaced white matter, which are zones that share practically similar diffusion characteristics. Figure 6.5 shows a schematic view of the white matter subregions; the first zone follows a circular path whereas the second one lays on an ellipsis. The crossing of both pathways occurs on the top and bottom of the tensor field where the designed paths meet.

Figure 6.6 depicts the spatial configuration of the synthetic DTI dataset, composed of 2500 voxels in a 50×50 tensor field. Depending on the zone, the general geometry was defined for each voxel and the diffusion signal was calculated with 21 gradient directions computed as the tessellation of the icosahedron of the unit sphere and b-value=1500 s/mm^2, using the accurate continuous approximation of the diffusion signal in Söderman and Jönsson restricted diffusion model by Barmpoutis [10]. The mean derived parameters for each zone, mean diffusivity and fractional anisotropy, are detailed in Table 6.1. Depending on the fractional anistopy for each zone, different degrees of random variation to the main pathway were introduced in each tensor to reflect the respective anisotropy with respect to the neighbors. Figure 6.7 and 6.8 show the tensor field visualized as classic ellipsoids.

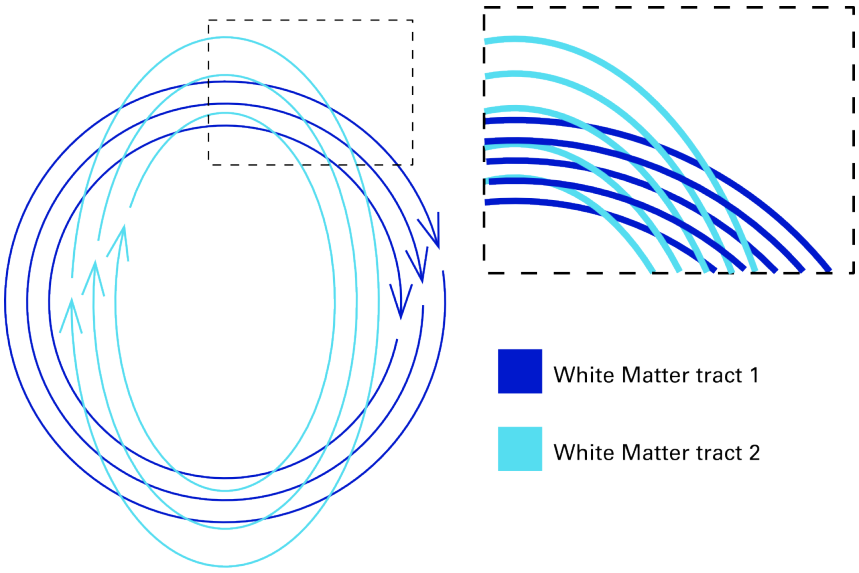


Fig. 6.5: Schematic of the two tracts composing the White Matter zone in the Synthetic DT dataset. The upper right rectangle shows an enlargement of the zone where the fibers cross. The signal of the crossing zone was calculated combining the diffusion signal of both tracts.

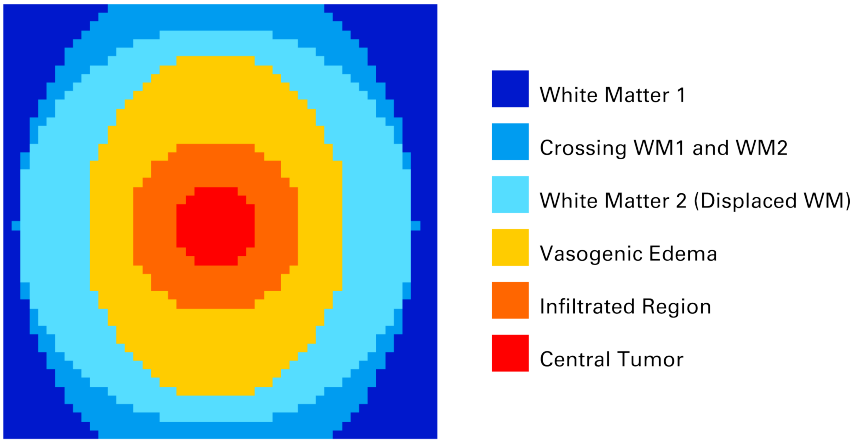


Fig. 6.6: The different zones composing the synthetic DTI dataset.

Region	**FA**	**MD** (mm^2/sec)
White Matter 1	0.74	7.63×10^{-4}
White Matter 2	0.70	7.98×10^{-4}
Crossing WM1 & WM2	0.72	7.81×10^{-4}
Vasogenic Edema	0.41	14×10^{-4}
Infiltrated zone	0.29	11×10^{-4}
Central glioma	0.15	11×10^{-4}

Table 6.1: Mean Fractional Anisotropy (FA) and Mean Diffusivity (MD) values for the different zones composing the synthetic DTI dataset.

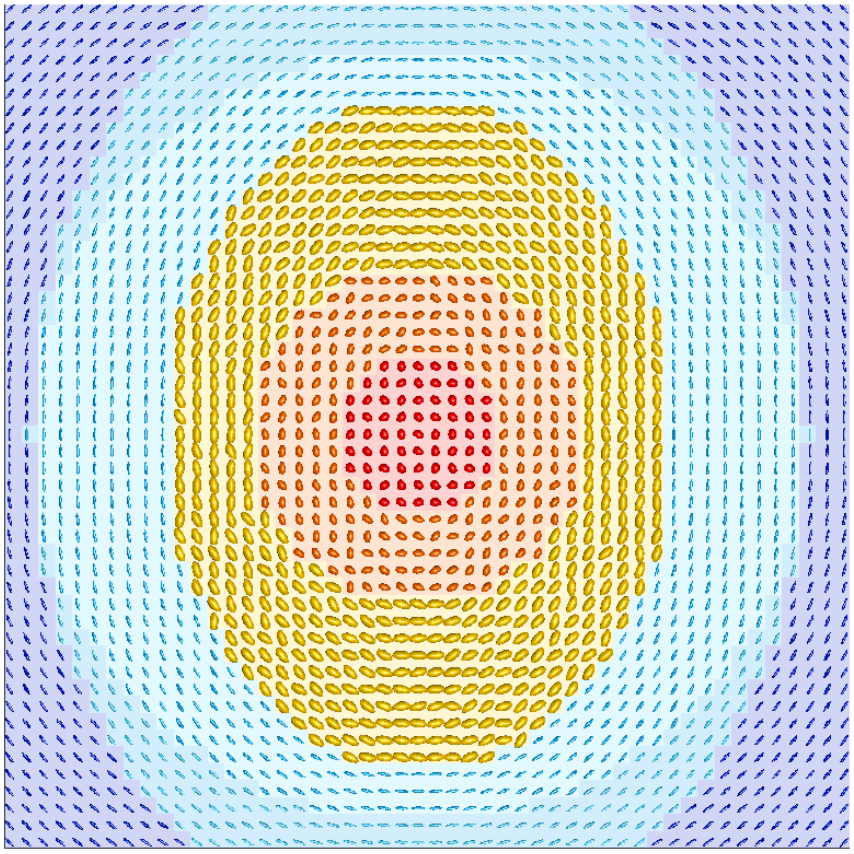


Fig. 6.7: Plot of the diffusion tensors visualized as ellipsoids in each one of the distinct zones of the synthetic dataset.

Fig. 6.8: Detail of the diffusion tensors, visualized as ellipsoids, corresponding to the zones in the synthetic DTI dataset.

6.9.4 Synthetic DCE-MRI dataset

In order to create a synthetic DCE-MRI dataset that corresponds spatially and functionally with the calculated diffusion tensor field we relied on the segmentation results and evaluation of a real glioma MRI volume. Following the methodology detailed in the previous chapter we performed a segmentation of a DCE-MRI volume. This volume was obtained using a pre-clinical mouse model of inoculated glioma cells.

The laboratory mouse shares extensive molecular and physiological similarities to humans and is a powerful tool for studying cancer. Unlike invertebrate model systems, tumor development in mice is accompanied by other complex processes such as angiogenesis and metastasis, similar to those in human cancer. More importantly, mouse tumor models provide temporally and genetically controlled systems for studying the tumorigenic process as well as response to treatment [26].

The DCE-MRI volume was acquired 35 days after inoculation and consisted on 60 timepoints. Just as in the previous chapter, a region of interest was drawn surrounding the lesion and a vectorial space was constructed using the voxel-wise dissimilarities between perfusion curves given by Equation 6.23. Furthermore, a second region of interest was added, this ROI was taken in the contralateral white matter in order to add a sizeable region of healthy white matter voxels to serve as reference for the. The corresponding vectorial space was clustered in 4 zones as in Section 5.3. From each zone the mean perfusion curve was extracted and used as a prototype of its corresponding region. A random variation was added systematically to each timepoint in the corresponding representative time-intensity curve in order to build a different perfusion curve for each voxel. The layout of the diverse zones is equal to that of the DTI synthetic dataset, as depicted in

Figure 6.9, with the main difference that in the DCE-MRI dataset there is no distinction between different subregions of white matter such as in the DTI dataset due to the fact that the contrast intake in white matter is the same indepently of the underlying geometry of the fibers. This model assumes a relation between the DTI and DCE information in co-registered voxels that might not exist in reality, however for the purposes of the methodological validation this presuposed relationship proves to be useful and valid.

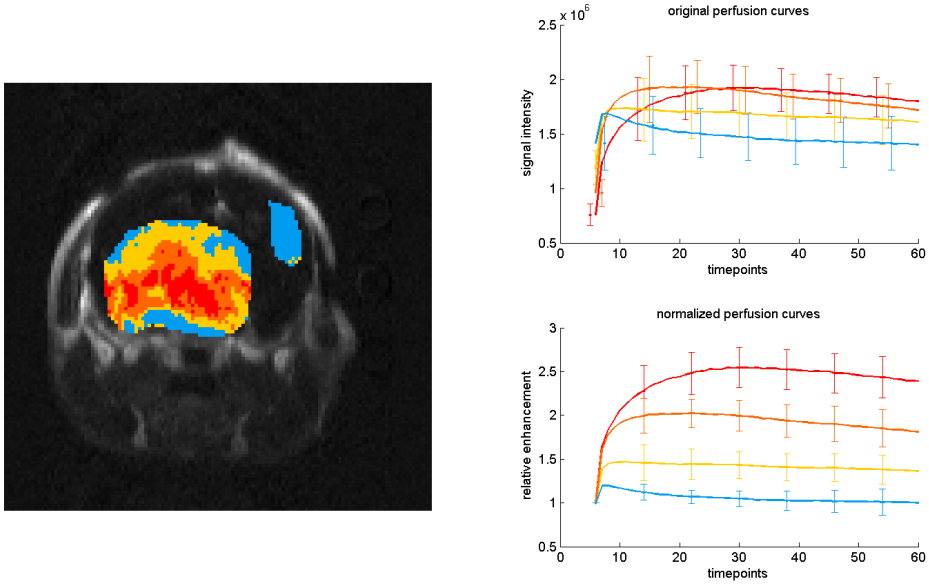


Fig. 6.9: The DCE-MRI volume from which the perfusion prototypes for the synthetic dataset were extracted (left). Original and normalized perfusion curves of the disntinct zones segmented in the volume (right).

