Bayesian Generative Learning of Brain and Spinal Cord Templates from Neuroimaging Data sets

Claudia Blaiotta

Dissertation submitted in partial fulfilment

of the requirements for the degree of

Doctor of Philosophy

of

University College London

Wellcome Trust Centre for Neuroimaging

Institute of Neurology

University College London

I, Claudia Blaiotta confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

In the field of neuroimaging, Bayesian modelling techniques have been largely adopted and recognised as powerful tools for the purpose of extracting quantitative anatomical and functional information from medical scans. Nevertheless the potential of Bayesian inference has not yet been fully exploited, as many available tools rely on point estimation techniques, such as maximum likelihood estimation, rather than on full Bayesian inference.

The aim of this thesis is to explore the value of approximate learning schemes, for instance variational Bayes, to perform inference from brain and spinal cord MRI data. The applications that will be explored in this work mainly concern image segmentation and atlas construction, with a particular emphasis on the problem of shape and intensity prior learning, from large training data sets of structural MR scans.

The resulting computational tools are intended to enable integrated brain and spinal cord morphometric analyses, as opposed to the approach that is most commonly adopted in neuroimaging, which consists in optimising separate tools for brain and spine morphometrics.

Contents

List of Figures

List of Tables

1

Introduction

The demand for automated image processing tools has increased dramatically over the last few years. This has been happening in parallel to the collection of large digital data sets, which is a prominent feature of the most recent phase of the information era.

Interestingly, the exponential increase in data storage capacity and in processing speed of computers are not sufficient to explain the phenomenon that is nowadays known as the big data revolution. Certainly, the availability of larger computational resources is a crucial factor, but it could not by itself sustain and motivate the collection of such massive volumes of data if there was not a concurrent effort towards the development of improved statistical and computational methods for data processing (Wu et al., 2014b). In fact, the expensive process of collecting new data becomes truly profitable only when tools are available to analyse and discern hidden patterns in the data themselves.

In the field of medical imaging, such a process has been further encouraged by the constantly improving performance of imaging devices. As a result, algorithms for medical image processing are currently expected to be able to extract information in an automated manner, so as to allow fast, quantitative and repeatable analyses in research as well as in clinical practice. Typical examples include image segmentation, registration, reconstruction and classification algorithms.

Medical image computing problems are addresses in this thesis from a Bayesian modelling perspective. In such a framework, mathematical models have to be formulated, fitted to the observed data, and compared for model selection (MacKay, 1992). Therefore, from a conceptual point of view, the development of image processing methods is indeed the search for the best models to represent imaging data. This last aspect might be sometimes overlooked, in spite of being the fundamental question underlying all probabilistic data science problems.

A number of advantages derive from the choice of relying on probabilistic modelling techniques. These include the possibility of describing uncertainty and noise in the data, the opportunity to make predictions and infer unknown quantities from experimental observations (Ghahramani, 2013), as well as the chance of comparing models to select the one that is most explanatory of the observed data (MacKay, 1992). On the other hand, exploiting the potential of full Bayesian inference is typically challenging both from a mathematical and computational point of view, thus often requiring the adoption of approximate inference schemes.

The central topic of this thesis is the application of variational Bayesian learning techniques to model structural neuroimaging data sets. The main fields of application are: image segmentation, image registration and atlas construction. Moreover, as opposed to much of the work done so far in neuroimaging, a systemic vision is proposed, to demonstrate that different parts of the central nervous system, such as the brain and the spinal cord, can be effectively represented within a single modelling framework.

1.1. Image segmentation

In computer vision the term image segmentation refers to the task of partitioning a digital image into subsets consisting of pixels, or voxels, that share common properties, such as colour, intensity or membership of the same object. The development of image segmentation techniques is generally motivated by the need to perform some form of object, or structure, recognition task, in a fast and automated manner. Therefore applications range over a wide spectrum. Nonetheless, among these, medical imaging applications are among the ones which have received the most attention over the last few years, because of the significant impact they can have on medical research and clinical practice.

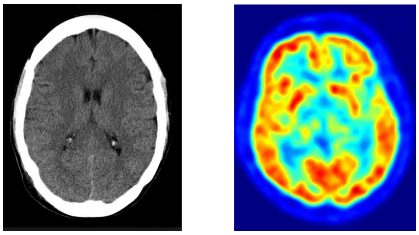
There are, in fact, a number of clinical applications where image segmentation could potentially be very useful. They include both diagnostic procedures, for example lesion or tumour detection (Mustaqeem et al., 2012), and therapeutic interventions, such as treatment or surgical planning (Gering et al., 1999). Moreover, in medical research, image segmentation algorithms have been recognised as particularly valuable in the field of neuroimaging, where they can be used, for example, as processing tools in the context of studying normal and pathological variability of brain anatomy (Ashburner and Friston, 2000), as well as for the purpose of mapping functional activations (Maldjian et al., 2003).

Particularly in medical imaging, where manual annotation of the data requires extensive training of the raters, the primary scope of using automated segmentation tools is to make the analyses less time consuming and more reproducible. However it has also been shown that, especially by exploiting simultaneously different image contrasts, the results of automated image segmentation can provide more accurate information compared to simple visual inspection (Bezdek et al., 1992).

A wide range of algorithmic methods have been exploited so far to perform medical image segmentation, among which are, clustering algorithms (Chuang et al., 2006), probabilistic generative models (Ashburner and Friston, 2005), multi-atlas segmentation (MAS) methods (Aljabar et al., 2009), region growing techniques (Pohle and Toennies, 2001), deformable contour models (He et al., 2008) and deep neural networks (Zhang et al., 2015b), just to provide some of the most relevant examples. An extensive review of the methods for medical image segmentation can be found in Norouzi et al. (2014); Pal and Pal (1993); Pham et al. (2000); Setarehdan and Singh (2012); Sharma and Aggarwal (2010). The work presented in this thesis is mainly set in the framework of probabilistic atlas-based methods, where in particular the term atlas refers to prior probabilistic maps encoding tissue composition.

1.1.1. Probabilistic tissue classification from MRI data

When analysing neuroimaging data, it is very helpful to partition the brain into different tissue types. This processing step often represents the first stage for performing brain volumetric and morphometric analyses, which are extremely valuable in research and potentially for clinical practice (Ashburner and Friston, 2000; Giorgio and De Stefano, 2013). In fact, quantifying neural tissue volume not only has a major role for unravelling the mechanisms underlying neurodegenerative and psychiatric disorders, but can also significantly help in disease diagnosis and treatment planning or monitoring (Mazzara



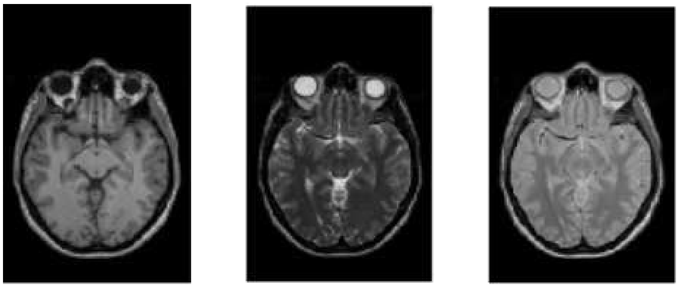
(a) CT (b) PET

Figure 1.1: Examples of CT (a) and PET (b) brain scans

et al., 2004).

For healthy subjects, the tissues of interest are typically gray matter, white matter and cerebrospinal fluid, while for patients, additional classes may be defined, such as tumour, oedema or necrosis (Moon et al., 2002; Prastawa et al., 2004). For this purpose, MRI is usually the most convenient imaging modality to work with, as, without using ionising radiation, it provides excellent soft tissue contrast and good signal to noise ratio, compared to other imaging techniques. For example, CT yields very good contrast between bone and soft tissue but is generally inadequate for correctly differentiating tissues within the brain (Loubele et al., 2006). In the case of PET, which is a functional rather than structural imaging technique, tissue segmentation is quite challenging due to poor spatial resolution, low signal to noise ratio, together with scatter and signal attenuation effects (Boellaard, 2009). Examples of CT and PET brain scans are depicted in Figure 1.1. Moreover magnetic resonance imaging opens up the possibility of enhancing contrast between specific tissues simply by adjusting the acquisition parameters (Figure 1.2).

As a result, a lot of work has been done to develop algorithms that are capable of automatically identifying tissue types from MRI data. Most of these methods rely on the contrast between the intensities of different tissues to assign a tissue label associated with each voxel. The problem can be effectively solved from a probabilistic generative modelling point of view. Essentially, this requires learning the intensity distributions of the tissues of interest. Once such distributions have been estimated the unknown tissue labels can be inferred making use of Bayes' rule. More precisely, if a parametric representation of the intensity distributions is adopted, where Θ denotes a set of parameters,



(a) T1w (b) T2w (c) PDw

Figure 1.2: Examples of MRI contrasts obtained by varying the scanning parameters. The three panels report a T1-weighted (a) a T2-weighted (a) and a PD-weighted (a) scan.

the probability of voxel j belonging to tissue k can be computed as

$$p(z_j = k|\mathbf{x}_j, \Theta) = \frac{p(z_j = k, \mathbf{x}_j|\Theta)}{\sum_{k=1}^{K} p(z_j = k, \mathbf{x}_j|\Theta)} = \frac{p(z_j = k)p(\mathbf{x}_j|z_j = k, \Theta)}{\sum_{k=1}^{K} p(z_j = k)p(\mathbf{x}_j|z_j = k, \Theta)} \ , \tag{1.1}$$

where $\mathbf{x}_j$ indicates the observed image intensity at location j and z_j is a discrete latent (unobserved) variable encoding class memberships. From this expression, it is clear that the crucial point is learning an optimal model of the observed intensities, which allows computing the conditional probabilities $p(\mathbf{x}_j|z_j, \Theta)$.

Various approaches have been proposed so far by different authors. In particular different parametrisations might be adopted, as well as different learning techniques. A widely used strategy consist in modelling the likelihoods of observed data (conditional probabilities of the data given the labels) as Gaussian distributions, that is to say

$$p(\mathbf{x}_j|z_j = k, \Theta) \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \ , \tag{1.2}$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the mean vector and covariance matrix of tissue class k. This naturally leads to a finite Gaussian mixture model (GMM), where data points (i.e. voxels intensities) are assumed to be statistically independent. Some authors (see for example Bricq et al. (2008); Held et al. (1997); Kapur et al. (1998); Van Leemput et al. (1999a); Warfield et al. (2004)) have argued that such a class of models might be lacking spatial constraints to enforce the piecewise homogeneity of tissue labels. A very well studied strategy to address this issue consists in introducing Markov random field priors to ensure spatial coherence (Cross and Jain, 1983). This approach is particularly useful for increasing robustness to noise, therefore it might become essential at very low signal to noise ratios (SNR) or in the presence of artifacts.

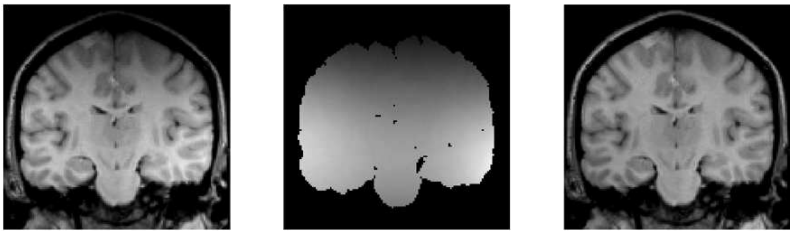
Despite being very well suited to perform intensity-based tissue segmentation, MRI data is often corrupted by artifacts that make the tissue classification task non-trivial. Among these, thermal noise, intensity non-uniformity and partial volume effects are probably the most common. These phenomena cause the intensity distributions of different tissues to partially overlap, thus making the classification problem more challenging (Zhang et al., 2001). Fortunately, one of the big advantages deriving from adopting a probabilistic perspective is that prior population-based anatomical knowledge can be easily incorporated within such models, for example in the form of smooth average-shaped tissue probability maps (Xu et al., 2014), thus helping to alleviate the above mentioned problems. Additionally, specific artifact correction strategies are available (Shattuck et al., 2001; Sled et al., 1998).

1.1.2. Modelling intensity inhomogeneities

Intensity inhomogeneities, in MR images, are smooth variations of intensity, which are not caused by random noise (Figure 1.3). Such a phenomenon is very common and occurs for multiple reasons. Among these, the most relevant ones are the inhomogeneity of the static (B0), radio-frequency (B1) and gradient fields, the non-uniformity of detector sensitivity and electrodynamic interactions between the magnetic field and the scanned object (RF penetration and standing wave effects) (Lewis and Fox, 2004; Rajapakse and Kruggel, 1998). Additional minor causes are: eddy currents driven by the switching of the gradient fields, mistuning of the RF coil and bandwidth filtering of the data (Mazziotta et al., 2001).

At low field strengths the imperfect spatial homogeneity of the static field B0 is the main cause of these slow intensity variations. At higher MR field strengths the contribution of B0 diminishes while other effects, such as tissue dependent distortions produced by MR gradients, start to become more significant (Ganzetti et al., 2016).

This type of artifact, can rarely make visual interpretation of the scans harder but, most importantly, strongly affects the results of quantitative MR analyses. Figure 1.4 shows, for example, the very poor results produced by a tissue classification algorithm based on Gaussian mixture models (Ashburner and Friston, 2005) after having disabled



(a) (b) (c)

Figure 1.3: MR image corrupted by intensity inhomogeneities (a), bias field (b) and corrected image (c).

the bias correction option.

According to the RF field mapping theory, intensity inhomogeneities can be modelled as a multiplicative bias (Ganzetti et al., 2016). This assumption is largely accepted, so most retrospective intensity inhomogeneity correction methods try to estimate a low frequency field that, once multiplied by the data, will compensate for the distortion. Methods where the bias is decomposed into a multiplicative and an additive component have also been proposed (Likar et al., 2001), even if the additive component is most often neglected, unless the data has been log transformed.

Many computational methods have been proposed to correct inhomogeneities in MR data (Ashburner and Friston, 2005; Brinkmann et al., 1998; Guillemaud and Brady, 1997; Lewis and Fox, 2004; Likar et al., 2001; Mihara et al., 1998; Rajapakse and Kruggel, 1998; Sled et al., 1998; Styner et al., 2000; Tustison et al., 2010; Van Leemput et al., 1999d; Wang et al., 1998). The techniques exploited by such methods include: non-parametric non-uniform intensity normalisation (N3/N4) (Sled et al., 1998; Tustison et al., 2010), Fourier domain filtering (Haselgrove and Prammer, 1986), histogram matching (Wang et al., 1998), homomorphic filtering (Guillemaud, 1998), information theoretic approaches based on entropy minimisation (Likar et al., 2000), generative models of image intensity distributions (Ashburner and Friston, 2005; Van Leemput et al., 1999d).

Empirical methods for bias correction based on measures of the inhomogeneity field (obtained for example on phantoms) such as in Tincher et al. (1993) have become less widely used over the past few years, due to their impractical nature as well as to the low validity of the assumption that the bias is subject independent.