6.10 Results

In order to test the multi-view methodology presented in this chapter and illustrated in Fig. 6.1, a synthetic MRI dataset was generated as described in Sec. 6.9.

Using the synthetic MRI dataset, distinct vectorial dissimilarity spaces were calculated using the DT metrics described in Sec. 6.5.1: ds_{FA}, ds_{MD}, d_{ang}, d_F, d_g, d_{LE}, d_{KL}. These vectorial spaces were created by the computation between every pair of diffusion tensors in the dataset, resulting in a $N \times N$ squared matrix for every dissimilarity space derived from any individual metric. For any DT-based dissimilarity measure D_{DT}, every voxel $x \in X$ belonging to the DT volume is represented in these dissimilarity spaces by the vector $D_{DT}(x, X) = [D_{DT}(x, x_1), D_{DT}(x, x_2), \dots D_{DT}(x, x_n)]^T$. Besides the use of DT metric-derived spaces, we considered the inclusion of the 6 DT components

(XX, XY, XZ, YY, YZ, ZZ) as another voxel-based input vector for the kernel manifold learning processing and subsequent clustering, just as suggested by Khurd et al. in [94]. It is important to remark that the direct vector representation using the 6 unprocessed DT components is not clustered directly, only their principal components resulting from the kernel manifold learning procedure. The first fifty principal components of the DTI-derived spaces were obtained by virtue of the KPCA methodology using a Gaussian Kernel and an appropriate kernel width calculated for each space [90].

From the DCE-MRI side, the time-intensity curves derived from the DCE synthetic dataset were processed in the same way as in Section 5.3, that is, using Eq. 5.1 as the dissimilarity function to form the corresponding vectorial spaces, which are diversely created varying the tuning parameter k_{DCE}, introduced in 6.6 to form Eq. 6.23. To this end k_{DCE} was varied from 1 to 5.

To create the ensemble of base clusterings we relied on three algorithms coming from different theoretical domains, described in Sec. 6.7: the classic K-means, Support Vector Clustering (SVC) and clustering by Affinity Propagation (AP) [56, 202]. The parameter k in the K-means algorithm was varied from 3 to 10, allowing for a heterogeneous ensemble encompasing diverse scales.

Three cases were generated for Affinity Propagation, with all self similarities set at -1, -5 and -10, without giving any preference to any data point to be regarded as an exemplar. These three values were selected with the same purpose as the diverse k values in K-means, i.e., to probe the data at diverse scales and add heterogeneity to the general ensemble.

All the vectorial spaces were clustered with these algorithms to create a comprehensive set of base clusterings. The resulting cluster ensemble was then used as input for the Connected-Triple link-based consensus function (Sec. 6.8) [79]. The output similarity matrix was partitioned hierarchically after which the final partition was assessed. A detailed diagram of all the steps followed in the methodology with DCE-MRI and DTI-MR is presented in Fig. 6.10.

For our analysis, we defined an assorted set of test cases to evaluate the effects of including base clusterings derived from different DT metrics to form the final partition through a consensus function. We hypothesized that the best results would be obtained using the base clusterings derived from metrics that use the full tensor information, an assumption that is based on the complex geometric nature of the DT-derived information and the expectations of the manifold learning techniques used in the methodology. The definition of the diverse test cases can be seen in Table 6.2. Besides the aforementioned DT metrics, all the test cases share the same set of base clusterings derived from the DCE-MRI volume.

The results of the tests performed with the diverse cases are presented in Table 6.3. To assess the results we have used the Classification Accuracy, Rand Index and the Adjusted Rand Index 3.8.2. As it was hypothesized the case using the DT metrics that make use of the full tensor information and manifold learning scored the best in the evaluation. This case was, thanks to an effective manifold learning, able to discern both white matter zones and their mutual crossing area, a task where the other cases failed in varying degrees.

Table 6.2: Definition of the diverse test cases used for evaluation of the methodology with synthetic datasets. Besides the using the DCE-derived base clusterings, these cases are composed of the individual base-clusterings obtained from a diverse set of DTI metrics. The last ones in particular make use of the metrics that use the full tensor information.

Method	ds_{FA}	ds_{MD}	d_{ang}	d_F	d_g	d_{LE}	dKL	DT elements
Case 1	x	x	x	x	x	x	x	
Case 2	x	x	x	x	x	x	x	x
Case 3	x	x	x		x	x	x	
Case 4	x	x	x		x	x	x	x
Case 5	x	x			x	x	x	
Case 6	x	x			x	x	x	x

Table 6.3: Results of the evaluation performed with synthetic datasets. The results are evaluated by means of the Classification Accuracy (AC), Rand Index (RI) and Adjusted Rand Index (ARI). For each of the defined test cases (Table 6.2) two different results were generated; one with the implementation of manifold learning techniques (labeled as KPCA) and one without. The use of manifold learning improved the results in all cases and, as it was expected, the case with the metrics that use the full tensor information proved to be the best among all.

Method	CA	RI	ARI
Case 1	0.719	0.826	0.515
Case 1 KPCA	0.864	0.915	0.750
Case 2	0.6848	0.801	0.489
Case 2 KPCA	0.842	0.884	0.662
Case 3	0.7312	0.831	0.533
Case 3 KPCA	0.866	0.911	0.7398
Case 4	0.726	0.833	0.529
Case 4 KPCA	0.852	0.900	0.7080
Case 5	0.7352	0.837	0.543
Case 5 KPCA	0.865	0.905	0.7275
Case 6	0.7312	0.834	0.532
Case 6 KPCA	**0.986**	**0.991**	**0.973**

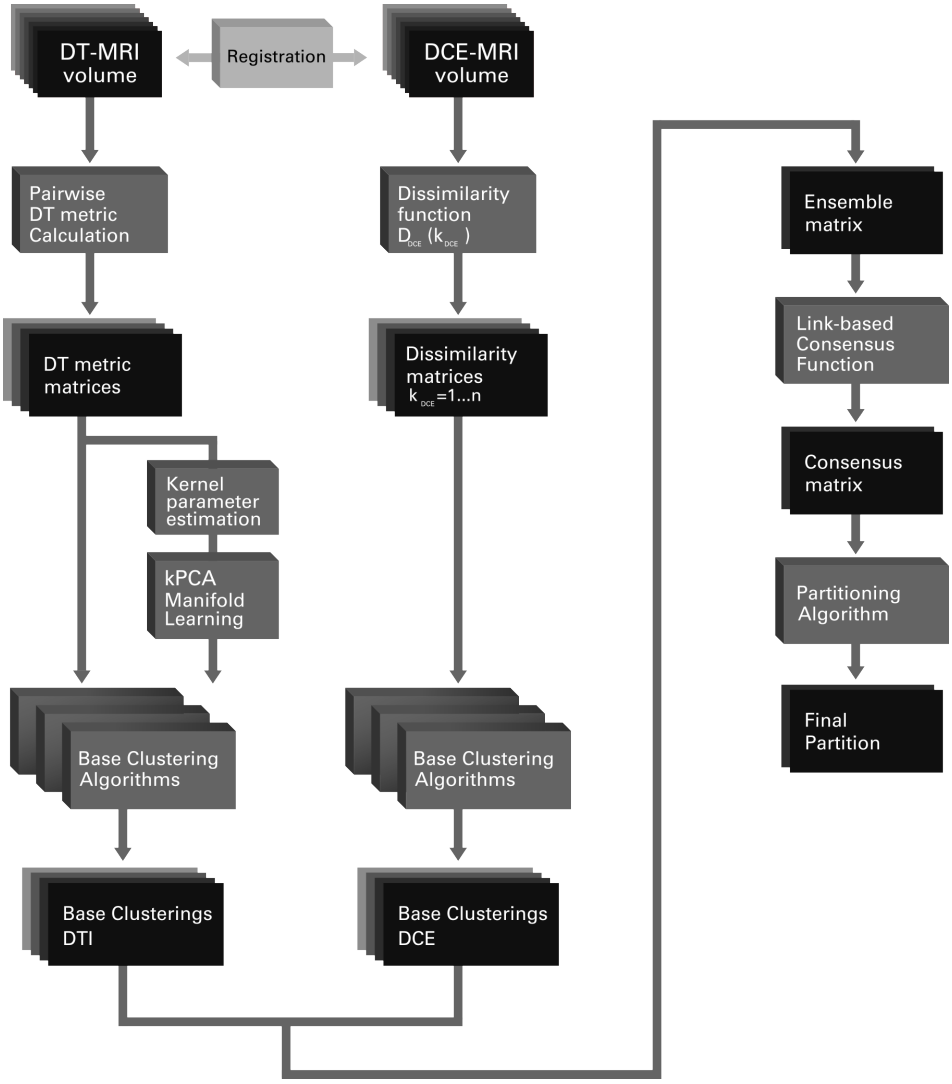


Fig. 6.10: Diagram of the multi-view cluster-ensemble methodology applied to DCE-MRI and DTI-MR volumes.

6.11 Tests with a software tumor simulator

To further the validation of our methodology we also relied on the simulator developed by Prastawa and colleagues [148].

Their proposed system combines physical and statistical modeling to generate synthetic multi-modal 3D brain MRI with tumor and edema, along with the underlying anatomical ground truth. They place emphasis on the simulation of the major effects known for tumor MRI, such as contrast enhancement, local distortion of healthy tissue, infiltrating edema adjacent to tumors, destruction and

deformation of fiber tracts, and multi-modal MRI contrast of healthy tissue and pathology.

The simulator synthesizes pathology in multi-modal MRI and diffusion tensor imaging (DTI) by simulating mass effect, warping and destruction of white matter fibers, and infiltration of brain tissues by tumor cells. It generates synthetic contrast-enhanced MR images by simulating the accumulation of contrast agent within the brain. The appearance of the the brain tissue and tumor in MRI is simulated by synthesizing texture images from real MR images.

In his published work Prastawa compares the manual segmentation of the simulated images by human experts with the reference ground truth masks generated by his simulator and finds a considerable agreement, arguing that the simulated images at least present a segmentation challenge similar to that of real pathological images for human experts.

Tumor and edema growth involves many concurrently occurring processes. The growth model may involve biomechanics, nutrient distribution, and metabolic processes. Since Prastawa's goal was not to model tumor growth in detail, they have chosen to simplify the model and use three separate sequential processes for efficiency, as shown in Figure 6.11. First, the simulator process the deformation that is due to tumor mass effect using a biomechanical model. It is then followed by the simulation of the infiltration process using reactiondiffusion. Finally, it computes the deformation that is due to tumor infiltration of brain tissue and the mass effect of edema [148].

6.11.1 Drawbacks of the tumor simulator

Despite its promised utility, the tumor simulator has certain drawbacks that need to be taken into consideration when using it for validation. First, it is important to note that with this simulator, the goal of Prastawa and colleagues was to generate sufficiently realistic MR images, or in other words, to generate MRI volumes that *appear to be realistic*, similar to real pathologic images. In the documentation this modified MRI images are specially used to assess the segmentation abilities of human experts. The accurate modeling of tumor growth and MR image synthesis are beyond the scope of the simulator, instead the focus falls on the generation of test images that empirically exhibits pathology seen in real images.

Our principal concern, however, is in the way the DTI is processed. As it can be seen in Figure 6.11, the DTI modification step is performed after the first body-mass simulation due to the initial tumoral seed and it is not influenced by the following step, which is the second body-mass simulation due to the infiltration process. The DTI modification process is based mainly in the observation that white matter fibers around a tumor tend to be displaced, and in regions near the tumor the mean diffusivity tends to increase while the fractional anisotropy usually decreases. Local volume explansion reduces tensor coherence, producing more isotropic tensors, while local volume compression or shrinking does not modify the tensor information. Starting from these basic assumptions, the simulator process the input DT volume using a combination of image warping and non-linear interpolation. Our main criticism of this methodology is that, although the assumptions regarding the changes in diffusivity are backed by the literature, the tensor modification process it too simplistic and relies only on the results of the first body-mass

iterations, without taking into account the results of the infiltration process. Since the final ground truth masks are mainly calculated by the tumor and infiltration processes, the modification of the DTI can be seen mainly as a byproduct that does not contribute to the final ground truth. This is particularly important for the validation of our methodology, which relies in the full tensor information and the relationships between all voxels when calculating KPCA as a manifold learning technique.

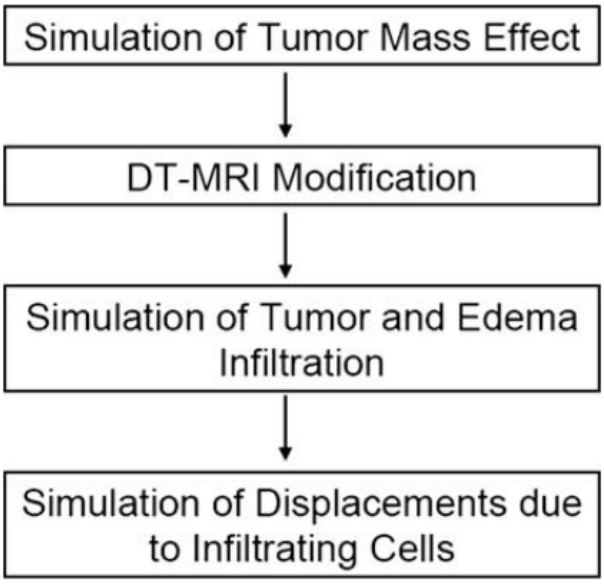


Fig. 6.11: Overview of the simplified tumor and edema growth model in the tumor simulator by Prastawa et al. [148]. The model is composed of four sequential processes, where we simulate the deformation due to tumor expansion, the modification of DTI due to the deformation, the infiltration of brain tissue by tumor cells and edema, and the displacements due to the infiltrating cells.

6.11.2 Test case

Despite the aforementioned drawbacks, we think that the tumor simulator is an interesting and useful tool, as long as we have in mind its limitations, specially when evaluating specific validation results. Another mundane reason that pushed us to its use is the known difficulty of obtaining real appropriate datasets for validation, a common problem in the world of medical imaging research. Using the tumor simulator, we generated a synthetic tumor with 20 iterations in the mass effect calculation of the tumor volume expansion on surrounding tissues and 20 iterations in infiltration, which simulates of the growth and spreading of tumor cells due to the infiltrating process. The full list of parameters is shown in Table

6.4. Most of the parameters refer to the biomechanical properties used to model the brain tissue and the interactions with the distinct pathological processes. The full description of the methodology used by the simulator to calculate every step of the process is described in the work by Prastawa [148].

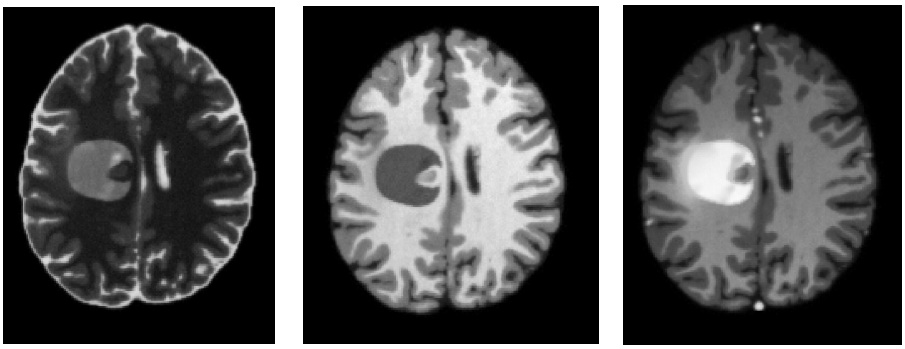
Table 6.4: Parameters used in the tumor simulator [148] to generate the test case.

Parameter	Value	Parameter	Value
Deformation Iterations	20	Infiltration Iterations	20
Infiltration body-force iterations	6	Deformation initial pressure	6
Deformation kappa coefficient	20	Deformation damping	0.05
Infiltration time-step	0.5	Infiltration early-time	6
Infiltration body-force coefficient	120	Infiltration body-force damping	0.25
Contrast-enhancement type	uniform	Brain young modulus	694
Brain poisson ratio	0.4	Falx young modulus	1200
Falx poisson ratio	0.4	Infiltration reaction coefficient	0.1
White matter tensor multiplier	200	Gray matter tensor multiplier	1
Gad noise stddev	20	T1 noise stddev	50
T2 noise stddev	50	Flair noise stddev	90
Gad max-bias degree	4	T1 max bias degree	4
T2 max bias degree	4	Flair max-bias degree	4
Disable background	1	Deformation solver iterations	8
Infiltration solver iterations	8	Number of threads	4

A sample of the resulting set of images from the test case can be seen in Figure 6.12, where the T1, T2 and T1 with contrast agent images exhibit a clear lesion zone and deformation. Figure 6.13 shows the resulting DTI volume as a visualization with diffusion ellipsoids. In this figure it can be clearly appreciated the effect described in the previous Subsection, i.e. the articial-looking swelling of the diffusion tensors that compose the initial tumoral seed in the middle of a more restricted tensor deformation surrounding that area. Such seed does not appear delineated in the final ground truth masks, as it is part of the region regarded as tumor. We applied to the results of the simulation the same procedure used in the previous synthetic dataset, as written in Section 6.10. It is worth noticing that the tumor simulator does not calculate a full DCE-MRI sequence, just a single image that simulates the contrast agent accumulation. The equations for the processing of DCE-MRI are used as if we had a single timpoint beyond the initial baseline volume.

The results shown in Figure 6.14 and Table fig:tumorsimout4 indicate an important agreement between the ground truth and the obtained segmentation. How-

ever the limitations of the simulation in the DTI processing lead us to think that a better result would be obtained in the case of a more comprehensive or relevant DTI volume. Nevertheless, this positive results warrant a further analysis of the methodology with real pathological datasets.



(a) T2 image (b) T1 image (c) Contrast-enhanced T1

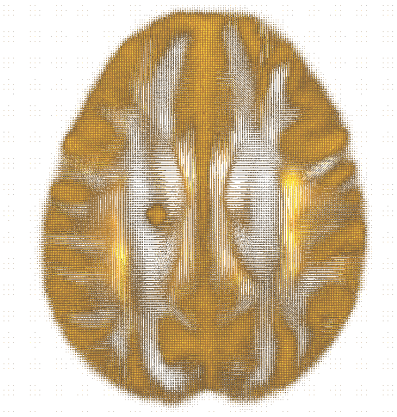
Fig. 6.12: Output volumes from the tumor simulator. The images correspond to a slice of the test case generated with the parameters shown in Table 6.4

Table 6.5: Results of the evaluation performed with the tumor simulator. The results are evaluated by means of the Classification Accuracy (AC), Rand Index (RI) and Adjusted Rand Index (ARI). Two different results were generated; one with the implementation of manifold learning techniques (labeled as KPCA) and one without. Despite the drawbacks inherent to the tumor simulator, explained in Section 6.11.1, the use of manifold learning improved the results.

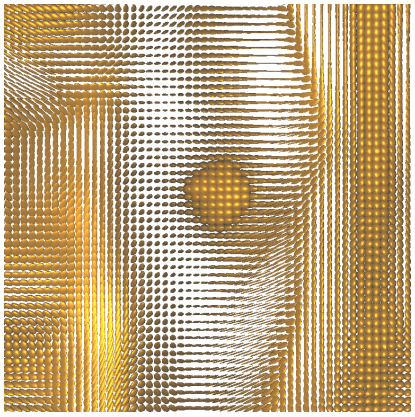
Method	CA	RI	ARI
Multi-view method	0.714	0.772	0.53
Multi-view method + KPCA	0.800	0.832	0.610

6.12 Conclusion

We have presented a general multi-view strategy for for creating a set of unsupervised base clusterings of lesions in multi-modal MRI images with the purpose to reach a unified consensus. The creation of multiple vctorial spaces from each MRI modality allow to focus on specific characteristics or views of the multi-dimensional

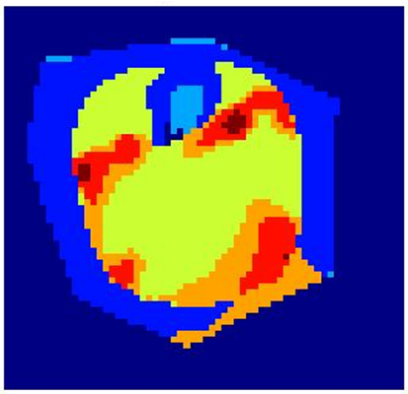


(a) DTI image, visualized as diffusion ellipsoids.

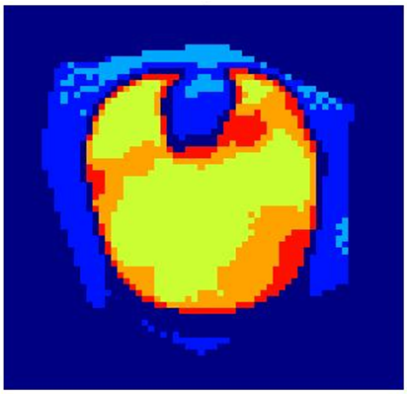


(b) Enlargement of the tumoral zone in the DTI image.