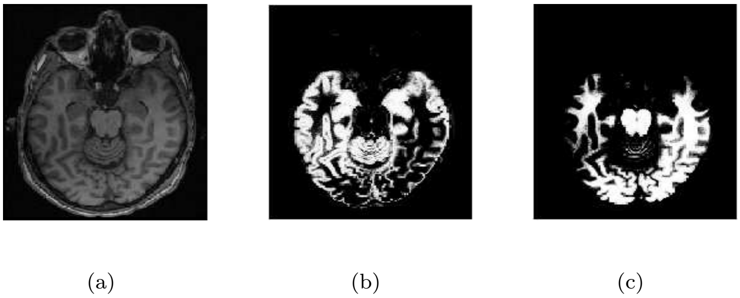

(a) (b) (c)

Figure 1.4: MR image corrupted by intensity inhomogeneities (a) and the resulting gray (b) and white (c) matter segmentations, produced with the segmentation algorithm implemented in the SPM12 software by disabling the bias correction option.

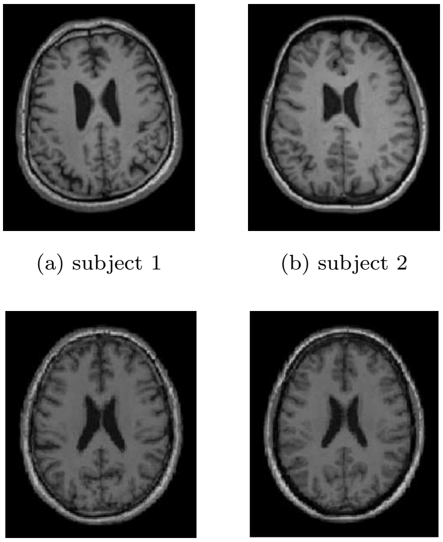
1.2. Spatial normalisation

When working with medical images, it is very often desirable to bring different individual scans into a common anatomical space. This task is often referred to as spatial normalisation, or registration, and it represents an essential processing step for a wide range of applications. From a very general perspective, the reason for this is that data acquired from different subjects cannot directly be compared without mapping individual anatomies into some form of or reference space, where morphological and functional correspondences are more likely.

This is, for example, a critical step in the context of functional neuroimaging studies, where common activation patterns across subjects are sought (Orchard et al., 2003). Other applications include atlas-based segmentation techniques (Cabezas et al., 2011), the realignment of scans of the same subject acquired at different times (Jenkinson et al., 2002), as well as the cross-sectional or longitudinal modelling of structural differences, or changes (Ashburner et al., 2003; Kipps et al., 2005; Scahill et al., 2003).

Due to its broad range of applications, the topic of medical image registration has been widely explored over the past few years. Numerous approaches and processing tools have been proposed and compared. For an extensive survey see Sotiras et al. (2013). Nevertheless many questions remain to be answered, especially when it comes to the biophysical plausibility and interpretability of the results.

Broadly speaking, the process of registering a set of images typically involves:



(a) subject 1

(b) subject 2

(c) subject 1 normalised

(d) subject 2 normalised

Figure 1.5: Example of spatially normalised images.

- Defining a spatial transformation model
- Identifying an objective function and a suitable optimisation strategy
- Selecting an interpolation strategy to resample the images at the locations specified by the underlying deformation model

A number of options are available with respect to each of the three points listed above. Therefore many alternative methods can be designed, which differ in the strategy adopted for at least one of such points.

With regard to the modelling of spatial transformations, there is a large number of available approaches, which can be broadly divided in global affine and local non-rigid methods (Lester and Arridge, 1999). The first ones allow only a limited number of degrees of freedom (typically from 6 to 12), while the second can be extremely highly dimensional and include up to a maximum of $3N$ free scalar parameters, with N equal to the number of voxels constituting the image (Modersitzki, 2004).

Affine transformations, which permit translation, rotation, scaling and skewing, are intrinsically global and cannot model local structural properties, unless a piecewise approach is adopted. On the contrary, non-rigid methods belonging to the second group are local in the sense that they can capture morphological differences at a much smaller spatial scale. Approaches of this second type can be further subclassified depending on how the deformations are modelled, or parametrised. In particular, two broad categories can be identified. The common feature of the methods belonging to the first one is the introduction of a small displacement field that is added to the identity transform in order to map between different anatomies. Instead, methods of a second sort aim at capturing larger shape variations, while preserving topological properties and, to do so, they rely on the construction of diffeomorphic deformation fields, essentially by integrating a velocity field over multiple time points (Ashburner, 2007; Joshi et al., 2004; Rueckert et al., 2006; Vercauteren et al., 2009).

The second crucial point is identifying a suitable objective function. This generally requires choosing a distance metric to quantify the similarity between two, or multiple, images. With respect to this, the choice of the the metric strongly depends on whether the images were acquired with the same or with different modalities. In the first case, because the scans share similar intensity distributions, suitable distance measures can be estimated based on voxel-wise intensity differences. Indeed, the most common solution for intra-modality registration involves computing the sum of the squared differences (SSD) as a similarity metric. Alternatively, local normalised cross-correlation can also be used (Avants et al., 2008), which is invariant to linear transforms of the intensities.

On the contrary, in the case of inter-modality registration, using SSD-based objective functions is not a viable option. In such cases, information theoretic approaches are more suitable, as they allow quantification of the amount of information shared by images, without relying on the difference between intensities, which in this case is not informative for measuring image similarity. In particular mutual information (MI) represents the most commonly adopted metric (Maes et al., 1997, 2003; Wells et al., 1996; Zitova and Flusser, 2003). Approaches based on the concept of image self-similarity have also been proposed (Heinrich et al., 2012).

Unfortunately, optimising the coordinate transformation just by minimising a distance metric is not feasible in practice, because the registration problem is inherently ill-posed, with non-unique and unstable solutions. As a result additional constraints (regularisation) have to be introduced (Hill et al., 2001). The regularising term is usually incorporated in the objective function together with the matching (similarity) term, so as to restrict the space of possible solutions. Only in a limited number of cases, for example when the transformation space is very low dimensional, the problem is implicitly regularised without having to add an explicit penalty term. Most regularisers are computed as combination of L_2 norms of the derivatives of the displacement and can be interpreted according to physical (most commonly mechanical) models (Burger et al., 2013; Sorzano et al., 2005). Examples include elastic, diffusion- and curvature-based regularisers. In addition to assuring a well-posed mathematical framework, regularisation is also exploited to enforce biophysical plausibility of the transformations. For example it might be reasonable to use known biomechanical properties to constrain the deformations (e.g. impose a local rigidity constraint in the presence of bony structures etc.). Topology preservation is also a desirable property, commonly enforced by restricting the space of solutions to locally invertible transformations.

Another aspect, which has to be taken into account to solve image registration problems, is how to interpolate images. In fact, digital images are discrete, sampled versions of an underlying continuous signal. Therefore, whenever a spatial transformation is applied, it is necessary to resample such a continuous signal at new locations specified by the transformation model. Most common interpolation schemes are linear (Lehmann et al., 1999) and spline interpolation (Hou and Andrews, 1978). These methods, in spite of the different nomenclature, belong indeed to the same family, as the linear approach is equivalent to first order B-spline interpolation. Nevertheless they can differ very much in computational complexity and time (Parker et al., 1983). The choice of a suitable interpolation scheme, however, does not only depend on computational convenience. In fact, it has been shown that different interpolation strategies can impact the accuracy of registration. For example, some interpolation approaches can cause the presence of a high number of spurious local optima in the objective function, thus compromising its smoothness and making the optimisation more challenging (Tsao, 2003). Another factor to be taken into account is that higher order interpolation might not always preserve existing constraints on image intensities; that's why, for instance, linear approaches are more suitable to interpolate probability maps, which are bounded in a probability simplex.

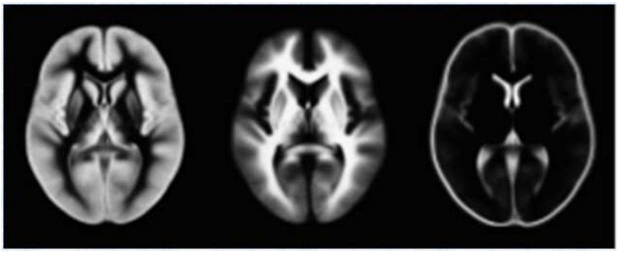


Figure 1.6: Example of probabilistic brain atlas. From left to right gray matter, white matter and cerebrospinal fluid tissue probability maps are illustrated.

1.3. Atlas-based methods: a link between image segmentation and registration

A great number of segmentation algorithms make use of prior information in the form of probabilistic atlases [1] (Ashburner and Friston, 2005; Fischl et al., 2002; Yeo et al., 2008). Indeed, as opposed to purely intensity driven clustering methods (Gerig et al., 1992), atlas-based strategies allow accurate differentiation of structures that have similar (i.e. overlapping) intensity distributions, in spite of belonging to different tissue types, or structures. Additionally, by incorporating prior anatomical knowledge, robustness to noise and imaging artifacts is increased (Pham et al., 2000). Further robustness is often achieved by introducing contextual information via Markov random fields (Bricq et al., 2008; Van Leemput et al., 1999b).

Figure 1.6 illustrates, as an example, a set of probabilistic templates of gray matter, white matter and cerebrospinal fluid. Atlases of this sort are widely used to perform automated tissue classification in neuroimaging (Ashburner and Friston, 2005; Bricq et al., 2008; Cabezas et al., 2011; Van Leemput et al., 1999b).

Other types of atlases, which, rather than tissue labels, carry cytoarchitectonic labels on an average-shaped anatomy (Fan et al., 2016), can instead be used to parcellate

[1] The term atlas is widely used in medical image computing. However, depending on the particular framework, atlases can encode different types of information. In the remaining chapters of this thesis the term atlas will refer to average-shaped tissue probability maps that indicate the prior probability of finding each tissue type at every location in a reference anatomical coordinate system. This approach should not be confused with the multi-atlas framework for image segmentation, which instead makes use of a set of labelled images of individual subjects, also referred to as atlases.

organs into different functional or structural areas. Atlases of these sort can even encode information on the intensity distributions relative to different anatomical structures, so as to ensure higher robustness and better classification performance (Fischl et al., 2002). An example of a brain atlas with cortical and subcortical region labels is provided in Figure 1.7.

In order to make use of such *a priori* information, atlas-based segmentation algorithms rely on the knowledge, or estimation, of a spatial transformation, which brings the template in register with a new individual image (Ashburner and Friston, 2005). The underlying idea being that, when an atlas is warped to match an individual scan, structural and functional correspondences are ensured (Figure 1.8).

A similar principle is at the basis of the so called multi-atlas segmentation (MAS) techniques. The main difference between these and conventional probabilistic atlas-based methods, which will be extensively explored throughout this thesis, is that, in the case of multi-atlas segmentation, information encoded in the training data is not summarised in a single population-based atlas. Instead, each training sample, consisting of a single subject image with an associated manual segmentation, constitutes an atlas, which is warped onto unseen individual scans (Klein et al., 2005). Propagation and then fusion of the labels provided by each atlas allows a single segmentation of the test data to be attained (Heckemann et al., 2006; Langerak et al., 2010). For a comprehensive review of multi-atlas segmentation strategies see Iglesias and Sabuncu (2015). Interestingly, probabilistic formulations of the label fusion problem have also been proposed (Iglesias et al., 2013b; Sabuncu et al., 2010).

The crucial point here is that image segmentation and registration problems are intrinsically interdependent. Nevertheless they are often solved as individual tasks. For instance, the most common approach is to first register a target image to one, or multiple, templates and, in a second processing step, obtain tissue labels by manipulation of the anatomical information encoded in the atlas(es).

From a theoretical point of view this corresponds to formulating two separate models. A first one to find morphological correspondences, which typically does not take into account the unknown tissue labels of the test image, and a second one that, making use of the deformation field estimated in the previous step, enables the estimation of anatomical labels. In spite of providing acceptable results, approaches of this sort are

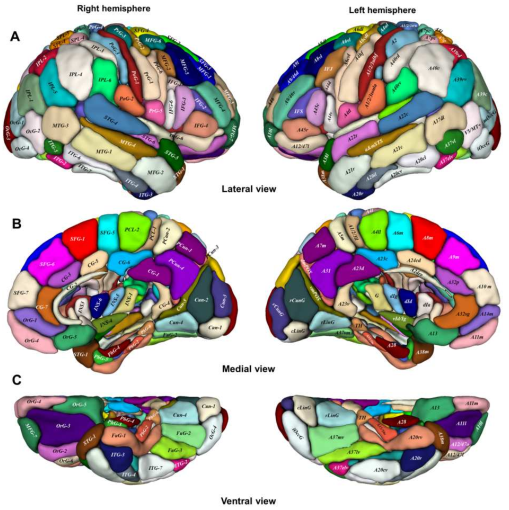
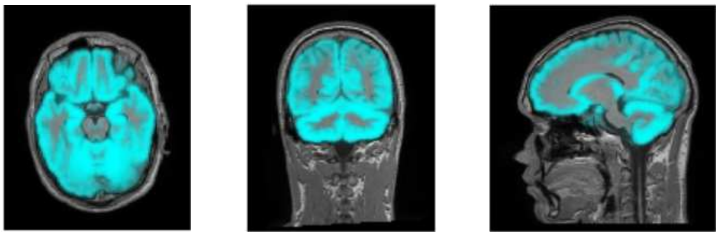


Figure 1.7: Example of brain atlas for anatomical and functional parcellation (Fan et al., 2016).



(a) (b) (c)

Figure 1.8: Example of gray matter tissue probability map overlaid on an individual scan. Axial (a), coronal (b) and sagittal (c) views.

somehow suboptimal (Ashburner and Friston, 2005). In fact, when the second step has been solved, the acquired information on tissue composition or structure location could be exploited to further refine the estimated spatial transformation (Mahapatra and Sun, 2012), which would in turn help to improve the results of the second (segmentation) step. In other words, it would be more convenient, and more accurate at the same time, to formulate exhaustive mathematical models, capable of capturing simultaneously both shape and tissue composition. Indeed, methods of this sort have already been explored by a number of authors (Ashburner and Friston, 2005; DAgostino et al., 2006; Pohl et al., 2006; Xiaohua et al., 2004b; Yezzi et al., 2001) and the results of their experiments seem to indicate that solving simultaneously image segmentation and registration tasks can provide more accurate solutions, compared to decoupling of the two problems (Pohl et al., 2006).

For this reason, joint modelling techniques should always be preferred, in spite of being potentially more computationally expensive, compared to tools that only solve one sub-problem at a time, due to the higher number of parameters to be estimated and to the greater complexity of the underlying models. Such drawbacks are in fact most often manageable with the computational resources available these days. Moreover they are counterbalanced by an increased reliability of the results, as well as by the practical convenience of solving multiple tasks within a single algorithmic framework (Ashburner and Friston, 2005).

1.3.1. Probabilistic tissue template construction

In its most simplistic implementation, the construction of average-shaped tissue probability maps involves averaging a number of individual segmentations, or label maps, after they have been spatially normalised. This indicates that the processes of segmenting images and generating atlases are intrinsically related in a circular manner, as to produce accurate segmentations it is desirable to have an adequate atlas and to construct a representative atlas it is necessary to have a set of accurately segmented images.

It is therefore natural to try to solve both problems simultaneously. For this purpose, one natural solution consist in trying to enforce, in a mathematical form, the fact that individual segmentations are realisations of a stochastic process governed by a prior anatomical model, which can be inferred from a large data set of individual observations. Along this line, the works of Bhatia et al. (2007) and Ribbens et al. (2010) provide interesting probabilistic formulations, relative to the problem of groupwise tissue template construction.