Fig. 6.13: DTI-MR output volumes from the tumor simulator. The images correspond to a slice of the test case shown in Figure 6.12, generated with the parameters shown in Table 6.4. In both images can be appreciated the artificial-looking swelling of the tensors composing the initial seed, which does not appear delineated in the final ground truth (Fig. 6.14).



(a) Ground truth.



(b) Resulting segmentation.

Fig. 6.14: Result of the clustering procedure performed on a test case calculated with the software tumor simulator.

information conveyed by the MRI data. The use of the dissimilarity representation paradigm allows us to represent the data in vectorial spaces created according to their relationships expressed by a specific distance or dissimilarity function. The results show a consistent improvement of the assessment scores when using KPCA as a manifold learning technique, a consideration taken specifically to address the geometric structure of the tensorial diffusion data. This outcome, together with the results showing a marked increase in performance with the test case that employs the DT metrics using the full tensor information, shows the importance of including problem-specific knowledge in choosing the appropriate set of metrics or dissimilarity functions. A drawback of this method is the high computational cost of KPCA, since using the standard eigendecomposition of the kernel matrix involves a time complexity of $\mathcal{O}(n^3)$. However there are considerable advances on alternative methods for reducing the complexity of KPCA, with special focus on the factorization of the Gram matrix [195, 206]

As we have mentioned, the positive results in the tests warrant a further analysis using real pathological datasets, either from human patients or, as it is more feasible, using pre-clinical rodent models with inoculated tumoral cells. In any case, with this work we have demonstrated that the use of Consensus Clustering techniques in multi-modal medical image segmentations is a promising strategy for assessing the heterogeneity of tumoral regions.

7

Conclusions

Medical imaging is one of the most powerful tools for gaining insight into normal and pathological processes that affect health. The role of image processing in medicine is expanding with the increasing importance of finding ways to improve workflow in reading environments where more images are being acquired in more acquisition modalities. Image processing has an important influence on the medical decision making process and even on surgical actions. The performance of image processing methods may have an important impact on the performance of the larger systems as well as on the human observer that needs to analyze all of the available image data and render a diagnostic or therapeutic decision. An emerging focus is the development of imaging biomarkers for drug or therapy response, and the development and application of sophisticated image analysis methods in order to improve the accuracy of diagnosis, or to better predict outcomes of disease or treatment and intervention strategies.

Regarding diagnosis and treatment of tumors, the act of differential diagnosis applies a label to the tumor as a whole, but therapy outcomes are increasingly recognized to depend only on part on this label but appear to have correlations with such local properties as cellularity, vascularity and oxygen delivery. Biopsy, in addition to providing a definitive label, can provide indications of these local properties as well as histochemical profiles. It would be very, very costly however to attempt to use biopsy to achieve comprehensive coverage of tumors. MRI on the other hand is well suited to such coverage, and can provide indicators of some of the tissue properties of interest: cell density via diffusion imaging, and a mix of vascularity and vessel permeability via DCE-MRI.

In a conventional radiological report however, the results of the MRI-derived mappings are typically reduced to minimalist statements that leave important questions in the face of therapy. The present work is a contribution towards addressing the generalized clinical desire to integrate the different domains of multimodal MRI into a single coherent framework to aid in diagnosis and therapy.

The aim of this thesis was to investigate processing strategies for the combination of multi-modal MRI images. Mainly we made special emphasis in the combination of perfusion and diffusion MRI, considering and exploiting the particularities of each modality.

We initiated this thesis with the relevant background in medical imaging, particularly the two MRI modalities used through this work. Furthermore, continuing with the background information, we presented details on unsupervised classification, specially the interesting and sometimes overlooked problem of data representation.

Ultimately we reviewed the approaches to medical imaging fusion for ulterior discrimination, and proposed two main methodologies for the combination and usupervised classification of MRI mdoalities for heterogeneity assessment.

The use of the first methodology (Chapter 5) for heterogeneity quantification that integrates information from diffusion (an indicator of cellularity) and perfusion MRI images was illustrated in application to ductal carcinoma. The demonstration illustrated multimodal clustering leads to improved selectivity and yields a greater refinement of the segmentation of tissues within the lesion than the separate processing of the two modalities.

By demonstrating that statistically consistent subgroups can be defined within tumors based on a combination of DCE-MRI and DWI-MRI data, we have indicated a means for objectively segmenting tumors that can be used for larger studies to examine clinical impact. Moreover, the appearance of statistically distinct perfusion regions within the tumor at moderate and low ADC weightings that in turn have statistically distinct ADC distributions suggests there is a useable distinction present that is not capitalized upon in present clinical practice.

This methodology could also benefit from the incorporation of a cluster ensemble strategy, like the one we used in the last Chapter, to make a consensus between the different parameters that control the combination and weighting of diffusion and perfusion information to the overall dissimilarity score. Since the combination of modalities creates a vectorial "meta-space" in which both modalities are represented, the resulting partitioned zones should be carefully evaluated by the medical experts in order to compare them to a wide range of their histological and clinical assessments, not just a reductionist label as is commonly done with a regular biopsy.

The issue of a more conscientious validation still remain, specially from the clinical perspective. We are now looking into robust methods for further validation of the processing pipeline that would enable a clinical exploitation of the multimodal analysis. Access to ground truth beyond radiological and biopsy evaluation is needed and likely requires voxel-wise comparison of with histology of resections, a process that requires modifications to the surgical procedure that were not justified for this first demonstration and research of the method. Even were histology image data available, a significant task remains in the spatially correlation of individual MRI voxels with the histological results in order to get the requisite voxel-scale validation.

Although the first proposed methodology was applied to the specific problem of breast ductal carcinoma although it was formulated in a general and flexible way, with focus on the dimensionality and definition of suitable criteria for the representation of the multi-modal MRI data. There is nothing in the methodology that makes it exclusively suitable for breast cancer, on the contrary, its flexibility makes it appropriate for a straightforward implementation and tuning specific to a variety of different clinical domains.

The second methodology we presented in Chapter 6 relied on a "multi-view" approach to represent different aspects of each modality in a series of unique dissimilarity spaces. We exploited this set of vectorial spaces through a cluster ensemble strategy that served to reach a consensus between the different decisions and eventually an unifying partition.

Particularly, as we have focused on DTI-MR, an analysis of the peculiar characteristics of the modality highlights the importance of the geometric characteristics of the high-dimensional data. This is an issue that we have addressed with the integration of a kernel-based manifold learning technique. Results obtained with simulated data have demonstrated an improvement of the results using the manifold learning stage against the same methodology without it. However, just as with the first proposed approach, a further validation is required. In the case of the multi-view method the first step would be to make further analysis with a more complex set of simulated data, which would serve to analyze the specific limits and drawbacks of the method. As a second step, following the methodology of Chapter 5, an analysis with clinical data would be optimal. This of course raises new concerns, specifically the selection of appropriate cases and the reduction of variables with the selection of patients with the same pathology, something which is difficult to obtain when dealing with clinical medical data.

A common suggestion when dealing with validation is that of allowing an expert radiologist to delineate the regions of inhomogeneity and measure the amount of overlap with the results in terms of any specific validation index. However, the delineation by a radiologist is a highly non-trivial task. We have little hope that delineations by an expert radiologist will provide useful reference material however. Such an approach will require pixel by pixel decision-making on their part, and reflect entirely their bounds on defining similarity between enhancement patterns. We are looking instead to a histological approach that can be based on less biased analysis methods, or at least established criteria that will be suitable for an analysis of the type suggested by the reviewer. Access to ground truth beyond radiological and biopsy evaluation is needed and likely requires voxel-wise comparison of with histology of resections, a process that requires modifications to surgical procedures that were not justified for this first demonstration of the method. Either approach requires much greater funding and effort commitments than we have been able to obtain for a novel technique without prior validation.

The issue of multi-modal MRI data combination for classification purposes is still relatively new, we are sure this will become an buoyant and interesting research area as the MRI methods keep progressing towards higher resolutions. We hope to have made an interesting and relevant contribution to the development of this field with this thesis.

References

1. AL Alexander, JE Lee, Mariana Lazar, and AS Field. Diffusion tensor imaging of the brain. *Neurotherapeutics*, 4(3):316–329, 2007.
2. Anil K. Jain Alexander Topchy. A mixture model of clustering ensembles. *Proc. SIAM Intl. Conf. on Data Mining*, pages 379–390, 2004.
3. Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Fast and simple calculus on tensors in the log-Euclidean framework. *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, 8(Pt 1):115–22, January 2005.
4. Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Log-Euclidean metrics for fast and simple calculus on diffusion tensors. *Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine*, 56(2):411–21, August 2006.
5. Yaniv Assaf and Ofer Pasternak. Diffusion tensor imaging (DTI)-based white matter mapping in brain research: a review. *Journal of molecular neuroscience : MN*, 34(1):51–61, January 2008.
6. Hanan Ayad and Mohamed Kamel. Finding natural clusters using multi-clusterer combiner based on shared nearest neighbors. *Multiple Classifier Systems*, pages 166–175, 2003.
7. Hanan G Ayad and Mohamed S Kamel. Cumulative voting consensus method for partitions with variable number of clusters. *IEEE transactions on pattern analysis and machine intelligence*, 30(1):160–73, January 2008.
8. C Baratti, AS Barnett, and C Pierpaoli. Comparative mr imaging study of brain maturation in kittens with t1, t2, and the trace of the diffusion tensor. *Radiology*, 210(1):133, 1999.
9. J Barcelo, J Vilanova, Antonio Luna, Ramón Ribes, and Jorge a. Soto. Diffusion MRI Outside the Brain. chapter DWI of the, pages 203–230. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.
10. Angelos Barmpoutis, Bing Jian, and Baba C Vemuri. Adaptive kernels for multi-fiber reconstruction. *Information processing in medical imaging : proceedings of the ... conference*, 21:338–49, January 2009.
11. P. J. Basser, J. Mattiello, and D. LeBihan. Estimation of the effective self-diffusion tensor from the NMR spin echo. *Journal of magnetic resonance. Series B*, 103(3):247–254, March 1994.
12. P. J. Basser and C. Pierpaoli. Microstructural and physiological features of tissues elucidated by quantitative-diffusion-tensor MRI. *Journal of magnetic resonance. Series B*, 111(3):209–219, June 1996.

13. P.J. Basser, J. Mattiello, and D. LeBihan. Mr diffusion tensor spectroscopy and imaging. *Biophysical Journal*, 66(1):259 – 267, 1994.
14. P G Batchelor, M Moakher, D Atkinson, F Calamante, and A Connelly. A rigorous framework for diffusion tensor calculus. *Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine*, 53(1):221–5, January 2005.
15. Christian Beaulieu. The basis of anisotropic water diffusion in the nervous system – a technical review. *NMR in Biomedicine*, 15(7-8):435–455, 2002.
16. Christian Beaulieu. The basis of anisotropic water diffusion in the nervous system - a technical review. *NMR Biomed*, 15(7-8):435–455, Nov-Dec 2002.
17. E Belogay. Calculating the Hausdorff distance between curves. *Information Processing Letters*, 64(1):17–22, October 1997.
18. Asa Ben-Hur, David Horn, Hava T. Siegelmann, and Vladimir Vapnik. Support vector clustering. *The Journal of Machine Learning Research*, 2:125–137, March 2002.
19. D Le Bihan, E Breton, D Lallemand, P Grenier, E Cabanis, and M Laval-Jeantet. Mr imaging of intravoxel incoherent motions: application to diffusion and perfusion in neurologic disorders. *Radiology*, 161(2):401–407, 1986.
20. In The Human Brain, Sinisa Pajevic, and Carlo Pierpaoli. Color schemes to represent the orientation of anisotropic tissues from diffusion tensor data: Application to white matter fiber tract mapping. 2008.
21. Gunnar Brix, Fabian Kiessling, Robert Lucht, Susanne Darai, Klaus Wasser, Stefan Delorme, and Jürgen Griebel. Microcirculation and microvasculature in breast tumors: pharmacokinetic analysis of dynamic MR image series. *Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine*, 52(2):420–9, August 2004.
22. Gunnar Brix, Robert Lucht, and Jürgen Griebel. Tracer kinetic analysis of signal time series from dynamic contrast-enhanced MR imaging. *Biomedizinische Technik. Biomedical engineering*, 51(5-6):325–30, January 2006.
23. David L Buckley. Uncertainty in the analysis of tracer kinetics using dynamic contrast-enhanced T1-weighted MRI. *Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine*, 47(3):601–6, March 2002.
24. John C Carter, Diane C Lanham, Laurie E Cutting, Amy M Clements-Stephens, Xuejing Chen, Muhamed Hadzipasic, Joon Kim, Martha B Denckla, and Walter E Kaufmann. A dual dti approach to analyzing white matter in children with dyslexia. *Psychiatry Research: Neuroimaging*, 172(3):215–219, Jan 2009.
25. Isabelle Catalaa, Roland Henry, William P Dillon, Edward E Graves, Tracy R McKnight, Ying Lu, Daniel B Vigneron, and Sarah J Nelson. Perfusion, diffusion and spectroscopy values in newly diagnosed cerebral gliomas. *NMR in biomedicine*, 19(4):463–75, June 2006.
26. Jian Chen, Renée M McKay, and Luis F Parada. Malignant glioma: lessons from genomics, mouse models, and stem cells. *Cell*, 149(1):36–47, March 2012.
27. Weijie Chen, Maryellen L. Giger, Ulrich Bick, and Gillian M. Newstead. Automatic identification and classification of characteristic kinetic curves of breast lesions on DCE-MRI. *Medical Physics*, 33(8):2878, 2006.
28. Thomas L. Chenevert, Lauren D. Stegman, Jeremy M. G. Taylor, Patricia L. Robertson, Harry S. Greenberg, Alnawaz Rehemtulla, and Brian D. Ross. Diffusion magnetic resonance imaging: an early surrogate marker of therapeutic efficacy in brain tumors. 2000.
29. Ahlame Douzal Chouakria and Panduranga Naidu Nagabhushan. Adaptive dissimilarity index for measuring time series proximity. *Advances in Data Analysis and Classification*, 1(1):5–21, January 2007.

30. Peter L Choyke, Andrew J Dwyer, and Michael V Knopp. Functional tumor imaging with dynamic contrast-enhanced magnetic resonance imaging. *Journal of magnetic resonance imaging : JMRI*, 17(5):509–20, May 2003.
31. O. Ciccarelli, T. E. Behrens, D. R. Altmann, R. W. Orrell, R. S. Howard, H. Johansen-Berg, D. H. Miller, P. M. Matthews, and A. J. Thompson. Probabilistic diffusion tractography: a potential tool to assess the rate of disease progression in amyotrophic lateral sclerosis. 2006.
32. W R Crum. Non-rigid image registration: theory and practice. *British Journal of Radiology*, 77(suppl_2):S140–S153, December 2004.
33. R. Cummins. *Meaning and Mental Representation*. MIT Press, 1991.
34. David L. Davies and Donald W. Bouldin. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, April 1979.
35. Hadassa Degani, Vadim Gusis, Daphna Weinstein, Scott Fields, and Shalom Strano. Mapping pathophysiological features of breast tumors by MRI at high spatial resolution. *Nature Medicine*, 3(7):780–782, July 1997.
36. K. I. Diamantaras and S. Y. Kung. *Principal Component Neural Networks: Theory and Applications (Adaptive and Learning Systems for Signal Processing, Communications and Control Series)*. Wiley-Interscience, 1996.
37. M Doran, J V Hajnal, N Van Bruggen, M D King, I R Young, and G M Bydder. Normal and abnormal white matter tracts shown by mr imaging using directional diffusion weighted sequences. *J Comput Assist Tomogr*, 14(6):865–873, Nov-Dec 1990.
38. Ahlame Douzal-Chouakria and Cecile Amblard. Classification trees for time series. *Pattern Recognition*, August 2011.
39. M.-P. Dubuisson and A.K. Jain. A modified Hausdorff distance for object matching. *Proceedings of 12th International Conference on Pattern Recognition*, pages 566–568, 1994.
40. Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2nd editio edition, 2000.
41. Robert P.W. Duin and Elbieta Pekalska. The dissimilarity space: Bridging structural and statistical pattern recognition. *Pattern Recognition Letters*, 33(7):826–832, May 2012.
42. M Dujardin, S Sourbron, R Luypaert, D Verbeelen, and T Stadnik. Quantification of renal perfusion and function on a voxel-by-voxel basis: a feasibility study. *Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine*, 54(4):841–9, October 2005.
43. J. C. Dunn. Well-Separated Clusters and Optimal Fuzzy Partitions. *Journal of Cybernetics*, 4(1):95–104, January 1974.
44. Shimon Edelman. *Representation and Recognition in Vision (Bradford Books)*. A Bradford Book, 1999.
45. A. Einstein, R. F
"urth, and A.D. Cowper. *Investigations on the theory of the Brownian movement*. Dover books on physics. Dover Publications, 1956.
46. V. Estivill-Castro and I. Lee. Amoeba: Hierarchical clustering based on spatial proximity using delaunay diagram. In *In Proc. of the 9th Int. Symposium on Spatial Data Handling*, pages 7a.26–7a.41, 2000.
47. V. Estivill-Castro and I. Lee. Criteria on proximity graphs for boundary extraction and spatial clustering. *Advances in Knowledge Discovery and Data Mining*, pages 348–357, 2001.
48. Brian Everitt and Torsten Hothorn. Cluster Analysis. In *An Introduction to Applied Multivariate Analysis with R Use R*, pages 163–200. Springer New York, New York, NY, 2011.

49. Xiaobing Fan, Milica Medved, Gregory S Karczmar, Cheng Yang, Sean Foxley, Sanaz Arkani, Wendy Recant, Marta a Zamora, Hiroyuki Abe, and Gillian M Newstead. Diagnosis of suspicious breast lesions using an empirical mathematical model for dynamic contrast-enhanced MRI. *Magnetic resonance imaging*, 25(5):593–603, June 2007.
50. Xiaoli Zhang Fern and Carla E. Brodley. Solving cluster ensemble problems by bipartite graph partitioning. *Twenty-first international conference on Machine learning - ICML '04*, page 36, 2004.
51. XZ Fern and CE Brodley. Random projection for high dimensional data clustering: A cluster ensemble approach. *MACHINE LEARNING-INTERNATIONAL ...*, 2003.
52. P Fletcher and Sarang Joshi. Principal geodesic analysis on symmetric spaces: Statistics of diffusion tensors. *... Methods in Medical and Biomedical Image Analysis*, pages 87–98, 2004.
53. R Fletcher. *Practical methods of optimization*. Wiley, 1987.
54. P. Franti. A heuristic k-means clustering algorithm by kernel pca. In *2004 International Conference on Image Processing, 2004. ICIP '04.*, volume 5, pages 3503–3506. IEEE.
55. Ana L N Fred and Anil K Jain. Combining multiple clusterings using evidence accumulation. *IEEE transactions on pattern analysis and machine intelligence*, 27(6):835–50, June 2005.
56. Brendan J Frey and Delbert Dueck. Clustering by passing messages between data points. *Science (New York, N.Y.)*, 315(5814):972–6, February 2007.
57. Yaniv Gal, Andrew Mehnert, Andrew Bradley, Dominic Kennedy, and Stuart Crozier. *Feature and Classifier Selection for Automatic Classification of Lesions in Dynamic Contrast-Enhanced MRI of the Breast*. IEEE, December 2009.
58. Yaniv Gal, Andrew Mehnert, Andrew Bradley, Kerry McMahon, and Stuart Crozier. An evaluation of four parametric models of contrast enhancement for dynamic magnetic resonance imaging of the breast. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, 2007:71–4, January 2007.
59. Joydeep Ghosh and Ayan Acharya. Cluster ensembles. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(4):305–315, July 2011.
60. Peter GIBBS, Gary P. LINEY, Martin D. PICKLES, Bashar ZELHOF, Greta RODRIGUES, and Lindsay W. TURNBULL. Correlation of ADC and T2 Measurements With Cell Density in Prostate Cancer at 3.0 Tesla. *Investigative radiology*, 44(9):572–576.
61. Mark Girolami. Orthogonal series density estimation and the kernel eigenvalue problem. *Neural Computation*, 688:669–688, 2002.
62. Sylvia Glaß er, Uta Preim, Klaus Tönnies, and Bernhard Preim. A visual analytics approach to diagnosis of breast DCE-MRI data. *Computers & Graphics*, 34(5):602–611, October 2010.
63. Andrey Goder and Vladimir Filkov. Consensus clustering algorithms: Comparison and refinement. *ALENEX'08: Procs. 10th Workshop on Algorithm ...*, 2008.
64. L. Goldfarb. A new approach to pattern recognition. *Progress in pattern recognition*, 2:241–402, 1985.
65. Fabio A. Gonzalez and Eduardo Romero. *Biomedical Image Analysis and Machine Learning Technologies: Applications and Techniques*. Medical Information Science Reference, 2009.
66. David M Greer, Walter J Koroshetz, Sean Cullen, R Gilberto Gonzalez, and Michael H Lev. Magnetic resonance imaging improves detection of intracerebral hemorrhage over computed tomography after intra-arterial thrombolysis. *Stroke*, 35(2):491–495, Feb 2004.