A considerable part of the work presented in the following chapters relies on the idea that, with a single hierarchical generative model of MR data, it is possible to capture morphological variability across a homogeneous population in the form of average-shaped probabilistic atlases. In particular, in the spirit of Ashburner and Friston (2005), it will be assumed that image intensities are drawn from multivariate Gaussian mixture distributions, with the incorporation of spatially varying tissue priors, which are unknown but, as the following chapters will illustrate, can be learned directly from large multispectral MR data sets by fitting generative latent variable models.

This approach, which will be explored in detail in the remainder of this thesis, defines a general computational framework, which could serve to learn representative and unbiased priors, for many different populations. Therefore, it could also open up the possibility of extending well-established image processing techniques to the analysis of data sets that are currently considered difficult to deal with, due to the lack of appropriate prior models (e.g. data relative to particular age groups, as well as animal or pathological data).

1.4. Bridging the gap between brain and spine imaging

As anticipated in the previous section, the work presented in this thesis aims to explore the potential of Bayesian generative modelling techniques to learn anatomical priors from cross-sectional imaging data sets. Given its generality, such a methodological framework could in principle be exploited to solve a diverse range of medical image computing problems. However the application studied in this thesis regards primarily the development of computational tools to process within a single modelling framework both brain and spinal cord MR data [2]. This is a rather unexplored research topic, as the most common approach adopted in neuroimaging consists in optimising different tools either for brain or spinal cord data. As a result, these two applications remain hard to integrate in practice, due to the lack of a common processing framework.

In addition, it should be noted that there exists a technical gap between brain and spine imaging. In fact, incredible effort has been put by the neuroscientific community into the development of computational techniques to enable morphometric brain studies. Such tools have been extensively tested, validated and improved during the past twenty five years, both for healthy subjects (Avants et al., 2011a; Ghosh et al., 2010; Gronenschild et al., 2012; Išgum et al., 2015; Klein et al., 2009; Wenger et al., 2014) and pathological populations (Ghosh et al., 2010; Pereira et al., 2010; Popescu et al., 2012; Wang et al., 2007), and are now easily accessible for researchers working on brain imaging data (Ashburner, 2007; Ashburner and Friston, 2005; Avants et al., 2011b; Cox, 1996; Fischl, 2012; Klein et al., 2010; Smith et al., 2004).

In the meanwhile, the progress in the field of spinal cord MR imaging has been slower. This is in part due to numerous challenges in the process of data acquisition (Cohen-Adad et al., 2011; Wilm et al., 2007), which have made large spinal cord imaging studies impractical and therefore less appealing compared to brain imaging experiments. In fact, many of the technical challenges encountered is spine MRI are directly related to the peculiar anatomy and to the geometrical properties of the spinal cord (Cohen-Adad et al., 2011; Lycklama et al., 2003). For example, the small cross-sectional area of the cord, whose diameter is around 1 cm, together with its large rostrocaudal extension (approximately 45 cm), make the image acquisition process much more challenging than in conventional brain MRI. In particular, achieving high resolution in the transverse plane (at least 1 mm × 1 mm) becomes crucial and, as a result, the amount of data that needs to be collected in order to cover long spinal portions can become incredibly large. In fact, the field of view (FOV) typically spans the entire body width, to minimise the impact of aliasing effects, which otherwise would need to be controlled using spatial suppression pulses to eliminate the signal from regions outside the FOV. Additionally,

[2] The PhD project discussed in this thesis was funded according to UCL Impact studentship scheme in partnership with Balgrist University Hospital in Zurich, which is a world leading institution for clinical research on traumatic spinal cord and musculoskeletal injuries.

the periodic pulsation of the cerebrospinal fluid (CSF), together with respiration, introduces motion artifacts, which need to be minimised in the acquisition phase or corrected during post-processing (Mohammadi et al., 2013; Taber et al., 1998). These are just some of the main difficulties encountered in spinal cord MRI. Additional issues can for example arise when imaging patients with orthopaedic implants (Petersilge et al., 1996; Rudisch et al., 1998).

Such a gap between brain and spinal cord imaging techniques is maybe among the reasons why not much research has been conducted on the development of image processing algorithms for spinal cord data, as opposed to the large effort invested into the design of processing solutions for brain scans.

In particular, many of the publicly available automated processing tools, which perform well on brain images, either are not applicable or have a poorer performance at the cord level (De Leener et al., 2016). As a result, most of the analyses performed on spinal cord images still require large amounts of manual editing (e.g. manual identification of the cord centre or manual delineation of the cord and its internal structure) (Horsfield et al., 2010; Yiannakas et al., 2012).

Bridging the gap between brain and spinal cord imaging will require further advance in both image acquisition and image processing techniques. Indeed, a number of research groups are currently working on the development of dedicated tools to analyse spinal cord MR data (a brief survey on their work will be presented the following section of this chapter). However, such tools are still hard to assimilate with brain image processing methods, while in principle it would be very helpful for the neuroimaging community to have a common modelling framework capable of handling simultaneously the diverse challenges presented by brain and spinal cord images, thus allowing to perform integrated brain and spine morphometric analyses. The potential impact of having general computational frameworks to deal with the entire central nervous system is incredibly promising. In fact, numerous studies have already shown that spine MRI may help in differential diagnosis and disease progression monitoring, as opposed to solely using brain MRI scans (Bot et al., 2004; Freund et al., 2016; Losseff et al., 1996). Such a problem will be addressed in the remaining chapters, by exploiting hierarchical generative models of MR data. In particular, the line adopted in this work consists in keeping the mathematical formulation of such models as unified and general as possible, so as to ensure maximal generalisation capability.

1.5. Spinal cord imaging and volumetry: the state-of-the-art

The spinal cord is a long and thin cylindrical structure of the central nervous system and it constitutes the main pathway for transmitting information between the brain and the rest of the body. Its peripheral region is constituted by white matter tracts, which contain sensory and motor axons, both ascending and descending, while the central region is formed by three grey matter columns containing nerve cell bodies. Just like the brain, the spinal cord is a major site of traumatic injury (Huber et al., 2015) and it can be affected by a number of neurodegenerative diseases, such as multiple sclerosis, amyotrophic lateral sclerosis, transverse myelitis and neuromyelitis optica (Rocca et al., 2015).

Understanding the degenerative processes underlying these pathologies represents a crucial step towards the development of effective therapeutic interventions, as well as towards the identification of sensitive and selective diagnostic criteria. In particular, quantification of spinal cord tissue loss (i.e. atrophy) has been regarded over the past two decades as a promising biomarker (Filippi et al., 1996; Freund et al., 2013a; Grabher et al., 2015; Kidd et al., 1993; Losseff and Miller, 1998; Losseff et al., 1996), which could potentially help in monitoring disease progression, predicting clinical outcome and understanding the mechanisms underlying neurological disability (e.g. demyelination, inflammation, axonal or neuronal loss), in a number of conditions that affect the central nervous system both at the brain and spinal cord level, such as multiple sclerosis (MS) and traumatic spinal cord injury (SCI) (Bakshi et al., 2005; Freund et al., 2013a,b; Grossman et al., 2000; Miller et al., 2002).

Neuroimaging techniques, particularly MRI, represent the most effective tools to investigate non-invasively and *in vivo* the structure and function of the spinal cord, both in physiological and pathological conditions. Figure 1.9 illustrates two examples of brain and cervical cord MR scans.

Unfortunately, spinal cord MRI is not immune from technical challenges. Some of

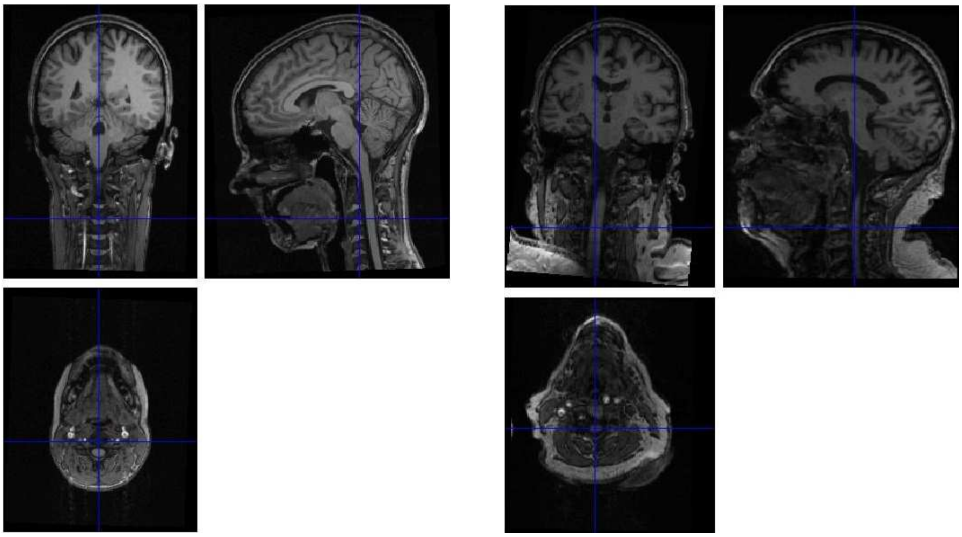


Figure 1.9: Examples of brain and cervical cord MR scans

them are intrinsic to MR imaging, such as the presence of intensity inhomogeneities, while others arise from the peculiar anatomy of the cord itself, for instance from its small cross-sectional area (Grossman et al., 2000; Stroman et al., 2014; Wheeler-Kingshott et al., 2014). Nevertheless, spinal cord imaging using MR techniques has improved significantly over the past few years, especially with the introduction of phased-array surface coils and fast spin-echo sequences (Stroman et al., 2014).

For spinal cord imaging studies, delineating the cord represents the first step before assessing atrophy or detecting any other morphometric change, or difference. This indicates that there is an urgent need not only for automated algorithmic solutions dedicated to spinal cord tissue classification and image registration (Chen et al., 2013; De Leener et al., 2017; Fonov et al., 2014; Levy et al., 2015; Taso et al., 2014; Van Uitert et al., 2005), but also for large, systematic and reproducible validation studies to objectively assess the performance of such tools (Prados et al., 2017).

Not surprisingly, the first methods that appeared in the literature to perform spinal cord image segmentation and the subsequent volumetric analyses were based on semi-automated algorithms. Among these, one of the earliest is described in the work of Coulon et al. (2002), where they introduce an algorithm for fitting a cylindrical cubic B-spline surface to MR spinal cord images, which requires the user to provide a set of landmarks that will define the medial axis of the initial surface.

Later on, few other semi-automated solutions have been presented by Van Uitert et al. (2005) and Horsfield et al. (2010). In particular, Van Uitert et al. (2005) proposed a semi-automated segmentation technique based on level set methods, which was preliminary validated only on 2D data. The method of Horsfield et al. (2010) is instead based active surface models and it was validated on T1-weighted images of healthy controls as well as MS patients acquired at 1.5 T. In both cases, the user has to approximatively mark the cord centre, so as to provide a reliable initialisation of the algorithm.

Only recently have fully automated spinal cord segmentation methods started to be proposed. Chen et al. (2013) introduced a fuzzy c-means algorithm with topological constraints to segment the cervical and thoracic spinal cord from MR images. Their method relies on a statistical atlas of the cord and the surrounding CSF, which is constructed from five manual segmentations. Instead, De Leener et al. (2014) proposed a fully automated method for delineating the contour of the spinal cord, in T1- and T2-weighted MR images, by warping of a deformable cylindrical model.

The first significant effort to define and introduce a standard anatomical space for spinal cord neuroimaging studies relates to the work of Fonov et al. (2014), who developed a standard stereotactic space for spinal cord imaging data, between the vertebral levels of C1 and T6 (MNI-Poly-AMU template).

Their template is generated using the image registration algorithm presented in Avants et al. (2008) and includes a T2-weighted average image, together with probabilistic gray and white matter maps. Such tissue probability maps were developed by Taso et al. (2014), via automated registration of manually labelled MRI scans of 15 subjects.

Within the approach presented by Fonov et al. (2014), registration of new subjects into the MNI-Poly-AMU template space is obtained in a semi automated fashion. To further parcellate the spinal cord into different fiber tracts, Levy et al. (2015) generated a single slice spinal cord white matter template, obtained from digitalisation of existing anatomical atlases (Gray, 2009). Such a template has then been registered to MNI-Poly-AMU space.

The work of Fonov et al. (2014); Levy et al. (2015); Taso et al. (2014) constitutes an important step towards the development of robust and reliable tools for analysing structural spinal cord data. Indeed, having a common anatomical framework can potentially allow the comparison of results obtained by different research groups on different data sets, thus speeding up the progress of spinal cord imaging research.

Nevertheless, there is still a crucial limitation that has not been addressed: the tools that they have designed are optimised for spinal cord MR images and therefore they neglect the brain to the same extent to which most brain image processing environments neglect the spinal cord. As a result, integrating brain and spinal cord morphometric analyses is not yet feasible with their approaches.

1.6. Contribution of this thesis and summary of remaining chapters

The main contribution of this thesis is the formulation of a general Bayesian modelling framework, which exploits variational inference techniques to capture the variability of both shape and intensity across large data sets of MRI scans. In fact, the work presented here embraces the vision of Ashburner and Friston (2005) and expands the hierarchical structure of their generative model so as to allow learning of both average-shaped tissue probability maps and intensity priors from observed cross-sectional imaginga data sets. While in principle the proposed approach could have many potential applications in the field of medical imaging, this thesis focuses primarily on neuroimaging applications, with particular emphasis on the problem of developing an integrated processing framework for both brain and spinal cord image data.

Chapter 2 introduces, first, the very general principles underlying generative and discriminative modelling techniques, then presents an overview of some of the possible applications of probabilistic generative models to solve medical image processing problems.

Chapter 3 describes the generative model of MR neuroimaging data that constitutes the theoretical foundation of this thesis. The method is based on a spatially varying Gaussian mixture model (GMM) that includes unknown, average-shaped tissue probability maps (TPMs), which can be learned from the observed data. Model fitting is performed via both maximum likelihood and maximum a posteriori techniques.

Experiments to test the performance of the framework introduced in Chapter 3 are illustrated in Chapter 4. In particular, fully unsupervised and semisupervised learning are compared, along with different deformation models. Test data consist of both synthetic brain data and real brain and spinal cord data.

Chapter 5 presents some background concepts on the theory of variational Bayesian inference and introduces a modelling and algorithmic framework that applies the variational Bayes approach to medical image segmentation problems. This approach can alleviate some of the limitations of maximum likelihood and maximum a posteriori estimation, with no significant increase of the computational cost. Additionally, a procedure is described, which allows learning of empirical intensity priors from large MR data sets, thus increasing the robustness of the proposed tools, which are validated on both synthetic and real brain MR scans.

In Chapter 6, the variational approach introduced in the previous chapter is extended, in order to derive a semisupervised groupwise atlas construction framework, where morphological variability is modelled by means of diffeomorphisms. Application of such a framework to simultaneously perform brain and spinal cord morphometric analyses is explored, with validation experiments performed on real and synthetic MR images, mostly from publicly available databases.

Chapter 7 concludes this thesis, with a general discussion on the contribution of the presented work, on its limitations, as well as on possible directions for future work.

2

Generative models in medical imaging

2.1. Introduction

This chapter describes the fundamental concepts underlying generative modelling techniques. These are compared to the principles behind discriminative methods and the advantages and limitations of the two approaches are discussed.

A general overview of the possible applications of generative models to solve medical image analysis problems is also presented.