67. F. Grinberg, P. Heitjans, and T. Ito. *Diffusion Fundamentals Leipzig 2005*. Leipziger Universit
"atsverlag, 2005.
68. Yujun Guo, Radhika Sivaramakrishna, Cheng-Chang Lu, Jasjit S Suri, and Swamy Laxminarayan. Breast image registration techniques: a survey. *Medical & biological engineering & computing*, 44(1-2):15–26, March 2006.
69. Stefan T. Hadjitodorov, Ludmila I. Kuncheva, and Ludmila P. Todorova. Moderate diversity for better cluster ensembles. *Information Fusion*, 7(3):264–275, September 2006.
70. John R. Hansen. Pulsed nmr study of water mobility in muscle and brain tissue. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 230(3):482 – 486, 1971.
71. P Hao, J Chiang, and Y Tu. Hierarchically SVM classification based on support vector clustering method and its application to document categorization. *Expert Systems with Applications*, 33(3):627–635, October 2007.
72. Carmel Hayes, Anwar R Padhani, and Martin O Leach. Assessing changes in tumour vascular function using dynamic contrast-enhanced magnetic resonance imaging. *NMR in biomedicine*, 15(2):154–63, April 2002.
73. E Henderson, BK Rutt, and TY Lee. Temporal sampling requirements for the tracer kinetics modeling of breast disease. *Magnetic resonance imaging*, 16(9):1057–1073, 1998.
74. Heiko Hoffmann. *Unsupervised Learning of Visuomotor Associations*. PhD thesis, Universität Bielefeld,, 2005.
75. Lawrence Hubert and P Arabie. Comparing partitions. *Journal of classification*, 218:193–218, 1985.
76. PS Hüppi, SE MAIER, S PELED, GP ZIENTARA, PD BARNES, FA JOLESZ, and JJ VOLPE. Microstructural development of human newborn cerebral white matter assessed in vivo by diffusion tensor magnetic resonance imaging. *Pediatric Research*, 44(4):584, 1998.
77. D.P. Huttenlocher, G.a. Klanderman, and W.J. Rucklidge. Comparing images using the Hausdorff distance, 1993.
78. Nola Hylton. Dynamic contrast-enhanced magnetic resonance imaging as an imaging biomarker. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 24(20):3293–8, July 2006.
79. N Iam-On, Tossapon Boongoen, and Simon Garrett. Refining pairwise similarity matrix for cluster ensemble problem with cluster relations. *Discovery Science. Lecture Notes in Computer Science.*, pages 222–233, 2008.
80. N Iam-on and Simon Garrett. LinkCluE: A MATLAB Package for Link-Based Cluster Ensembles. *Journal of Statistical Software*, 36(9), 2010.
81. T Inoue, K Ogasawara, T Beppu, A Ogawa, and H Kabasawa. Diffusion tensor imaging for preoperative evaluation of tumor grade in gliomas. *Clinical neurology and neurosurgery*, 107(3):174–180, 2005.
82. Takashi Inoue, Kuniaki Ogasawara, Takaaki Beppu, Akira Ogawa, and Hiroyuki Kabasawa. Diffusion tensor imaging for preoperative evaluation of tumor grade in gliomas. *Clinical neurology and neurosurgery*, 107(3):174–80, May 2005.
83. A Jain and M Law. Data clustering: A user's dilemma. *Pattern Recognition and Machine Intelligence*, pages 1–10, 2005.
84. AK Jain, MN Murty, and PJ Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3), 1999.
85. Anil K. Jain. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8):651–666, June 2010.
86. Anil K. Jain and Richard C. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, 1988.

87. Michael D Jenkinson, Daniel G du Plessis, Trevor S Smith, Andrew R Brodbelt, Kathy A Joyce, and Carol Walker. Cellularity and apparent diffusion coefficient in oligodendroglial tumours characterized by genotype. *Journal of neuro-oncology*, 96(3):385–92, February 2010.
88. Line R Jensen, Benjamin Garzon, Mariann G Heldahl, Tone F Bathen, Steinar Lundgren, and Ingrid S Gribbestad. Diffusion-weighted and dynamic contrast-enhanced MRI in evaluation of early treatment effects during neoadjuvant chemotherapy in breast cancer patients. *Journal of magnetic resonance imaging : JMRI*, 34(5):1099–109, November 2011.
89. H. Johansen-Berg and T.E.J. Behrens. *Diffusion MRI: from quantitative measurement to in-vivo neuroanatomy*. Academic Press. Elsevier/Academic Press, 2009.
90. Kasper W. Jorgensen and Lars Kai Hansen. Model Selection for Gaussian Kernel PCA Denoising. *IEEE transactions on Neural Networks and Learning Systems,*, 23(1):163–168, 2012.
91. George Karypis, Rajat Aggarwal, Vipin Kumar, and Shashi Shekhar. Multilevel hypergraph partitioning. In *Proceedings of the 34th annual conference on Design automation conference - DAC '97*, pages 526–529, New York, New York, USA, June 1997. ACM Press.
92. George Karypis and Vipin Kumar. A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs. *SIAM Journal on Scientific Computing*, 20(1):359–392, January 1998.
93. P. Khurd, R. Verma, and C. Davatzikos. On Characterizing and Analyzing Diffusion Tensor Images by Learning their Underlying Manifold Structure. *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, pages 61–61, 2006.
94. Parmeshwar Khurd, Ragini Verma, and Christos Davatzikos. Kernel-based manifold learning for statistical analysis of diffusion tensor images. *Information processing in medical imaging : proceedings of the ... conference*, 20:581–93, January 2007.
95. J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, March 1998.
96. Arno Klein, Satrajit S Ghosh, Brian Avants, B T T Yeo, Bruce Fischl, Babak Ardekani, James C Gee, J J Mann, and Ramin V Parsey. Evaluation of volume-based and surface-based brain image registration methods. *NeuroImage*, 51(1):214–20, May 2010.
97. Stefan Klink, Patrick Reuther, and Alexander Weber. Analysing social networks within bibliographical data. *Database and Expert ...*, pages 234–243, 2006.
98. Michael V. Knopp, Frederik L. Giesel, and Hani Marcos. Dynamic Contrast Enhanced Magnetic Resonance Imaging in Oncology: Theory, Data Acquisition, Analysis, and Examples. *Topics in Magnetic Resonance Imaging*, 12(4):301–308, May 2001.
99. C. K. Kuhl. MRI of breast tumors, 2000.
100. C K Kuhl, P Mielcareck, S Klaschik, C Leutner, E Wardelmann, J Gieseke, and H H Schild. Dynamic breast MR imaging: are signal intensity time course data useful for differential diagnosis of enhancing lesions? *Radiology*, 211(1):101–10, April 1999.
101. Ritwik Kumar and Angelos Barmpoutis. A Physical basis for multi-fiber reconstruction from DW-MRI data. *... Imaging: From Nano ...*, pages 2–5, 2009.
102. L.I. Kuncheva and S.T. Hadjitodorov. Using diversity in cluster ensembles. *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)*, 2:1214–1219, 2004.
103. K A Kvistad, J Rydland, J Vainio, H B Smethurst, S Lundgren, H E Fjø sne, and O Haraldseth. Breast lesions: evaluation with dynamic contrast-enhanced T1-

weighted MR imaging and with T2*-weighted first-pass perfusion MR imaging. *Radiology*, 216(2):545–53, August 2000.

104. Cristina Lavini, Milko C de Jonge, Marleen G H van De Sande, Paul P Tak, Aart J Nederveen, and Mario Maas. Pixel-by-pixel analysis of DCE MRI curve patterns and an illustration of its application to the imaging of the musculoskeletal system. *Magnetic resonance imaging*, 25(5):604–12, June 2007.
105. Cristina Lavini, Branko P Pikaart, Milko C de Jonge, Gerard R Schaap, and Mario Maas. Region of interest and pixel-by-pixel analysis of dynamic contrast enhanced magnetic resonance imaging parameters and time-intensity curve shapes: a comparison in chondroid tumors. *Magnetic resonance imaging*, 27(1):62–8, January 2009.
106. Mariana Lazar, David M. Weinstein, Jay S. Tsuruda, Khader M. Hasan, Konstantinos Arfanakis, M. Elizabeth Meyerand, Benham Badie, Howard A. Rowley, Victor Haughton, Aaron Field, and Andrew L. Alexander. White matter tractography using diffusion tensor deflection. *Human Brain Mapping*, 18(4):306–321, 2003.
107. D Le Bihan, E Breton, D Lallemand, M L Aubin, J Vignaud, and M Laval-Jeantet. Separation of diffusion and perfusion in intravoxel incoherent motion MR imaging. *Radiology*, 168(2):497–505, August 1988.
108. Denis Le Bihan, Cyril Poupon, Alexis Amadon, and Franck Lethimonnier. Artifacts and pitfalls in diffusion mri. *Journal of Magnetic Resonance Imaging*, 24(3):478–488, 2006.
109. George Lee, Scott Doyle, Michael D Feldman, R Stephen, and John E Tomaszewski. A knowledge representation framework for integration, classification of multi-scale imaging and non-imaging data: Preliminary results in predicting prostate cancer. *Biomedical Imaging: ...*, pages 77–80, 2009.
110. Jianchao Liang. *Dynamic contrast enhanced MRI at high and ultra high fields*. PhD thesis, Ohio State University, 2008.
111. Gilad Liberman, Yoram Louzoun, Orna Aizenstein, Deborah T Blumenthal, Felix Bokstein, Mika Palmon, Benjamin W Corn, and Dafna Ben Bashat. Automatic multi-modal MR tissue classification for the assessment of response to bevacizumab in patients with glioblastoma. *European journal of radiology*, null(null), September 2012.
112. Laura Liberman and Jennifer H. Menell. Breast imaging reporting and data system (BI-RADS). *Radiologic Clinics of North America*, 40(3):409–430, May 2002.
113. G.J.S. Litjens, M. Heisen, J. Buurman, and B.M. ter Haar Romeny. *Pharmacokinetic models in clinical practice: What model to use for DCE-MRI of the breast?* IEEE, April 2010.
114. A Y Liu, J A Maldjian, L J Bagley, G P Sinson, and R I Grossman. Traumatic brain injury: diffusion-weighted mr imaging findings. *AJNR Am J Neuroradiol*, 20(9):1636–1641, Oct 1999.
115. Mark C Lloyd, Pushpa Allam-Nandyala, Chetna N Purohit, Nancy Burke, Domenico Coppola, and Marilyn M Bui. Using image analysis as a tool for assessment of prognostic and predictive biomarkers for breast cancer: How reliable is it? *Journal of pathology informatics*, 1:29, January 2010.
116. Richard G P Lopata, Walter H Backes, Paul P J van Den Bosch, and Natal A W van Riel. On the identifiability of pharmacokinetic parameters in dynamic contrast-enhanced imaging. *Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine*, 58(2):425–9, August 2007.
117. Stanley Lu, Daniel Ahn, Glyn Johnson, and Soonmee Cha. Peritumoral diffusion tensor imaging of high-grade gliomas and metastatic brain tumors. *American Journal of ...*, (May):937–941, 2003.