2.2. Generative versus discriminative models

By definition, generative models are statistical models that explicitly represent the probability distribution from which the observed (i.e. measured) data is assumed to be drawn. In other words, they capture the stochastic process underlying data generation, by means of probabilistic inference. This typically requires the introduction of unobserved random variables, referred to as latent variables, which correspond to hidden states of the modelled system (Bishop et al., 2007).

Having denoted the observed data by $\mathbf{x}$ and the hidden, or latent, data by $\mathbf{z}$, the core of a generative model is in its definition of the joint probability distribution $p(\mathbf{x}, \mathbf{z})$, which fully encodes the data generation process (Bishop, 2006).

Introducing a generic parametric representation, governed by a vector of parameters $\boldsymbol{\theta}$, the likelihood of the complete (i.e. observed and unobserved) data can be expressed as

$$p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) = p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta}) \ , \tag{2.1}$$

by making use of the product rule of probability theory.

Additionally, a prior probability distribution on the model parameters $p(\boldsymbol{\theta})$ can be incorporated, to give

$$p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta}) = p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \ . \tag{2.2}$$

By applying Bayes' rule it is possible to compute the posterior probability distribution of the unobserved latent variables and model parameters, given the observed data

$$p(\mathbf{z}, \boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta})}{p(\mathbf{x})} = \frac{p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta})}{\int\int p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta'})\mathrm{d}\boldsymbol{\theta'}\mathrm{d}\mathbf{z}} \ . \tag{2.3}$$

Indeed, this is usually one of the main quantities of interest for solving machine learning problems, since $\mathbf{z}$ commonly encodes an unknown property of the observed data (e.g. class labels) that the experimenter is trying to make predictions on (Bishop, 2006).

Equation 2.3 essentially indicates that the posterior probability over the unknown latent variables and model parameters is proportional to the joint distribution $p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta})$, as the term $p(\mathbf{x})$, referred to as evidence, does not depend on the unobserved variables.

Performing exact inference on $\mathbf{z}$ would require computing the posterior distribution $p(\mathbf{z}|\mathbf{x})$, which can be evaluated by marginalising $p(\mathbf{z}, \boldsymbol{\theta}|\mathbf{x})$ across the parameter space, as follows

$$p(\mathbf{z}|\mathbf{x}) = \int p(\mathbf{z}, \boldsymbol{\theta}|\mathbf{x})\mathrm{d}\boldsymbol{\theta} = \frac{\int p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}{\int\int p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta'})\mathrm{d}\boldsymbol{\theta'}\mathrm{d}\mathbf{z}} = \int p(\mathbf{z}|\boldsymbol{\theta}, \mathbf{x})p(\boldsymbol{\theta}|\mathbf{x})\mathrm{d}\boldsymbol{\theta} \ . \tag{2.4}$$

Unfortunately, the integrals in equations (2.4) and (2.3) are most often intractable in analytical form and too complex to solve numerically (Bishop, 2006). However, for many applications it is quite reasonable to assume that the posterior distribution $p(\boldsymbol{\theta}|\mathbf{x})$ is very peaked around its mode and therefore well approximated by a Dirac delta function centred on the maximum a posteriori estimate $\boldsymbol{\theta}_{MAP}$.

With this assumption

$$p(\mathbf{z}|\mathbf{x}) = \int p(\mathbf{z}|\boldsymbol{\theta}, \mathbf{x})p(\boldsymbol{\theta}|\mathbf{x})\mathrm{d}\boldsymbol{\theta} \approx p(\mathbf{z}|\boldsymbol{\theta}_{MAP}, \mathbf{x}) \ , \tag{2.5}$$

where

$$\boldsymbol{\theta}_{MAP} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{x}) \ , \tag{2.6}$$

that is to say, an approximate posterior probability on the hidden data $\mathbf{z}$ can be computed making use of point estimates of the model parameters, obtained by maximising the posterior $p(\boldsymbol{\theta}|\mathbf{x})$. In particular, since

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}, \boldsymbol{\theta})}{p(\mathbf{x})} = \frac{\int p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta})\mathrm{d}\mathbf{z}}{\int\int p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta'})\mathrm{d}\boldsymbol{\theta'}\mathrm{d}\mathbf{z}} \ , \tag{2.7}$$

maximising $p(\boldsymbol{\theta}|\mathbf{x})$ with respect to $\boldsymbol{\theta}$ is equivalent to maximising the joint probability distribution $p(\mathbf{x}, \boldsymbol{\theta})$.

As opposed to generative models, discriminative models do not represent the joint probability distribution of $\mathbf{x}, \mathbf{z}$ and $\boldsymbol{\theta}$. Instead, they make use of pairs of observed data vectors $\bar{\mathbf{x}} = \{\mathbf{x}_i\}$ and training hidden labels $\bar{\mathbf{z}} = \{\mathbf{z}_i\}$, to compute the probability of the labels given the input features and the model parameters $p(\bar{\mathbf{z}}|\bar{\mathbf{x}}, \boldsymbol{\theta})$.

The posterior $p(\boldsymbol{\theta}|\bar{\mathbf{x}}, \bar{\mathbf{z}})$ can then be expressed as

$$p(\boldsymbol{\theta}|\bar{\mathbf{x}}, \bar{\mathbf{z}}) = \frac{p(\bar{\mathbf{z}}|\bar{\mathbf{x}}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\bar{\mathbf{x}})}{p(\bar{\mathbf{z}}|\bar{\mathbf{x}})} = \frac{p(\bar{\mathbf{z}}|\bar{\mathbf{x}}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\bar{\mathbf{x}})}{\int p(\bar{\mathbf{z}}|\bar{\mathbf{x}}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\bar{\mathbf{x}})\mathrm{d}\boldsymbol{\theta}} \ , \tag{2.8}$$

where $p(\boldsymbol{\theta}|\bar{\mathbf{x}})$ is a prior on model parameters, as in (2.2).

For making predictions on unseen test data $\mathbf{x}$, the following posterior needs to be computed $p(\mathbf{z}|\mathbf{x}, \bar{\mathbf{x}}, \bar{\mathbf{z}})$, by integrating out the model parameters, as follows

$$p(\mathbf{z}|\mathbf{x}, \bar{\mathbf{x}}, \bar{\mathbf{z}}) = \int p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\bar{\mathbf{x}}, \bar{\mathbf{z}})\mathrm{d}\boldsymbol{\theta} \ . \tag{2.9}$$

Similarly to equations (2.3) and (2.4), also equations (2.8) and (2.9) involve the computation of integrals which are likely to be computationally very challenging. As a result MAP approximations are often preferred to a fully Bayesian learning scheme (Bishop et al., 2007).

A non-exhaustive list of discriminative techniques commonly used in machine learning includes logistic regression (Hosmer Jr et al., 2013), support vector machines (Hearst et al., 1998), linear regression (Neter et al., 1996), artificial neural networks (Yegnanarayana, 2009) and random forests (Breiman, 2001).

It has been shown that discriminative models can achieve very high predictive performance, as long as abundant training data is available. For instance, Jordan (2002) compared discriminative and generative learning as represented by logistic regression and naive Bayes to conclude that, for increasing training size, discriminative methods have a lower asymptotic prediction error, which they reach more slowly though, compared to generative techniques.

In other words, the main limitation of discriminative models is that they cannot be trained on unlabelled data (Bishop et al., 2007), thus making their performance heavily dependent on the number of labelled training examples.

This is unfortunately limiting the application of discriminative techniques for medical image analysis, since the availability of training labels is often quite scarce (Koch et al., 2015), as compared to the large amount of unlabelled data (Schmah et al., 2008). Therefore, as opposed to other fields of data science, where discriminative methods are now consolidated, in medical imaging these types of model have only recently started to be proposed (Ciresan et al., 2012; Fung and Stoeckel, 2007; Hoi et al., 2009; Li et al., 2006; Ronneberger et al., 2015; Yi et al., 2009; Zhang et al., 2015b), and nevertheless, *ad hoc* data augmentation strategies are often necessary to ensure success of these methods (Ciresan et al., 2012; Ronneberger et al., 2015).

With respect to that, generative models offer an interesting advantage, which derives directly from the fact that they explicitly model the relationship between observed and latent data, that is the possibility of augmenting the training set with unlabelled images. If no manual or ground truth labels are used for training of the model, the resulting computational learning scheme is said to be unsupervised. In the opposite case, that is to say when all training examples are provided with output labels, learning is instead fully supervised. Hybrid training strategies are also available, which combine labelled and unlabelled data, in a semisupervised fashion (Bishop et al., 2007; Kingma et al., 2014; Zhu, 2006).

However, the dichotomy between generative and discriminative models does not nec-

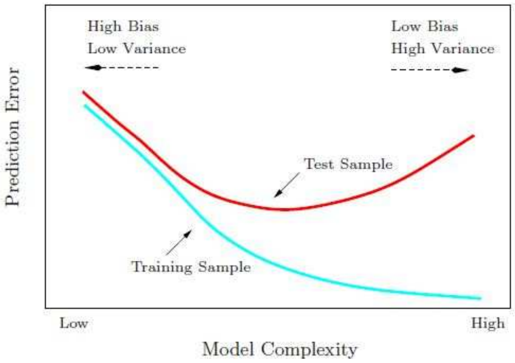


Figure 2.1: Prediction error as a function of model complexity. Qualitative illustration of the trade off between variance and bias from Friedman et al. (2001).

essarily imply that the two methods are mutually exclusive. There is, in fact, strong interest in the scientific community towards learning techniques that can combine the two approaches, thus yielding trade off solutions (Batmanghelich et al., 2012; Bosch et al., 2008; Jaakkola et al., 1999; Lasserre et al., 2006; McCallum et al., 2006; Raina et al., 2003), which ideally would provide more accurate predictions than pure generative approaches, with a lower amount of required labelled data, compared to pure discriminative methods. For instance, a possible approach involves synthesising data using a generative model and exploiting it to enhance the classification performance of a discriminative algorithm (Enzweiler and Gavrila, 2008). From a general perspective, such an effort could be seen as a possible way to cope with the bias-variance dilemma (Figure 2.1), which is a well known question arising in all classification problems, where simplistic models have poor predictive performance due to a high bias (underfitting), while overly complex models loose accuracy due to the high variance of their predictions (overfitting) (Bouchard and Triggs, 2004).

2.3. Generative models in MR imaging

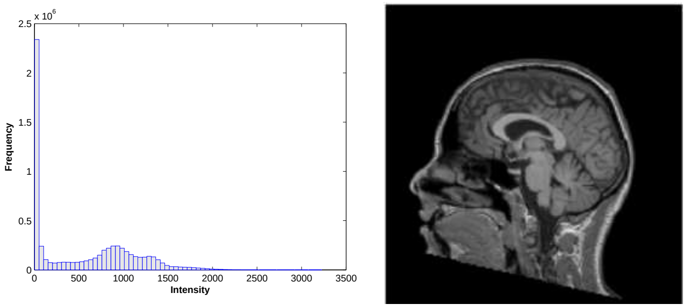
Generative models have been widely adopted to represent structural MRI data (Allassonnière et al., 2006; Ashburner and Friston, 2005; Cardoso et al., 2015; Gooya et al., 2012; Iglesias et al., 2011, 2012a; Lê et al., 2015; Maji and Bruchez, 2012; Menze et al., 2010; Pohl et al., 2006; Rajapakse and Kruggel, 1998; Sabuncu et al., 2010; Sharma et al., 2001; Sudre et al., 2015; Van Leemput et al., 2001; Wu et al., 2014a; Zhang and Fletcher, 2014; Zhang et al., 2001). In this framework, the most natural formulation considers voxel intensity values as observed data and voxel labels as hidden variables. Such labels can indicate tissue composition, membership of an anatomical structure or presence of a physiological or pathological feature. These models provide a very convenient framework for the development of automated image processing algorithms. Applications that have been explored so far include image segmentation, image registration, contrast synthesis and atlas construction. Hybrid generative and discriminative models have also been proposed (Batmanghelich et al., 2012; Tu et al., 2008), as an attempt to simultaneously maximise predictive performance and biological interpretability.

2.3.1. Generative models for image segmentation

Many generative models in medical imaging rely on learning a parametric (or non-parametric) probability density representation of the observed intensities. This is a relevant research problem, especially in MR imaging, since the observed image intensities are far from being standardised but depend heavily on the pulse sequence and acquisition parameters. Therefore, being able to capture the statistical properties of the observed data (see Figure 2.2 for an example of MR T1-weighted scan with its associated intensity histogram) and possibly relate them to the properties of previously seen data, is a critical question for the development of powerful MR image processing algorithms.

In this context, mixture models are particularly useful as they provide a natural framework for capturing the tissue specific properties of MRI signal intensities, by associating each mixture component, or a small number of them, to a specific tissue class. The Gaussian mixture model (GMM), in particular, has become established as a classical modelling framework for the quantitative analysis of MRI signal intensities. In fact, it represents a general and flexible approach to fit the intensity distribution of images and, for the same reason, it has been used profusely in computer vision, to model natural color images as well as video sequences (Belongie et al., 1998; Delignon et al., 1997; Friedman and Russell, 1997; Gupta and Sortrakul, 1998; Nikou et al., 2010).

Gaussian mixture models have a direct and rather intuitive application in medical image processing for the implementation of automated segmentation algorithms to



(a) (b)

Figure 2.2: Intensity histogram (a) of an MRI scan of the head (b)

identify tissue types or delineate anatomical structures. The accuracy and validity of such tools has been studied extensively, not only for classifying healthy brain tissues (Ashburner and Friston, 2005; Iglesias et al., 2012a; Liang et al., 1992; Rajapakse and Kruggel, 1998; Van Leemput et al., 1999c) but also for brain lesion and tumor segmentation (Menze et al., 2015; Prastawa, 2003; Sudre et al., 2015; Van Leemput et al., 2001).

Figure 2.4 shows the graphical representation of a simple Gaussian mixture model that could be used to segment MR images. The model in panel (a) makes use of global mixing proportions π, meaning that for every voxel j the prior probability of that voxel belonging to a certain tissue class does not depend on its spatial location. This assumption is normally too uninformative for imaging data, as there is a strong prior belief on where specific labels are more likely to occur. Therefore, many probabilistic segmentation methods make use of spatially varying mixing proportions π_j (Ashburner and Friston, 1997; Lorenzo-Valdés et al., 2004), as in panel (b) of Figure 2.4.

These local tissue weights encode population-specific information on anatomical variability; therefore they are often referred to as probabilistic atlases or tissue probability maps (Figure 2.3).

Another way of encoding prior anatomical information for solving image segmentation problems is provided by the so called multi-atlas label fusion framework (Rohlfing and Maurer, 2005; Wang et al., 2013). In this case rather than encoding information on anatomical variability in the form of probabilistic maps, the idea is to use a set of labelled images (atlases) as training examples in order to estimate the unknown labels

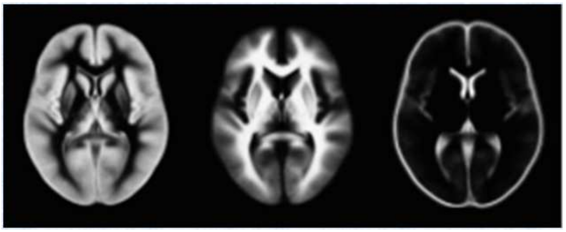


Figure 2.3: Example of probabilistic brain atlas generated as part of the International Consortium for Brain Mapping (ICBM) project (Mazziotta et al., 2001, 1995). From left to right gray matter, white matter and cerebrospinal fluid tissue probability maps are illustrated.

of the test data. For this purpose spatial (anatomical) correspondences between training and test data have to be estimated first and then used to propagate the training labels and intensities onto the test images. It should be noted that this algorithmic framework was not originally conceived as a probabilistic model fitting scheme (Rohlfing and Maurer, 2005), nevertheless probabilistic generative interpretations of this type of techniques have been proposed recently (Iglesias et al., 2012a; Sabuncu et al., 2010).