118. Stanley Lu, Daniel Ahn, Glyn Johnson, Meng Law, David Zagzag, and Robert I Grossman. Diffusion-tensor MR imaging of intracranial neoplasia and associated peritumoral edema: introduction of the tumor infiltration index. *Radiology*, 232(1):221–8, July 2004.
119. Anant Madabhushi, Shannon Agner, Ajay Basavanhally, Scott Doyle, and George Lee. Computer-aided prognosis: predicting patient and disease outcome via quantitative fusion of multi-scale, multi-modal data. *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society*, 35(7-8):506–14, 2011.
120. M Maeda, Y Kawamura, Y Tamagawa, T Matsuda, S Itoh, H Kimura, T Iwasaki, N Hayashi, K Yamamoto, and Y Ishii. Intravoxel incoherent motion (ivim) mri in intracranial, extraaxial tumors and cysts. *J Comput Assist Tomogr*, 16(4):514–518, Jul-Aug 1992.
121. Yoshitsugu Matsumoto, Masahiro Kuroda, Ryohei Matsuya, Hirokazu Kato, Koichi Shibuya, Masataka Oita, Atsushi Kawabe, Hidenobu Matsuzaki, Junichi Asaumi, Jun Murakami, Kazunori Katashima, Masakazu Ashida, Takanori Sasaki, Tetsuro Sei, Susumu Kanazawa, Seiichi Mimura, Seiichiro Oono, Takuichi Kitayama, Seiji Tahara, and Keiji Inamura. In vitro experimental study of the relationship between the apparent diffusion coefficient and changes in cellularity and cell morphology. *Oncology reports*, 22(3):641–8, September 2009.
122. Kathryn M. McMillan, Baxter P. Rogers, Cheng Guan Koay, Angela R. Laird, Ronald R. Price, and M. Elizabeth Meyerand. An objective method for combining multi-parametric MRI datasets to characterize malignant tumors. *Medical Physics*, 34(3):1053, 2007.
123. C. Andrés Mendez, F Pizzorni Ferrarese, Paul Summers, Giuseppe Petralia, and Gloria Menegaz. Multimodal MRI-based tissue classification in breast ductal carcinoma. In *In Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on*, pages 142–145, 2012.
124. C Andrés Méndez, Francesca Pizzorni Ferrarese, Paul Summers, Giuseppe Petralia, and Gloria Menegaz. DCE-MRI and DWI Integration for Breast Lesions Assessment and Heterogeneity Quantification. *International journal of biomedical imaging*, 2012:676808, January 2012.
125. Peter J Moate, Lawrence Dougherty, Mitchell D Schnall, Richard J Landis, and Raymond C Boston. A modified logistic model to describe gadolinium kinetics in breast tumors. *Magnetic resonance imaging*, 22(4):467–73, May 2004.
126. S Monti, P Tamayo, J Mesirov, and T Golub. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning*, (i):91–118, 2003.
127. Ken-ichi Morita, Hitoshi Matsuzawa, Yukihiko Fujii, Ryuichi Tanaka, Ingrid L Kwee, and Tsutomu Nakada. Diffusion tensor analysis of peritumoral edema using lambda chart analysis indicative of the heterogeneity of the microstructure within edema. *Journal of neurosurgery*, 102(2):336–41, February 2005.
128. M E Moseley, Y Cohen, J Kucharczyk, J Mintorovitch, H S Asgari, M F Wendland, J Tsuruda, and D Norman. Diffusion-weighted MR imaging of anisotropic water diffusion in cat central nervous system. *Radiology*, 176(2):439–445, 1990.
129. M. E. Moseley, Y. Cohen, J. Mintorovitch, L. Chileuitt, H. Shimizu, J. Kucharczyk, M. F. Wendland, and P. R. Weinstein. Early detection of regional cerebral ischemia in cats: Comparison of diffusion- and t2-weighted mri and spectroscopy. *Magnetic Resonance in Medicine*, 14(2):330–346, 1990.
130. Stavros Mussurakis, David L Buckley, and Anthony Horsman. Dynamic MRI of Invasive Breast Cancer: Assessment of Three Region-of-Interest Analysis Methods. *Journal of Computer Assisted Tomography*, 21(3), 1997.

131. Morton Nadler and Eric P. Smith. *Pattern Recognition Engineering*. Wiley-Interscience, 1993.
132. Alissar Nasser, Denis Hamad, and Chaiban Nasr. K-means Clustering Algorithm in Projected Spaces. In *2006 9th International Conference on Information Fusion*, pages 1–6. IEEE, July 2006.
133. Nam Nguyen and Rich Caruana. Consensus Clusterings. *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 607–612, October 2007.
134. S Ono, K Takahashi, Y Fukuoka, K Jinnai, F Kanda, H Kurisaki, and S Mitake. Intracytoplasmic inclusion bodies of the substantia nigra in myotonic dystrophy:: Immunohistochemical observations. *Journal of the neurological sciences*, 148(2):193–198, 1997.
135. Savannah C. PARTRIDGE, Wendy B. DEMARTINI, Brenda F. KURLAND, Peter R. EBY, Steven W. WHITE, and Constance D. LEHMAN. Quantitative Diffusion-Weighted Imaging as an Adjunct to Conventional Breast MRI for Improved Positive Predictive Value. *American journal of roentgenology*, 193(6):1716–1722.
136. Ofer Pasternak, Ragini Verma, Nir Sochen, and PJ Basser. On what manifold do diffusion tensors live. *MICCAI Workshop*, 2008.
137. Rodrigues Paulo R. *Homogeneity based segmentation and enhancement of Diffusion Tensor Images - A white matter processing framework*. PhD thesis, Eindhoven University of Technology, 2011.
138. Robia G. Pautler, Afonso C. Silva, and Alan P. Koretsky. In vivo neuronal tract tracing using manganese-enhanced magnetic resonance imaging. *Magnetic Resonance in Medicine*, 40(5):740–748, 1998.
139. T.H.J.M. Peeters, P.R. Rodrigues, A. Vilanova, and B.M. ter Haar Romeny. Analysis of Distance/Similarity Measures for Diffusion Tensor Imaging. In *Visualization and Processing of Tensor Fields*, pages 113–138. 2009.
140. E. Pekalska and R.P.W. Duin. Beyond Traditional Kernels: Classification in Two Dissimilarity-Based Representation Spaces. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(6):729–744, November 2008.
141. E. Pekalska, P. Paclik, and R.P.W. Duin. A generalized kernel approach to dissimilarity-based classification. *The Journal of Machine Learning Research*, 2:211, 2002.
142. Elzbieta Pekalska and Robert P. W. Duin. *The dissimilarity representation for pattern recognition: foundations and ...* World Scientific, 2005.
143. Elzbieta Pekalska, Robert P.W. Duin, and P. Paclik. Prototype selection for dissimilarity-based classifiers. *Pattern Recognition*, 39(2):189–208, February 2006.
144. Dan Pelleng and Andrew W. Moore. x-means: extending k-means with efficient estimation of the number of clusters. In *ICML '00 Proceedings of the Seventeenth International Conference on Machine Learning*, pages 727–734, 2000.
145. C Pierpaoli and PJ Basser. Toward a quantitative assessment of diffusion anisotropy. *Magnetic Resonance in Medicine*, 36(6):893–906, 1996.
146. C. Pierpaoli, P. Jezzard, P. J. Basser, A. Barnett, and G. Di Chiro. Diffusion tensor MR imaging of the human brain. *Radiology*, 201(3):637–648, December 1996.
147. Josien P W Pluim, J B Antoine Maintz, and Max a Viergever. Mutual-information-based registration of medical images: a survey. *IEEE transactions on medical imaging*, 22(8):986–1004, August 2003.
148. Marcel Prastawa, Elizabeth Bullitt, and Guido Gerig. Simulation of brain tumors in MR images for evaluation of segmentation efficacy. *Medical image analysis*, 13(2):297–311, April 2009.

149. S J Price, R Jena, N G Burnet, P J Hutchinson, a F Dean, a Peña, J D Pickard, T a Carpenter, and J H Gillard. Improved delineation of glioma margins and regions of infiltration with the use of diffusion tensor imaging: an image-guided biopsy study. *AJNR. American journal of neuroradiology*, 27(9):1969–74, October 2006.
150. Stephen J Price, Alonso Peña, Neil G Burnet, Raj Jena, Hadrian a L Green, T Adrian Carpenter, John D Pickard, and Jonathan H Gillard. Tissue signature characterisation of diffusion tensor abnormalities in cerebral gliomas. *European radiology*, 14(10):1909–17, October 2004.
151. Dustin K. Ragan. *Measurement of the vascular input function in mice for DCE-MRI*. PhD thesis, University of Texas, 2010.
152. William M. Rand. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66(336):846–850, December 1971.
153. TPL Roberts and Andrea Kassner. Imaging Tumor Biology. *High-Grade Gliomas*, pages 141–159, 2007.
154. T Rohlfing, a Pfefferbaum, E V Sullivan, and C R Maurer. Information fusion in biomedical image analysis: combination of data vs. combination of interpretations. *Information processing in medical imaging : proceedings of the ... conference*, 19:150–61, January 2005.
155. P Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(1):53–65, 1987.
156. D Rueckert, L I Sonoda, C Hayes, D L Hill, M O Leach, and D J Hawkes. Nonrigid registration using free-form deformations: application to breast MR images. *IEEE transactions on medical imaging*, 18(8):712–21, August 1999.
157. Daniel Rueckert and Paul Aljabar. Nonrigid Registration of Medical Images: Theory, Methods, and Applications [Applications Corner. *IEEE Signal Processing Magazine*, 27(4):113–119, July 2010.
158. Arthur Schatzkin and Mitchell Gail. The promise and peril of surrogate end points in cancer research. *Nature reviews. Cancer*, 2(1):19–27, January 2002.
159. Volker J Schmid, Brandon Whitcher, Anwar R Padhani, N Jane Taylor, and Guang-Zhong Yang. Bayesian methods for pharmacokinetic models in dynamic contrast-enhanced magnetic resonance imaging. *IEEE transactions on medical imaging*, 25(12):1627–36, December 2006.
160. Volker J Schmid, Brandon Whitcher, Anwar R Padhani, and Guang-Zhong Yang. Quantitative analysis of dynamic contrast-enhanced MR images based on Bayesian P-splines. *IEEE transactions on medical imaging*, 28(6):789–98, June 2009.
161. Bernhard Schölkopf and Alexander J Smola. *Learning with kernels*. Cambridge, MA : MIT,, 2001.
162. L H Schwamm, W J Koroshetz, A G Sorensen, B Wang, W A Copen, R Budzik, G Rordorf, F S Buonanno, P W Schaefer, and R G Gonzalez. Time course of lesion development in patients with acute stroke: serial diffusion- and hemodynamic-weighted magnetic resonance imaging. *Stroke*, 29(11):2268–2276, Nov 1998.
163. Armin Schwartzman. *Random ellipsoids and false discovery rates: statistics for diffusion tensor imaging data*. PhD thesis, Stanford University, 2006.
164. K Shanmuganathan, RP Gullapalli, J Zhuo, and SE Mirvis. Diffusion tensor mr imaging in cervical spine trauma. *American Journal of Neuroradiology*, 29(4):655, 2008.
165. V Sharifi-Salamatian, B Pesquet-Popescu, J Simony-Lafontaine, and J P Rigaut. Index for spatial heterogeneity in breast cancer. *Journal of microscopy*, 216(Pt 2):110–22, November 2004.
166. Saurabh Sinha, Mark E Bastin, Ian R Whittle, and Joanna M Wardlaw. Diffusion tensor MR imaging of high-grade cerebral gliomas. *AJNR. American journal of neuroradiology*, 23(4):520–7, April 2002.