The number of tissue classes to be included in a generative model of MR data depends on the anatomical region as well as on the imaging modality. For brain imaging, the tissue types that are mostly of interest are gray matter, white matter and cerebrospinal fluid (at least for healthy subjects). However additional classes must also be included to model bone, soft tissues and air in the background as well as inside anatomical cavities, while pathological tissue types, such as tumour or lesion, may be added for patients data. Within the Gaussian mixture framework, an intuitive choice would be to associate each tissue class with one single Gaussian, however it turns out that for many tissues the distribution of intensities is more complex, due to both biological properties and and partial volume effects (Cardoso et al., 2011), therefore multiple Gaussians should generally be used to represent each tissue type, so that each tissue class is itself modelled as a Gaussian mixture (Ashburner and Friston, 2005).

The problem of how to determine the optimal number of Gaussian components is a typical model selection problem, where overly complex models should be avoided to prevent overfitting, while simplistic models might not be able to capture all the relevant patterns in the data, thus introducing large biases. Different model selection strategies have been proposed to solve this problem, such as the Bayesian information criterion

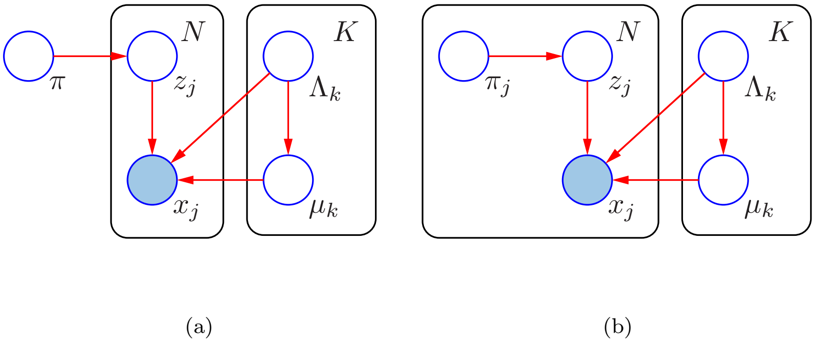


Figure 2.4: A simple graphical Gaussian mixture model for image segmentation. Large filled circles indicate the observed data (image intensities $\mathbf{X}$). Unfilled circles represent unobserved random variables (latent variables $\mathbf{Z}$, which encode class memberships, and model parameters Θ). The observed intensities are assumed to be drawn from a Gaussian mixture distribution consisting of K components with means $\{\boldsymbol{\mu}_k\}_{k=1,\ldots,K}$ and covariance matrices $\{\boldsymbol{\Sigma}_k\}_{k=1,\ldots,K}$. The model in panel (a) uses global mixing proportions $\boldsymbol{\pi}$, as opposed to that in panel (b), which has local mixing weights $\{\boldsymbol{\pi}_j\}_{j=1,\ldots,N}$.

(BIC) (Sudre et al., 2015), the minimum message length (MML) criterion (Wu et al., 2003), the alternating kernel and mixture (AKM) method (Priebe and Marchette, 2000). Such a topic will be discussed in greater detail in Chapter 5, where it will be shown how the variational Bayes framework can be exploited to automatically select the optimal number of classes.

2.3.2. Generative models for image registration and atlas construction

Image registration and atlas estimation problems can also be formulated in Bayesian generative framework (Allassonnière and Kuhn, 2010; Allassonnière et al., 2007; Ashburner and Friston, 2009; Risholm et al., 2010; Van Leemput, 2009; Zhang and Fletcher, 2014; Zhang et al., 2013; Zöllei et al., 2007a). In fact, optimal deformation fields, which align multiple images to an unbiased group average, can be elegantly obtained as MAP estimates within a Bayesian inference setting. In such a case, the negative log likelihood function of the data (image intensities) acts as a dissimilarity or distance metric, while priors on the deformations implement the regularisation, so as to encourage smoothness of the spatial transformations. The resulting log posterior function is therefore a Bayesian equivalent of the classical energy objective functions, which have been thoroughly used to solve image registration problems.

In particular, under the hypothesis of independent and identically distributed Gaussian noise, the corresponding Gaussian likelihood function leads to the sum of squared differences (SSD) distance metric (Zhang et al., 2013), which is commonly used to estimate templates having the same image contrast as the training data. Furthermore the work of Zöllei et al. (2003) demonstrates that the mutual information objective function for image registration (Maes et al., 1997) has a local optimum about the point of correct alignment under a generative latent variable model of the observed intensities.

Formulations of this sort have been proposed both in the small deformation (Allassonnière and Kuhn, 2010; Loic le Folgoc, 2016; Risholm et al., 2010; Simpson et al., 2015, 2012; Zöllei et al., 2007a) and diffeomorphic setting (Vialard et al., 2012; Zhang and Fletcher, 2014; Zhang et al., 2013).

Within such models, the variables encoding the deformation fields are treated as unobserved random variables, which should ideally be integrated out from the model by marginalisation (Risholm et al., 2010; Zhang et al., 2013). In fact, the performance of common mode approximations has been questioned for image registration problems, especially in heavy noise conditions (Allassonnière et al., 2007; Iglesias et al., 2012b). Unfortunately, integration under the true posterior of the deformations is a particularly challenging task, due to the lack of analytical solutions and to the high dimensionality of non-linear deformation models (Allassonnière and Kuhn, 2010). The effectiveness of sampling techniques, like Markov chain Monte Carlo methods (MCMC), to solve such problems has been investigated in Allassonnière and Kuhn (2010); Iglesias et al. (2013c); Risholm et al. (2010); Zhang et al. (2013), but further work needs to be done in order to make such techniques computationally convenient, compared to standard MAP methods.

Lately some authors have proposed generative mixture models (Allassonnière and Kuhn, 2010; Zhang et al., 2015a) to simultaneously solve the problems of template construction and image clustering. This requires the incorporation of additional latent random variables, encoding a cluster label for each observed image. In other words, each observed image is considered to be generated by applying a stochastic deformation field to a template image, drawn from a set of morphometrically distinct templates.

2.3.3. Other applications of generative models

One of the advantages of generative modelling techniques is that a single model might be suitable for solving multiple processing tasks, without the need to design task specific models, which will have, by definition, poor generalisation capability. An interesting example can be found in Cardoso et al. (2015), where the authors show how the same generative model of multimodal MR data can be exploited for both segmentation and image synthesis applications with minimal adjustments.

Along the same line, performing the correction of intensity inhomogeneities together with image segmentation in a probabilistic generative setting has become established as standard practice within the neuroimaging community (Ashburner and Friston, 2005; Greenspan et al., 2006; Peng et al., 2006; Pohl et al., 2005; Van Leemput et al., 1999d; Wells III et al., 1996; Zhang et al., 2001).

These types of method, which are implemented in a number of publicly available image processing softwares (Ashburner and Friston, 2005; Fischl, 2012; Smith et al., 2004), have proved effective and computationally convenient (Hou, 2006). They typically integrate the bias estimation procedure within an expectation-maximisation framework, where the computations to obtain class labels (i.e. segmentations) and intensity distribution parameters (e.g. Gaussian mixture parameters) are interleaved with optimisation of the bias, in an iterative fashion. Moreover it should be noted that constraints to enforce smoothness of the estimated non-uniformity fields can be easily incorporated in such a Bayesian framework, through the introduction of appropriate prior models (Ashburner and Friston, 2005).

Interestingly, it has recently been shown that also a widely used, non-probabilistic bias correction scheme, N3 (Sled et al., 1998), can be interpreted as the implementation of a generative model of MRI data (Larsen et al., 2015).

2.4. Summary

In this chapter, the general principles behind generative and discriminative modelling techniques have been discussed. Additionally, some of the possible applications of generative models in medical imaging have been introduced.

The following chapter will present in detail a generative model of MR neuroimaging data, which can be used to solve the problem of constructing probabilistic, average-shaped tissue templates from large cross-sectional data sets.

3

A Bayesian framework for groupwise atlas construction

3.1. Introduction

This chapter will introduce the modelling framework that constitutes the foundation of the work within this thesis. Such a model represents MR imaging data from a generative perspective, for the purpose of capturing the variability of both shape and image intensity across large MRI data sets.

The most direct application of this framework is related to the construction of anatomical tissue probability maps, which is a problem that arises naturally in medical image computing, for example when structural data is used to perform group morphometric analyses.

The topic of template construction and its importance for medical imaging applications will be introduced in Section 3.2 from a general point of view. Section 3.3 instead will present details on the mathematical formulation of the model underlying the proposed method, whereas the computational strategy adopted to estimate the model parameters and the resulting algorithm will be discussed in Section 3.4.

Finally, the main limitations of the presented work will be discussed in Section 3.5, along with possible directions for future work, some of which will be explored in the following chapters.

3.2. Constructing anatomical atlases: motivations and challenges

The impressive growth of interest towards neuroscience that has occurred during the last fifty years has been accompanied and sustained by massive collection of neuroscientific data, ranging from the molecular and cellular scales to the macroscopic level. In this context of neuroscientific information proliferation, imaging techniques have been widely exploited for the *in vivo* investigation of brain anatomy and physiology (Mazziotta et al., 1995), due to their non- (or minimally) invasive nature. As a result, the development of tools for the automated processing, or mining, of neuroimaging data has become, and remains, a critical research topic.

Among the challenges that arise when working with neuroimaging data, there is one, which is nearly ubiquitous, that is having to deal with wide morphological variability across individuals, as well as across populations (Toga and Thompson, 2000).

In functional imaging studies, this translates into the necessity to map common functional activation sites onto individual anatomical scans (DeYoe et al., 1994). More generally, inter-subject anatomical variability represents a crucial factor to be taken into account when performing statistical group analyses or comparing experimental results coming from different laboratories for meta-analysis (Laird et al., 2011; Mazziotta et al., 2001).

For this reason, the neuroimaging community has put considerable effort into the construction of digital brain atlases, whose natural application is to provide a population-based stereotactic space, for spatially normalising data, whenever there is a need to compensate for individual shape differences (Ashburner and Friston, 2000; Friston et al., 1995; Mechelli et al., 2005).

Nevertheless, this fact is not just a drawback for the neuroimaging community. Indeed, if it is possible to construct models that allow to compensate for shape differences, this means that the same models of imaging data can be used to investigate intra-population anatomical variability and inter-population shape variations (Thompson et al., 2000; Xu et al., 2014). Analyses of this sort could eventually serve to answer clinically relevant questions, for example by offering decisional support for distinguishing

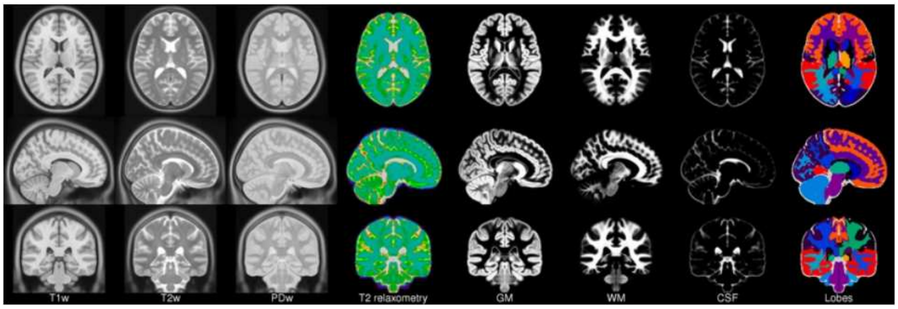


Figure 3.1: Non-linear ICBM 152 (International Consortium for Brain Mapping) T1-, T2- and PD-weighted models. The templates include also T2 relaxometry maps, tissue probability maps and a lobe atlas. Additional information can be found at `http://www.loni.usc.edu/atlases`*.*

normal anatomical features from pathological ones.

From a technical point of view, another interesting aspect is that, when an atlas is warped to match data of a single subject, any form of information stored in the atlas, which might regard for example tissue composition (Ashburner and Friston, 2005), cytoarchitecture (Eickhoff et al., 2005), vascular architecture or neurochemical content, is automatically projected onto the particular individual anatomy (Thompson et al., 2000). This is, for instance, the very general principle behind the many atlas-based segmentation methods developed during the past few decades (Aljabar et al., 2009; Cuadra et al., 2004; Lawes et al., 2008).

Atlas-guided computerised segmentation is indeed well established, as a technique to perform tissue or structure classification, in an automated fashion. The validity and robustness of the currently available tools has been investigated and assessed by many authors (Cabezas et al., 2011; Collins and Evans, 1997; Wang et al., 2005).

With respect to this, a fundamental question arises though, as to how an ideal atlas should be constructed. This is in fact far from being a purely theoretical problem, as the atlas generation procedure directly impacts segmentation or classification results. A quantitative evaluation of the influence of atlas construction and selection methods on the performance of segmentation algorithms can be found in Aljabar et al. (2009); Avants et al. (2010); Rohlfing et al. (2004); Zöllei et al. (2007b).

A number of approaches have been proposed so far to address the problem of constructing population-based atlases (Ashburner and Friston, 2009; Avants and Gee, 2004;

Bhatia et al., 2004; De Craene et al., 2004; Guimond et al., 2000; Joshi et al., 2004; Ribbens et al., 2010; Shattuck et al., 2008; Thompson and Toga, 1997; Van Leemput, 2009; Wang et al., 2005; Xu et al., 2014). A critical aspect, as recognized in many previous works, is that of avoiding a bias in the shape of the template (Lorenzen et al., 2005), a circumstance which typically occurs when a reference anatomy is chosen *a priori* (Thompson et al., 2000). Such an issue is commonly solved by introducing a hidden reference space and by formulating the optimisation problem in terms of a simultaneous, groupwise estimation of the set of transformations that minimise a distance measure between the atlas and the individual images (Balci et al., 2007; Bhatia et al., 2004; Joshi et al., 2004; Mahapatra, 2013), while ensuring that the warps are as small as possible (Avants et al., 2010).

As with most image registration problems, additional constraints (i.e. regularisation) on the deformations have to be introduced in order to preserve as many topological properties as possible (Simpson et al., 2012; Stefanescu et al., 2004). From a probabilistic perspective this is equivalent to preventing or penalising implausible and overly complex solutions.

Various similarity measures, as well as different deformation models have been proposed by different authors. Moreover some works aim at constructing templates that represent an average shape (Avants and Gee, 2004) or an average shape and intensity (Ashburner et al., 1999; Bhatia et al., 2004; Guimond et al., 2000; Joshi et al., 2004; Sabuncu et al., 2008), while others provide methods for estimating probabilistic tissue maps (Ashburner and Friston, 2009; De Craene et al., 2004; Kuklisova-Murgasova et al., 2011; Petrovic et al., 2007; Ribbens et al., 2010; Shattuck et al., 2008; Van Leemput, 2009; Xu et al., 2014). Exemplars of brain atlases, created as an initiative of the International Consortium for Brain Mapping (ICBM), are illustrated in Figure 3.1

Finally, a research topic, which is closely related to the construction of anatomical atlases, is the problem of structural image clustering, whose aim is to identify subgroups of individuals sharing common morphological features. Such a problem was explored, for instance, in the work of Sabuncu et al. (2008) and Ribbens et al. (2010), but will not be addressed in this thesis.

3.3. Generative groupwise model of MR data

The remainder of this chapter will present a computational framework that can serve to learn tissue probability maps (TPMs) from large data sets of multispectral MR images.

The method relies on the formulation of a single generative groupwise model, where observed image intensities are assumed to be drawn from Gaussian mixture distributions. In particular, the standard Gaussian mixture model is adapted in order to incorporate unknown deformable tissue priors, which can be learned from training data, either in a completely unsupervised or in a semisupervised manner. Intensity non-uniformity correction is also performed, within the same framework, by modelling the bias field as a combination of low spatial frequency basis functions.

In practice, the method introduced here will enable tissue classification (i.e. segmentation), atlas construction, bias field correction and image registration to be performed simultaneously in the same computational framework.

Treating image segmentation and registration within a single statistical model is an approach that has already been explored by a number of authors (Ashburner and Friston, 2005; DAgostino et al., 2006; Pohl et al., 2006; Wyatt and Noble, 2003; Xiaohua et al., 2004a; Xu et al., 2014; Yezzi et al., 2001). Indeed, numerous experimental findings support the underlying hypothesis that solving the two problems in a coupled manner benefits the results of both (Ashburner and Friston, 2005; DAgostino et al., 2006; Pohl et al., 2006). An additional advantage is that methods developed according to this unifying perspective tend to be general and powerful enough to deal with a wider range of applications, compared to most bottom up approaches.

Nonetheless integrated approaches of this sort have been mainly exploited for the processing of individual images rather then in the context of groupwise prior learning. The algorithms proposed in Bhatia et al. (2007); Petrovic et al. (2007); Ribbens et al. (2010); Riklin-Raviv et al. (2010) are among the few ones that extend this approach to the modelling of population data, for the purpose of constructing average-shaped tissue templates.

Estimation of optimal model parameter is formulated, in this chapter, as a mixed maximum likelihood (ML) and maximum a posteriori (MAP) problem. Therefore, an expression for the joint probability of the data and the model parameters will be derived in following paragraphs of this section, whereas the optimisation scheme, adopted for maximising such probability function, will be illustrated in Section 3.4.

3.3.1. Distribution of image intensities

Each image is treated as function of space $f_j : \Omega_j \to \mathbb{R}$, where the domain Ω_j is a compact subset $\Omega_j \subset \mathbb{R}^3$. If different image contrasts are available for the same subject then $f_j : \Omega_j \to \mathbb{R}^D$, with D equal to the number of imaging modalities.

Let us first consider data of one subject and let us denote by $\mathbf{x}_j$ a D-dimensional vector of signal intensities at voxel j, with $j \in \{1, \dots, N\}$. As demonstrated by numerous works in the relevant literature (Ashburner and Friston, 2005; Iglesias et al., 2012a; Liang et al., 1992; Rajapakse and Kruggel, 1998; Van Leemput et al., 1999c; Zhang et al., 2001), such intensities can be modelled as being drawn from a Gaussian mixture distribution of K components, having mean vectors $\Theta_\mu = \{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K\}$ and covariance matrices $\Theta_\Sigma = \{\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_K\}$.

Such an assumption is valid only for large signal-to-noise ratios (SNR). At lower SNR (i.e. when the ratio between signal intensity and noise standard deviation approaches one), the magnitude of the MR signal can no longer be modelled by a Gaussian probability density function but becomes Rician distributed instead (Aja-Fernández and Tristán-Vega, 2013). This results from the non-linear nature of the transformation applied to compute image magnitudes from the raw complex data, which, in the absence of MR signal, consists of Gaussian noise with zero mean and uncorrelated real and imaginary parts (Gudbjartsson and Patz, 1995). In regions where signal is zero, such as the air-filled background, noise can be modelled by a Rayleigh distribution, which is a special case of the Rician distribution.

Having denoted by π_k the prior probability of any signal intensity, irrespective of its value and location, being generated from class k, the joint probability of observing $\mathbf{x}_j$ and voxel j belonging to class k, can be computed, by making use of Bayes' rule, as follows

$$p(\mathbf{x}_j, z_{jk} = 1) = \pi_k \, \mathcal{N}(\mathbf{x}_j | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \prod_{c=1}^{K} \left[\pi_c \, \mathcal{N}(\mathbf{x}_j | \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)\right]^{z_{jc}} \ , \tag{3.1}$$

where $\mathbf{z}_j$ is a K-dimensional discrete latent variable, encoding class memberships, whose scalar components are equal to

$$z_{jc} = \begin{cases} 1, & \text{if } c = k \ . \\ 0, & \text{otherwise} \ . \end{cases} \tag{3.2}$$

A model of this sort can be applied to each image of a cross-sectional data set that includes M subjects, under the assumption that data acquired from different individuals have different intensity distributions, due to the lack of consistency of conventional MR signal intensities across different scans (Jovicich et al., 2009). As a result, for each subject i, with $i \in \{1, \ldots, M\}$, different mean vectors $\{\boldsymbol{\mu}_{i1}, \ldots, \boldsymbol{\mu}_{iK}\}$ and covariance matrices $\{\boldsymbol{\Sigma}_{i1}, \ldots, \boldsymbol{\Sigma}_{iK}\}$ have to be introduced.

A standard Gaussian mixture model would make use of global mixing proportions $\Theta_\pi = \{\pi_1, \ldots, \pi_K\}$ as in equation (3.1). On the contrary, the approach adopted in this chapter involves defining local, unknown mixing coefficients, that is to say, for each voxel j and class k, a spatial tissue prior $\pi_{jk} \in [0, 1]$ is introduced, which represents the prior probability of any signal, at a spatial location indexed by j, being drawn from class k.

Thus, neglecting for now the problem of template warping in order to account for individual anatomical differences, and assuming that all data points are independent, the log likelihood function for the entire data set can be written as

$$\mathcal{J} = \log \prod_{i=1}^{M} \prod_{j=1}^{N} p\left(\mathbf{x}_{ij} | \Theta_\pi, \Theta_\mu, \Theta_\Sigma\right) = \sum_{i=1}^{M} \sum_{j=1}^{N} \log \left(\sum_{k=1}^{K} \pi_{jk} \ p\left(\mathbf{x}_{ij} | \boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik}\right) \right) \ . \tag{3.3}$$

3.3.2. Modelling intensity non-uniformities

As discussed Chapter 1, an artifact very commonly found in MR images is the one referred to as intensity non-uniformity, or bias field. It consist of a smooth, low frequency signal that does not originate from magnetic tissue properties but instead is caused by factors such as the inhomogeneity of the radio frequency pulse, the disuniformity of reception coil sensitivity, eddy currents induced by field gradients, as well as the electromagnetic interaction between the scanned object and the RF field (Sled and Pike, 1998).

As a result, any intensity-based model of MRI data should take into account the presence of such a distortion. This is particularly important for automated segmentation methods. Many studies have in fact demonstrated that the accuracy of segmentation algorithms can be considerably enhanced when a correction for intensity non-uniformity is applied (Ashburner and Friston, 2005; Dawant et al., 1993; Held et al., 1997).

Therefore, for each subject i, a multiplicative bias vector field $\mathbf{b}(\boldsymbol{\beta}_i)$ is introduced, where $\boldsymbol{\beta}_i$ encodes a set of parameters. Each of the D components of the bias represents a non-uniformity field, which corrects the image of the corresponding channel, and is modelled as the exponential of a linear combination of three dimensional discrete cosine transform (DCT) basis functions. Only a small number of low frequency basis functions are considered, in order to ensure spatial smoothness of the resulting field.

Equation (3.1) can therefore be rewritten as

$$p(\mathbf{x}_{ij}, c = k) = \det\left(\mathbf{B}_{ij}\right) \pi_{jk} \mathcal{N}\left(\mathbf{B}_{ij}\mathbf{x}_{ij} | \boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik}\right) \ , \tag{3.4}$$

where $\mathbf{B}_{ij} = \operatorname{diag}\left(\mathbf{b}_j\left(\boldsymbol{\beta}_i\right)\right)$ and $\mathbf{b}_j(\boldsymbol{\beta}_i)$ is a D-dimensional vector denoting the bias at voxel j for subject i.

The objective function $\mathcal{J}$ in (3.3) becomes instead

$$\mathcal{J} = \sum_{i=1}^{M} \sum_{j=1}^{N} \log \left(\det(\mathbf{B}_{ij}) \sum_{k=1}^{K} \pi_{jk} \ p\left(\mathbf{B}_{ij}\mathbf{x}_{ij} | \boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik}\right) \right) \ . \tag{3.5}$$

3.3.3. Deformable anatomical priors

To account for anatomical variability, the prior class membership probabilities, denoted by $\{\boldsymbol{\pi}_j\}_{j=1,\dots,N}$, can be warped, in order to match the morphology of each subject.

Thus, from a modelling point of view, the Gaussian mixing proportions become, for every subject, functions of a different coordinate mapping, which specifies correspondences between the voxel centres of that subject's volume and a set of locations in the space of the atlas.

Each of such coordinate transformations is controlled by a vector of parameters $\mathbf{a}$, so that, for a population of M subjects, all the mappings together define a set $\Theta_a = \{\mathbf{a}_1, \dots, \mathbf{a}_M\}$.

Many transformation models have been explored to solve medical image registration problems (McInerney and Terzopoulos, 1996) and potentially all of them could be integrated in the modelling framework that is being presented. As to be expected, the choice on what family of models to prefer is dependent, to some extent, on the type of application. For example, the number of degrees of freedom of the deformations should vary according to the amount of morphological variability present in the data (Denton et al., 1999), which is in turn dependent on the represented anatomical structure, as well as on the homogeneity of the population of interest, in terms of age, ethnicity, health status etc.

The availability of computational resources and time is also a non-negligible factor. Obviously, very complex representations, which are likely to provide more accurate results, will most often require a longer processing time and a larger memory cost (Wollny and Kruggel, 2002). In particular, it should be noted that this aspect is particularly relevant for those technological solutions that are intended for integration in the clinical routine, as speed and robustness become priorities in that situation (Otake et al., 2012; Rueckert et al., 2016).

Different deformation models are explored and compared in this thesis. In particular, the formulation presented in this chapter makes use of affine transformations in combination with a non-linear small deformation model.

Rigid body and affine transformations

Rigid body and affine transformations are linear functions that map between two spaces while preserving straight lines and planes. Indeed, rigid body transforms are a subset of affine transforms, with a lower number of degrees of freedom (in three-dimensional space six instead of twelve). Therefore, the mathematical treatment of these two types of transformations is, except for a different number of free parameters, very similar (Jenkinson and Smith, 2001).

Let us first define the identity transform on a continuous domain $I_d : \Omega \rightarrow \Omega$

$$I_d(\boldsymbol{y}) = \boldsymbol{y},\ \forall \boldsymbol{y} \in \Omega \tag{3.6}$$

where Ω is a compact subset of $\mathbb{R}^3$.

Since digital images can be thought of as continuous functions, sampled on a discrete domain, the following notation [1] will be used to indicate the coordinates of the geometric

[1] Throughout this manuscript different font styles will be used to distinguish continuous and discrete vector fields. For instance, $\mathbf{y}$ is a discrete field obtained by sampling of the continuous field $\boldsymbol{y}$.

centre of each voxel j

$$\mathbf{y}_j = \begin{bmatrix} \mathrm{y}_{j,1} & \mathrm{y}_{j,2} & \mathrm{y}_{j,3} \end{bmatrix}^T . \tag{3.7}$$

An affine transformation $\Delta : \Omega \to \Omega$ is fully described by a transformation matrix $\mathbf{T}(\mathbf{a})$, parameterised by $\mathbf{a} \in \mathbb{R}^{12}$, so that

$$\Delta(\mathbf{y}_j) = \mathbf{y}'_j(\mathbf{a}) = \mathbf{T}(\mathbf{a}) \cdot \begin{bmatrix} \mathbf{y}_j \\ 1 \end{bmatrix} , \tag{3.8}$$

where $\mathbf{y}'_j$ identifies the location, in the space of the moving image, which corresponds to $\mathbf{y}_j$ in the static (i.e. target) image. It should be noted that, for the model adopted in this work, the moving images are the tissue probability maps, while the target images are the individual scans. Moreover, $\mathbf{T}(\mathbf{a})$ is constructed as the exponential map of a matrix $\boldsymbol{Q}(\mathbf{a}) \in \mathfrak{ga}(\mathit{3})$ (Ashburner and Ridgway, 2013), where $\mathfrak{ga}(\mathit{3})$ represents the Lie algebra [2] of the 3D affine group $GA(3)$.

$$\boldsymbol{Q}(\mathbf{a}) = \begin{bmatrix} a_7 & a_6 + a_{10} & -a_5 + a_{11} & a_1 \\ -a_6 + a_{10} & a_8 & a_4 + a_{12} & a_2 \\ a_5 + a_{11} & -a_4 + a_{12} & a_9 & a_3 \\ 0 & 0 & 0 & 1 \end{bmatrix} . \tag{3.9}$$

For matrix Lie groups, the exponential map coincides with matrix exponential, therefore $\mathbf{T}(\mathbf{a})$ can be computed as

$$\mathbf{T}(\mathbf{a}) = \exp\left(\boldsymbol{Q}(\mathbf{a})\right) = \sum_{k=0}^{\infty} \frac{(\boldsymbol{Q}(\mathbf{a}))^k}{k!} . \tag{3.10}$$

The parameters controlling a rigid body transform are $\{a_1, a_2, a_3\}$ for the translational component and $\{a_4, a_5, a_6\}$ for the rotational component, while affine transformations also allow zooms, governed by $\{a_7, a_8, a_9\}$, and shears, controlled by $\{a_{10}, a_{11}, a_{12}\}$. Such a formulation, by using the concept of matrix exponential and exponential map, permits linear treatment of the parameters, which are defined in a vector space tangent to the identity element of the group. In turn this allows a rigorous mathematical definition of the notions of shape average and shape distance (Woods, 2003). For instance,

[2] Any Lie group (smooth, differentiable manifold) $\mathcal{G}$ can be associated with a Lie algebra $\mathfrak{g}$, which is a tangent vector space that captures the local structure of the group. In the case of real matrix groups, the Lie algebra $\mathfrak{g}$ consists of those matrices $\boldsymbol{Q}$ for which $\exp(x\boldsymbol{Q}) \in \mathcal{G}$ for all real numbers x, where exp is the exponential map.

within this framework, unbiasedness of a group average requires that the set of affine parameters across a population sum up to zero.