167. LR Skelly, V Calhoun, SA Meda, J Kim, DH Mathalon, and GD Pearlson. Diffusion tensor imaging in schizophrenia: Relationship to symptoms. *Schizophrenia Research*, 2007.
168. John J Smith, A Gregory Sorensen, and James H Thrall. Biomarkers in imaging: realizing radiology's future. *Radiology*, 227(3):633–8, June 2003.
169. OLLE SÖDERMAN and BENGT JÖNSSON. Restricted Diffusion in Cylindrical Geometry. *Journal of Magnetic Resonance, Series A*, 117(1):94–97, November 1995.
170. Steven Sourbron. Technical aspects of MR perfusion. *European journal of radiology*, April 2010.
171. Olaf Sporns, Giulio Tononi, and Rolf Kötter. The human connectome: A structural description of the human brain. *PLoS Comput Biol*, 1(4):e42, 09 2005.
172. Rujirutana Srikanchana, David Thomasson, Peter Choyke, and Andrew Dwyer. A Comparison of Pharmacokinetic Models of Dynamic Contrast Enhanced MRI. *Proceedings of the 17th IEEE Symposium on Computer-Based Medical Systems (CBMS04)*, pages 0–5, 2004.
173. E. O. Stejskal and J. E. Tanner. Spin diffusion measurements: spin echoes in the presence of a time dependent field gradient. *J Chem Phys*, 42(1):288–92, 1964.
174. Alexander Strehl and J Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research*, 3:583–617, 2003.
175. Min-Ying Su, Yun-Chung Cheung, John P Fruehauf, Hon Yu, Orhan Nalcioglu, Eugene Mechetner, Ainura Kyshtoobayeva, Shin-Cheh Chen, Swei Hsueh, Christine E McLaren, and Yung-Liang Wan. Correlation of dynamic contrast enhancement MRI parameters with microvessel density and VEGF for assessment of angiogenesis in breast cancer. *Journal of magnetic resonance imaging : JMRI*, 18(4):467–77, October 2003.
176. P C Sundgren, Q Dong, D Gómez-Hassan, S K Mukherji, P Maly, and R Welsh. Diffusion tensor imaging of the brain: review of clinical applications. *Neuroradiology*, 46(5):339–50, May 2004.
177. Stephen Swift, Allan Tucker, Veronica Vinciotti, Nigel Martin, Christine Orengo, Xiaohui Liu, and Paul Kellam. Consensus clustering and functional interpretation of gene-expression data. *Genome biology*, 5(11):R94, January 2004.
178. J Taylor. Evolution from empirical dynamic contrast-enhanced magnetic resonance imaging to pharmacokinetic MRI. *Advanced Drug Delivery Reviews*, 41(1):91–110, March 2000.
179. Sergios Theodoridis and Konstantinos Koutroumbas. *Pattern Recognition, Fourth Edition*. Academic Press, 2008.
180. PM Thompson, RA Dutton, KM Hayashi, A Lu, SE Lee, JY Lee, OL Lopez, HJ Aizenstein, AW Toga, and JT Becker. 3d mapping of ventricular and corpus callosum abnormalities in hiv/aids. *NeuroImage*, 31(1):12–23, 2006.
181. C Thomsen, O Henriksen, and P Ring. In vivo measurement of water self diffusion in the human brain by magnetic resonance imaging. *Acta Radiol*, 28(3):353–361, May-Jun 1987.
182. Pallavi Tiwari, Satish Viswanath, George Lee, and Anant Madabhushi. Multimodal data fusion schemes for integrated classification of imaging and non-imaging biomedical data. In *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 165–168. IEEE, March 2011.
183. Paul S. Tofts, Gunnar Brix, David L. Buckley, Jeffrey L. Evelhoch, Elizabeth Henderson, Michael V. Knopp, Henrik B.W. Larsson, Ting-Yim Lee, Nina A. Mayr, Geoffrey J.M. Parker, Ruediger E. Port, June Taylor, and Robert M. Weisskoff. Estimating kinetic parameters from dynamic contrast-enhanced t1-weighted MRI of a diffusable tracer: Standardized quantities and symbols. *Journal of Magnetic Resonance Imaging*, 10(3):223–232, September 1999.

184. Paul S. Tofts and Allan G. Kermode. Measurement of the blood-brain barrier permeability and leakage space using dynamic MR imaging. 1. Fundamental concepts. *Magnetic Resonance in Medicine*, 17(2):357–367, February 1991.
185. Nermin Tuncbilek, Hakki Muammer Karakas, and Ozerk Omur Okten. Dynamic contrast enhanced MRI in the differential diagnosis of soft tissue tumors. *European journal of radiology*, 53(3):500–5, March 2005.
186. R Turner, D Le Bihan, and A Chesnick.... Echo-planar imaging of diffusion and perfusion. *Magn Reson Med*, Jan 1991.
187. Ragini Verma, Parmeshwar Khurd, and Christos Davatzikos. On analyzing diffusion tensor images by identifying manifold structure using isomaps. *IEEE transactions on medical imaging*, 26(6):772–8, June 2007.
188. Hongjun Wang, Hanhuai Shan, and Arindam Banerjee. Bayesian cluster ensembles. *Statistical Analysis and Data Mining*, 4(1):54–70, February 2011.
189. Zhizhou Wang and Baba C Vemuri. DTI segmentation using an information theoretic tensor dissimilarity measure. *IEEE transactions on medical imaging*, 24(10):1267–77, October 2005.
190. ZJ Wang, Z Han, KJR Liu, and Y Wang. Simultaneous estimation of kinetic parameters and the input function from DCE-MRI data: theory and simulation. *Symposium on Biomedical Imaging: Nano to*, (i):996–999, 2004.
191. M L White, Y Zhang, F Yu, and S a Jaffar Kazmi. Diffusion tensor MR imaging of cerebral gliomas: evaluating fractional anisotropy characteristics. *AJNR. American journal of neuroradiology*, 32(2):374–81, March 2011.
192. UC Wieshmann and CA Clark. Reduced anisotropy of water diffusion in structural cerebral abnormalities demonstrated with diffusion tensor imaging. *Magnetic resonance imaging*, 17(9):1269–1274, 1999.
193. Rui Xu and Don Wunsch. *Clustering*, volume 2008. John Wiley & Sons, 2008.
194. Rui Xu and Donald Wunsch. Survey of clustering algorithms. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, 16(3):645–78, May 2005.
195. Yong Xu, David Zhang, Fengxi Song, Jing-Yu Yang, Zhong Jing, and Miao Li. A method for speeding up feature extraction based on KPCA. *Neurocomputing*, 70(4-6):1056–1061, January 2007.
196. Hidetake Yabuuchi, Yoshio Matsuo, Takashi Okafuji, Takeshi Kamitani, Hiroyasu Soeda, Taro Setoguchi, Shuji Sakai, Masamitsu Hatakenaka, Makoto Kubo, Noriaki Sadanaga, Hidetaka Yamamoto, and Hiroshi Honda. Enhanced mass on contrast-enhanced breast MR imaging: Lesion characterization using combination of dynamic contrast-enhanced and diffusion-weighted MR images. *Journal of magnetic resonance imaging : JMRI*, 28(5):1157–65, November 2008.
197. Fumiyuki Yamasaki, Kazuhiko Sugiyama, and Kaoru Kurisu. Brain Tumors: Apparent Diffusion Coefficient at Magnetic Resonance Imaging. *Methods of Cancer Diagnosis, ...*, 8:pp 279–296, 2010.
198. D Yang, Y Korogi, T Sugahara, M Kitajima, Y Shigematsu, L Liang, Y Ushio, and M Takahashi. Cerebral gliomas: prospective comparison of multivoxel 2D chemical-shift imaging proton MR spectroscopy, echoplanar perfusion and diffusion-weighted MRI. *Neuroradiology*, 44(8):656–66, August 2002.
199. Jian Yang, Alejandro F Frangi, Jing-Yu Yang, David Zhang, and Zhong Jin. KPCA plus LDA: a complete kernel Fisher discriminant framework for feature extraction and recognition. *IEEE transactions on pattern analysis and machine intelligence*, 27(2):230–44, February 2005.
200. Jianhua Yang. *Algorithmic Engineering of Clustering and Cluster Validity with Applications to Web Usage Mining.* PhD thesis, University of Newcastle, 2002.

201. Xiangyu Yang and Michael V Knopp. Quantifying tumor vascular heterogeneity with dynamic contrast-enhanced magnetic resonance imaging: a review. *Journal of biomedicine & biotechnology*, 2011:732848, January 2011.
202. Chalup S. Yang, Jianhua, Estivill- Castro V. Support vector clustering through proximity graph modelling. *Neural Information Processing, 2002. ICONIP '02. Proceedings of the 9th International Conference on*, 2:898–903, 2002.
203. Thomas E Yankeelov, Martin Lepage, Anuradha Chakravarthy, Elizabeth E Broome, Kenneth J Niermann, Mark C Kelley, Ingrid Meszoely, Ingrid A Mayer, Cheryl R Herman, Kevin McManus, Ronald R Price, and John C Gore. Integration of quantitative DCE-MRI and ADC mapping to monitor treatment response in human breast cancer: initial results. *Magnetic resonance imaging*, 25(1):1–13, January 2007.
204. Zhang Yili, Huang Xiaoyan, Du Hongwen, Zhang Yun, Chen Xin, Wang Peng, and Guo Youmin. The value of diffusion-weighted imaging in assessing the ADC changes of tissues adjacent to breast carcinoma. *BMC cancer*, 9(1):18, January 2009.
205. L. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965.
206. Li Zeng, Bin Chen, Linping Du, and Kejia Xu. Manifold based kernel optimization for KPCA. In *2011 IEEE 3rd International Conference on Communication Software and Networks*, pages 69–72. IEEE, May 2011.
207. B Zimmermann, A Moegelin, P de Souza, and J Bier. Morphology of the development of the sagittal suture of mice. *Anatomy and embryology*, 197(2):155–165, 1998.