Having introduced affine deformable tissue priors, the log likelihood function of equation (3.5) must be reformulated as follows

$$\mathcal{J} = \sum_{i=1}^{M}\sum_{j=1}^{N} \log\left(\det(\mathbf{B}_{ij})\sum_{k=1}^{K}\pi_k(\mathbf{y}'_j(\mathbf{a}_i))\; p\left(\mathbf{B}_{ij}\mathbf{x}_{ij}|\boldsymbol{\mu}_{ik},\boldsymbol{\Sigma}_{ik}\right)\right)\ . \tag{3.11}$$

In particular, the computation of $\pi_k(\mathbf{y}'_j(\mathbf{a}_i))$ requires two stages. First the set of transformed coordinate vectors $\{\mathbf{y}'_j\}_{j\in\{1,\ldots,N\}}$ must be evaluated, according to equation (3.8), then the discrete tissue priors $\{\boldsymbol{\pi}_k\}_{k=1,\ldots,K}$ have to be be interpolated and resampled. In the work presented here a trilinear interpolation scheme is adopted. In fact, as opposed to higher order approaches, the linear approach ensures that the warped tissue priors satisfy the following constraint

$$\begin{gathered}\forall i\in\{1,\ldots,M\},\ \forall j\in\{1,\ldots,N\},\ \forall k\in\{1,\ldots,K\},\\ \pi_k(\mathbf{y}'_j(\mathbf{a}_i))\in[0,1]\wedge\sum_{k=1}^{K}\pi_k(\mathbf{y}'_j(\mathbf{a}_i))=1\ ,\end{gathered} \tag{3.12}$$

as long as the tissue probability maps are normalised in their native space parametrisation.

Non-linear small deformations

Affine transformations are low dimensional deformation models, which allow compensating for the variability in object positioning (via translations and rotations) and for limited global shape and size differences (via zooming and shearing).

For an accurate matching of anatomical structures, which are generally non-rigid and morphologically variable across individuals, or within the same subject in the presence of physiopathological dynamical processes, the affine approach is most often inadequate, in spite of being robust and efficient from a computational point of view (Crum et al., 2014). With respect to this, evidence will be provided in the next chapter, where the results of affine template construction experiments will indicate the need to resort to higher order models for the purpose of simultaneously aligning brain and spinal cord data. Indeed, even for capturing solely pose and size variations, the affine model is not sufficient in this case, as it is not capable of encoding head flexion and extension movements.

A class of higher order non-linear deformation models, which have been widely exploited in the field of medical image processing, for the purpose of image alignment, are the so called small deformation models (Johnson and Christensen, 2002).

A small deformation field $\boldsymbol{\phi} : \Omega \rightarrow \Omega$ is defined as

$$\boldsymbol{\phi}(\boldsymbol{y}) = I_d + \boldsymbol{u}(\boldsymbol{y}) \ , \tag{3.13}$$

with $\Omega \in \mathbb{R}^3$ and $||\boldsymbol{u}|| \ll \epsilon$, $\forall \boldsymbol{y} \in \Omega$. In other words, the mapping $\boldsymbol{\phi}$ is obtained by adding a small vector field $\boldsymbol{u}$ to the identity transform I_d.

Even though larger deformation fields would in principle produce lower registration residuals, the constraint $||\boldsymbol{u}|| \ll \epsilon$ cannot be relaxed with such a simple additive model, as described by equation (3.13), without sacrificing biophysical plausibility of the warps (Christensen et al., 1996). In fact, assuming that $\boldsymbol{u}$ is not a constant vector field, only for sufficiently small displacements, the following model constitutes an acceptable approximation to compute inverse deformation fields

$$\boldsymbol{\psi}(\boldsymbol{y}) = \boldsymbol{\phi}^{-1}(\boldsymbol{y}) \approx I_d - \boldsymbol{u}(\boldsymbol{y}) \ , \tag{3.14}$$

that is to say

$$\boldsymbol{\phi} \circ \boldsymbol{\psi} = (I_d + \boldsymbol{u}) \circ (I_d - \boldsymbol{u}) \approx I_d \approx (I_d - \boldsymbol{u}) \circ (I_d + \boldsymbol{u}) = \boldsymbol{\psi} \circ \boldsymbol{\phi} \ , \tag{3.15}$$

where $\circ$ denotes the composition operation.

The larger the displacement fields, the less accurate the approximation in equations (3.14) and (3.15) become. The reason why invertibility is such a highly desirable property in the context of medical image registration (Chun and Fessler, 2008) is primarily that non invertible deformation fields, in addition to being less elegant mathematical objects, can disrupt topological properties, for example by causing folds or tears, and to introduce modelling biases, by capturing certain deformation trajectories systematically better than others, for instance sensitivity might be higher to detect shrinkage rather than growth effects, or *vice versa* (Cachier and Rey, 2000).

For this reason, application of almost all non-linear registration techniques must be preceded by an initial affine, or at least rigid body, alignment step. In fact, even within large deformation settings, such as the LDDMM framework (Beg et al., 2005), higher computational stability and faster convergence can be ensured if the input data is in rough alignment prior to model estimation. In alternative, an overall deformation model can be defined

$$\boldsymbol{\xi}(\boldsymbol{y}) = \mathbf{T}(\mathbf{a}) \cdot \begin{bmatrix} \boldsymbol{y} + \boldsymbol{u}(\boldsymbol{y}) \\ 1 \end{bmatrix} , \tag{3.16}$$

by composing a discrete small non-linear deformation field $\boldsymbol{u}$ and an affine transformation, which can be jointly optimised, so as to make sure that optimal alignment is achieved by means of the smallest possible non-linear displacement field $\boldsymbol{u}$. The model of equation (3.16) is adopted both in this chapter and in the following one. With such an approach the data does not need to be affine registered prior to model fitting and the resulting algorithm can be applied with minimal pre-processing of the input scans. Indeed, the only pre-processing step that might be required is intra-subject coregistration of the different modalities, which here are assumed to be already in alignment.

3.3.4. Regularising the model

One of the advantages from adopting a Bayesian modelling perspective is that prior knowledge on the variability of model parameters can be easily incorporated by making use of Bayes' rule. This involves defining suitable prior probability distributions, which summarise information acquired through previous experiments, or observations, and inform during the process of statistical inference when new experimental data is analysed.

Point estimates of model parameters can therefore be obtained by maximising the posterior probability distribution $p(\Theta|\mathbf{X}) \propto p(\mathbf{X}|\Theta)p(\Theta)$, with $p(\mathbf{X}|\Theta)$ being the likelihood of the observed data and $p(\Theta)$ the prior term. This approach is well known as maximum a posteriori (MAP) estimation, and it results in a trade-off between maximising adherence of the model to the experimental data, as in a pure maximum likelihood fashion, and finding solutions that agree as much as possible with the priors.

From a numerical programming perspective, such an approach is equivalent to providing the objective function with an additional penalty term, whose function is to prevent unreasonable or undesirable parameter values (Williams, 1995). In other words, maximum a posteriori estimation is a regularised form of maximum likelihood and therefore it becomes essential to solve ill-posed mathematical problems, which in the context of data modelling most often arise as inverse problems, while additionally improving numerical stability (Davies and Anderssen, 1986).

Intensity non-uniformity field

With regard to intensity non-uniformities, the underlying assumption that the bias field is a low spatial frequency, smooth signal can for example be enforced by introducing a regularisation term $\mathcal{R}$ based on the squared euclidean norm of the Laplacian of the bias (Fan et al., 2003)

$$\mathcal{R} = \int_{\Omega} \|\Delta_{\boldsymbol{y}} \boldsymbol{b}\|^2 \, \mathrm{d}\boldsymbol{y} = \int_{\Omega} \left(\sum_{l=1}^{D} \left(\sum_{d=1}^{3} \frac{\partial^2 b_l}{\partial {y_d}^2} \right)^2 \right) \mathrm{d}\boldsymbol{y} \; , \tag{3.17}$$

where $\boldsymbol{b} : \Omega \rightarrow \mathbb{R}^D$ is a continuous D-dimensional vector field, whose components, indexed by $l \in \{1, \ldots, D\}$, represent intensity non-uniformities for each imaging modality.

An equivalent discretised version of equation (3.17) can be derived by first sampling $\boldsymbol{b}$ over a regular lattice, to give the discrete vector field $\mathbf{b}$ and then making use of a finite difference approximation to compute the second derivatives. Having denoted by $\mathbf{L}_b$ a sparse Toeplitz matrix representing the discrete three dimensional Laplacian operator, $\mathcal{R}$ can be expressed as

$$\mathcal{R} = \sum_{l=1}^{D} \|\mathbf{L}_b \mathrm{b}_l\|^2 = \sum_{l=1}^{D} (\mathbf{L}_b \mathrm{b}_l)^T \mathbf{L}_b \mathrm{b}_l = \sum_{l=1}^{D} \mathrm{b}_l^T \mathbf{L}_b^T \mathbf{L}_b \mathrm{b}_l \; . \tag{3.18}$$

The sparse matrix $\mathbf{L}_b$ con be obtained as

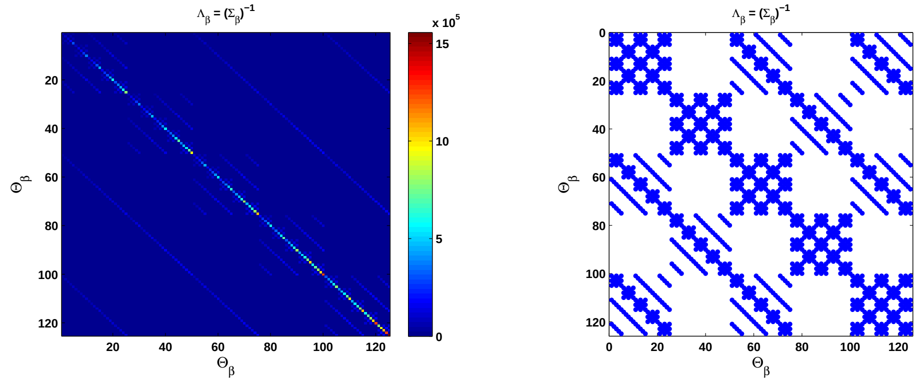
$$\mathbf{L}_b = \mathbf{D_z} \otimes \mathbf{I_y} \otimes \mathbf{I_x} + \mathbf{I_z} \otimes \mathbf{D_y} \otimes \mathbf{I_x} + \mathbf{I_z} \otimes \mathbf{I_y} \otimes \mathbf{D_x} \; . \tag{3.19}$$

where $\mathbf{I_x}, \mathbf{I_y}$ and $\mathbf{I_z}$ are identity matrices of appropriate size while $\mathbf{D_x}, \mathbf{D_y}$ and $\mathbf{D_z}$ represent one dimensional discrete differential operators, which allow computing central difference approximations of the second derivatives along the three Cartesian axes. The symbol $\otimes$ indicates the Kronecker product.

For the model of intensity inhomogeneities adopted here, which uses discrete cosine transform (DCT) basis functions, the regularisation term can be expressed as a function of the bias field parameters by

$$\mathcal{R}(\Theta_\beta) = \sum_{l=1}^{D} \sum_{i=1}^{M} \boldsymbol{\beta}_{il}^T \boldsymbol{\Phi}^T \mathbf{L}_b^T \mathbf{L}_b \boldsymbol{\Phi} \boldsymbol{\beta}_{il} = \frac{1}{2} \sum_{l=1}^{D} \sum_{i=1}^{M} \boldsymbol{\beta}_{il}^T \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} \boldsymbol{\beta}_{il} \; , \tag{3.20}$$

where $\Theta_\beta = \{\boldsymbol{\beta}_i\}_{i=1,\ldots,M}$ with i being an index over subjects and $\boldsymbol{\Phi}$ is a matrix of three dimensional DCT basis functions. Such a matrix, given the separable nature of the



(a)

(b)

Figure 3.2: Prior precision matrix (a) for implementing a Gaussian regularisation of the bias field, which penalises the Euclidean norm of the Laplacian, and its corresponding sparsity pattern (b).

discrete cosine transform, can be obtained as follows

$$\mathbf{\Phi} = \boldsymbol{\phi}_{\mathbf{z}} \otimes \boldsymbol{\phi}_{\mathbf{y}} \otimes \boldsymbol{\phi}_{\mathbf{x}} \ , \tag{3.21}$$

with $\boldsymbol{\phi}_{\mathbf{x}}, \boldsymbol{\phi}_{\mathbf{y}}, \boldsymbol{\phi}_{\mathbf{z}}$ representing matrices of one dimensional basis functions.

The regularisation matrix $\mathbf{\Sigma}_{\boldsymbol{\beta}}^{-1}$, which is illustrated in Figure 3.2, in a Bayesian setting would be interpreted as a prior precision matrix. In fact, the quadratic form in (3.20) is equivalent, except for the presence of an additive constant, to the negative logarithm of a multivariate normal distribution. The computation of $\mathbf{\Sigma}_{\boldsymbol{\beta}}^{-1}$ can be carried out much more efficiently if the high dimensional quadratic form $\mathbf{\Phi}^T \mathbf{L}_b^T \mathbf{L}_b \mathbf{\Phi}$ is further expanded by exploiting the mixed product property of the Kronecker product, to give the following lower dimensional decomposition

$$\begin{aligned}\mathbf{\Sigma}_{\boldsymbol{\beta}}^{-1} = {} & \boldsymbol{\phi}_{\mathbf{z}}^T \mathbf{D}_{\mathbf{z}}{}^T \mathbf{D}_{\mathbf{z}} \boldsymbol{\phi}_{\mathbf{z}} \otimes \boldsymbol{\phi}_{\mathbf{y}}^T \boldsymbol{\phi}_{\mathbf{y}} \otimes \boldsymbol{\phi}_{\mathbf{x}}^T \boldsymbol{\phi}_{\mathbf{x}} + \boldsymbol{\phi}_{\mathbf{z}}^T \boldsymbol{\phi}_{\mathbf{z}} \otimes \boldsymbol{\phi}_{\mathbf{y}}^T \mathbf{D}_{\mathbf{y}}{}^T \mathbf{D}_{\mathbf{y}} \boldsymbol{\phi}_{\mathbf{y}} \otimes \boldsymbol{\phi}_{\mathbf{x}}^T \boldsymbol{\phi}_{\mathbf{x}} + \\ & \boldsymbol{\phi}_{\mathbf{z}}^T \boldsymbol{\phi}_{\mathbf{z}} \otimes \boldsymbol{\phi}_{\mathbf{y}}^T \boldsymbol{\phi}_{\mathbf{y}} \otimes \boldsymbol{\phi}_{\mathbf{z}}^T \mathbf{D}_{\mathbf{z}}{}^T \mathbf{D}_{\mathbf{z}} \boldsymbol{\phi}_{\mathbf{z}} + 2 \boldsymbol{\phi}_{\mathbf{z}}^T \mathbf{D}_{\mathbf{z}}{}^T \boldsymbol{\phi}_{\mathbf{z}} \otimes \boldsymbol{\phi}_{\mathbf{y}}^T \mathbf{D}_{\mathbf{y}} \boldsymbol{\phi}_{\mathbf{y}} \otimes \boldsymbol{\phi}_{\mathbf{x}}^T \boldsymbol{\phi}_{\mathbf{x}} + \\ & 2 \boldsymbol{\phi}_{\mathbf{z}}^T \mathbf{D}_{\mathbf{z}}{}^T \boldsymbol{\phi}_{\mathbf{z}} \otimes \boldsymbol{\phi}_{\mathbf{y}}^T \boldsymbol{\phi}_{\mathbf{y}} \otimes \boldsymbol{\phi}_{\mathbf{x}}^T \mathbf{D}_{\mathbf{x}} \boldsymbol{\phi}_{\mathbf{x}} + 2 \boldsymbol{\phi}_{\mathbf{z}}^T \boldsymbol{\phi}_{\mathbf{z}} \otimes \boldsymbol{\phi}_{\mathbf{y}}^T \mathbf{D}_{\mathbf{y}}{}^T \boldsymbol{\phi}_{\mathbf{y}} \otimes \boldsymbol{\phi}_{\mathbf{x}}^T \mathbf{D}_{\mathbf{x}} \boldsymbol{\phi}_{\mathbf{x}} \ .\end{aligned} \tag{3.22}$$

In other words, this is equivalent to assuming that the *a priori* probability distribution of the bias field parameters is a multivariate normal distribution $\mathcal{N}(0, \mathbf{\Sigma}_{\boldsymbol{\beta}})$. With such a model however it is not possible to control the stiffness of the bias.

Deformation models

In the context of image registration, which is an inherently ill-posed problem (Fischer and Modersitzki, 2008; Modersitzki, 2004), the main purpose of regularisation is to enforce biophysical plausibility of the deformation fields. Indeed, this is a very complex and multidisciplinary research topic, which acts like a bridge between mathematical and biological sciences, given the impact that regularisation has on the biological interpretability of image registration results, especially around areas of low intensity contrast (Ciardo et al., 2013; Fischer and Modersitzki, 2008). Therefore, a lot of research has been conducted to determine suitable mathematical formulations, capable of preserving the topology of anatomical structures, while realistically capturing the underlying morphometric changes (Ashburner, 2007; Christensen and Johnson, 2001; Noblet et al., 2005).

Rigid models represent probably the only exception, since in that case having a penalty term is not strictly necessary, given the explicitly constrained nature of the transformations. For affine models instead, a simple Gaussian regularisation, that is to say assuming that the vector of parameters is *a priori* normally distributed, $\mathbf{a} \sim \mathcal{N}(0, \mathbf{\Sigma}_a)$, helps to prevent implausible scaling and skewing and at the same time it benefits numerical stability during the optimisation process.

The prior model adopted in the remainder of this chapter to regularise non-linear deformations is heavily based on the work of Ashburner (2007), even if it should be noted that the parametrisation adopted here is based on a small deformation approach, as opposed to the diffeomorphic formulation of Ashburner (2007). However, irrespectively of the particular form of regularisation adopted, the objective function of equation (3.11) can be reformulated as the logarithm of a joint probability, to give

$$\begin{aligned}
\mathcal{F}(\Theta) &= \log p(\Theta_\beta, \Theta_a, \Theta_u | \mathbf{X}, \Theta_\pi, \Theta_\mu, \Theta_\Sigma) \\
&= \log \left[p(\mathbf{X} | \Theta_\pi, \Theta_\mu, \Theta_\Sigma, \Theta_\beta, \Theta_a, \Theta_u) p(\Theta_\beta) p(\Theta_a) p(\Theta_u) \right] \\
&= \log p(\mathbf{X} | \Theta_\pi, \Theta_\mu, \Theta_\Sigma, \Theta_\beta, \Theta_a, \Theta_u) + \log p(\Theta_\beta) + \log p(\Theta_a) + \log p(\Theta_u) \\
&= \mathcal{J}(\Theta) + \mathcal{R}(\Theta) + \text{const} \ ,
\end{aligned} \tag{3.23}$$

with $\mathcal{J}(\Theta)$ being the log likelihood of the observed data. Here the entire parameter set has been denoted by $\Theta = \{\Theta_\pi, \Theta_\mu, \Theta_\Sigma, \Theta_\beta, \Theta_a, \Theta_u\}$ with $\{\Theta_\pi, \Theta_\mu, \Theta_\Sigma\}$ representing the Gaussian mixture parameters (i.e. mixing proportions, means and covariances),

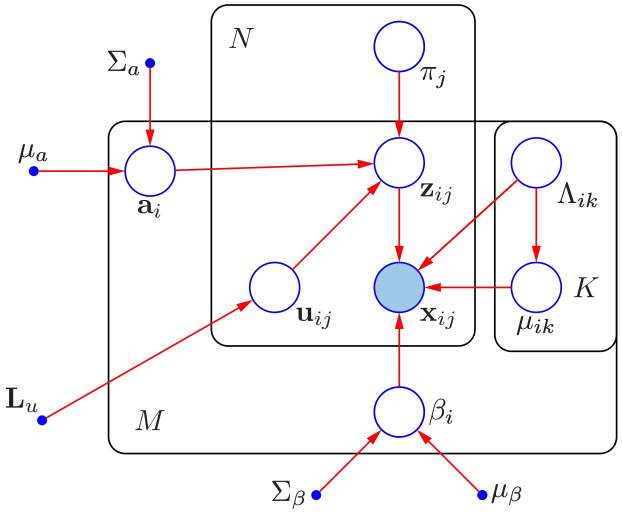


Figure 3.3: Directed acyclic graph representing the generative model discussed in this chapter. Filled circles indicate the observed data while unfilled circles represent unobserved random variables (latent variables $\mathbf{Z}$*, which encode class memberships, and model parameters* Θ*). Blue dots correspond to fixed hyperparameters, which are not estimated during model fitting.*

Θ_β the bias field parameters, Θ_a and Θ_u the affine and small deformation parameters respectively. A directed acyclic graph corresponding to the generative model presented in this chapter is illustrated in Figure 3.3.

3.4. Model fitting

As to be expected, given the complexity of the model presented here, finding a closed form solution to the problem of maximising $\mathcal{F}$ (or $\mathcal{J}$) is not possible. In principle, this would involve solving the system of equations obtained by differentiating $\mathcal{F}$ with respect to Θ and setting these derivatives to zero. In practice though, even for the canonical Gaussian mixture model, explicit solutions of the maximum likelihood (or maximum a posteriori) problem do not exist. For such a model, parameter estimation can be performed very efficiently using the expectation-maximisation (EM) algorithm, which is an optimisation strategy that combines an elegant probabilistic formulation, highly stable convergence and, quite often, a relatively low computational cost (Dempster et al., 1977). The main idea behind the EM algorithm is to optimise a log likelihood function in an iterative manner, by alternating between estimating a posterior distribution on the unobserved variables (E-step) and maximising (M-step) a lower bound on the log likelihood with respect to the model parameters.

For the model introduced in this chapter, the problem is slightly more complex as the Gaussian mixture, bias field and deformation parameters cannot be estimated simultaneously in a pure EM fashion. In cases of this sort, a natural approach consists in trying to reduce the complexity of an intractable optimisation problem by replacing it with multiple, simpler subproblems, corresponding to conditional estimations. In other words, since the model parameters strongly depend on one another, it is convenient to adopt a computational scheme that breaks down the problem into a number of separate constrained optimisations. This involves partitioning the parameter set into subsets and iteratively updating each one of them, while keeping the others fixed at their current estimates.

A general optimisation scheme, which relies on this sort of strategy, has been rigorously formulated in Meng and Rubin (1993), within an expectation-maximisation framework. The resulting algorithm is named expectation-conditional-maximisation (ECM), since it replaces a complicated maximisation step with a series of conditional optimisations, and, interestingly, it possesses convergence properties very similar to those of the EM algorithm. The ECM framework constitutes actually a special case of the generalised EM (GEM) algorithm. In fact, the GEM approach, rather than maximising a lower bound on the log likelihood during the M-step, seeks instead parameter values that improve such a lower bound, without necessarily maximising it (Neal and Hinton, 1998).

The computational scheme that will be presented in detail in the remainder of this section is indeed an ECM algorithm and it can be summarised by the following pseudocode (Algorithm 1)

```
Input: A cross-sectional data set of MR images $\mathbf{X}$
Output: Estimates of model parameters $\hat{\Theta}$
begin
    initialise model parameters $\Theta$;
    while objective function $\mathcal{F}$ has not converged do
        for each subject $i$ do
            for $iter = 1..., I_n$ do
                E-step:
                compute sufficient statistics;
                M-step:
                update $\{\boldsymbol{\mu}_{i1}, \ldots, \boldsymbol{\mu}_{iK}\}$ ;
                update $\{\boldsymbol{\Sigma}_{i1}, \ldots, \boldsymbol{\Sigma}_{iK}\}$;
                for $iter = 1..., I_\beta$ do
                    update $\boldsymbol{\beta}_i$;
                end
                for $iter = 1..., Iter_a$ do
                    update $\mathbf{a}_i$;
                end
                for $iter = 1..., Iter_u$ do
                    update $\mathbf{u}_i$;
                end
            end
        end
        update $\{\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_K\}$;
    end
end
```

Algorithm 1: optimisation algorithm for generating population-based tissue probability maps

3.4.1. Estimating the Gaussian mixture parameters

As discussed above, the parameters of a Gaussian mixture model can be conveniently estimated using the expectation-maximisation (EM) scheme. The EM algorithm is a general optimisation technique, which can be used to find maximum likelihood (or maximum a posteriori) solutions, for probabilistic models that make use of latent variables to explain the observed data. In the case of Gaussian mixtures, latent variables $\{\mathbf{z}_j\}_{j\in 1,\dots,N}$ are used to encode membership of the observed data points $\{\mathbf{x}_j\}_{j\in 1,\dots,N}$ with respect the K model components. Therefore $\mathbf{z}_j$ can be expressed as a K-dimensional binary variable.

As anticipated, the likelihood function $\mathcal{J} = \log p(\mathbf{X}|\Theta)$ cannot be maximised in closed form, but interestingly, the optimisation problem becomes considerably easier if, instead of considering the likelihood of the observed data, the problem is shifted towards maximising the joint probability of the observed and unobserved variables, given the model parameters (Bishop, 2006). The EM algorithm takes advantage of this circumstance, by defining a lower bound $\mathcal{L}$ on the objective function $\mathcal{J}$, which is computed making use of the complete data log likelihood. Therefore, it yields a much easier optimisation. In practice, this leads to an iterative computational scheme, which loops over generating a lower bound, given the current estimates of the model parameters, and updating the parameters, by assigning them values that maximise the current lower bound.

To derive such a lower bound $\mathcal{L}$, let us first decompose the likelihood function, as follows

$$\log p(\mathbf{X}|\Theta) = \log p(\mathbf{X},\mathbf{Z}|\Theta) - \log p(\mathbf{Z}|\mathbf{X},\Theta)\ . \tag{3.24}$$

If an arbitrary distribution $q(\mathbf{Z})$ over the set of latent variables $\mathbf{Z}$ is introduced, the previous equality can be reformulated as

$$\log p(\mathbf{X}|\Theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\Theta) - \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{Z}|\mathbf{X},\Theta)\ , \tag{3.25}$$

or equivalently as

$$\log p(\mathbf{X}|\Theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log\left(\frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})}\right) + \sum_{\mathbf{Z}} q(\mathbf{Z}) \log\left(\frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X},\Theta)}\right)\ . \tag{3.26}$$

The second term on the right end side of equation (3.26) is the Kullback-Leibler divergence [3] $D_{KL}(p\|q)$ between $q(\mathbf{Z})$ and the posterior distribution over the latent variables $p(\mathbf{Z}|\mathbf{X},\Theta)$.

Since $D_{KL}(q\|p) \geq 0$, equation (3.26) is the proof that a lower bound[4] on the likelihood function is given by

$$\mathcal{L}(q,\Theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left(\frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \right) , \tag{3.27}$$

where the same result could have also been derived from

$$\begin{aligned} \log p(\mathbf{X}|\Theta) = \log \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\Theta) &= \log \sum_{\mathbf{Z}} q(\mathbf{Z}) \left(\frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \right) \\ &\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left(\frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \right) , \end{aligned} \tag{3.28}$$

by applying Jensen's inequality.

The EM algorithm is an iterative procedure, consisting of two stages. In the first one, namely E-step, the functional $\mathcal{L}$ is maximised with respect to the function $q(\mathbf{Z})$.

A closer examination of equation (3.26) should indicate that this variational optimisation problem is almost straightforward. In fact, since the log likelihood does not depend on $q(\mathbf{Z})$, and since the Kullback-Leibler divergence is non-negative, maximising $\mathcal{L}(q,\Theta^{(n)})$, with respect to $q(\mathbf{Z})$, corresponds to minimising the Kullback-Leibler divergence between $q(\mathbf{Z})$ and the posterior distribution $p(\mathbf{Z}|\mathbf{X},\Theta^{(n)})$, where n indicates the current iteration. A global maximum of the lower bound occurs, in particular, when $D_{KL}(p\|q) = 0$. Therefore the solution of the E-step, at iteration n, is given by

$$q^{(n+1)}(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X},\Theta^{(n)}) . \tag{3.29}$$

In the subsequent M-step, a new lower bound $\mathcal{L}(q(\mathbf{Z})^{(n+1)},\Theta)$ is maximised with respect to Θ, that is to say

$$\Theta^{(n+1)} = \arg \max_{\Theta} \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta^{(n)}) \log \left(\frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{p(\mathbf{Z}|\mathbf{X},\Theta^{(n)})} \right) . \tag{3.30}$$

It is easy to prove that

$$\Theta^{(n+1)} = \arg \max_{\Theta} \mathcal{Q}(\Theta,\Theta^{(n)}) , \tag{3.31}$$

[3] The Kullback-Leibler divergence is a non-negative information theoretic measure that indicates the proximity of two probability distributions, nonetheless, because of its non-symmetric nature, it should not be considered as a proper distance metric.

[4] Such a lower bound also plays a crucial role in the neuroscientific theory of active inference (Friston, 2010), where it is referred to as variational free-energy.

with

$$\mathcal{Q}(\Theta, \Theta^{(n)}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta^{(n)}) \log p(\mathbf{X}, \mathbf{Z}|\Theta) \ . \tag{3.32}$$

The function $\mathcal{Q}(\Theta, \Theta^{(n)})$ represents the expectation of the log likelihood of the complete data $\{\mathbf{X}, \mathbf{Z}\}$, under the posterior probability distribution of the latent variables, which was computed in the previous E-step.

For the generative model presented in this chapter the E-step involves computing the posterior distribution of the tissue labels, given the observed image intensities and the current estimates of the model parameters $\Theta^{(n)}$, that is

$$\begin{aligned} q^{(n+1)}(\mathbf{z}_{ij}) &= p(\mathbf{z}_{ij}|\mathbf{x}_{ij}, \Theta^{(n)}) = p(z_{ijk} = 1|\mathbf{x}_{ij}, \Theta^{(n)}) \\ &= \prod_{c=1}^{K} \left(\frac{p(\mathbf{x}_{ij}, \mathbf{z}_{ij}|\Theta^{(n)})}{\sum_{\mathbf{z}} p(\mathbf{x}_{ij}, \mathbf{z}_{ij}|\Theta^{(n)})} \right)^{z_{ijc}} \\ &\overset{\circ}{=} \prod_{c=1}^{K} (\gamma_{ijk})^{z_{ijc}} \ . \end{aligned} \tag{3.33}$$

where

$$z_{ijc} = \begin{cases} 1, & \text{if } c = k \ . \\ 0, & \text{otherwise} \ . \end{cases} \tag{3.34}$$

The joint probability of $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ can be computed making use of Bayes' rule

$$p(\mathbf{x}_{ij}, \mathbf{z}_{ij}|\Theta^{(n)}) = p(\mathbf{z}_{ij}|\Theta_{\pi}^{(n)}, \Theta_{a}^{(n)}, \Theta_{u}^{(n)}) \ p(\mathbf{x}_{ij}|\mathbf{z}_{ij}, \Theta_{\mu}^{(n)}, \Theta_{\Sigma}^{(n)}, \Theta_{\beta}^{(n)}) \ , \tag{3.35}$$

where Θ_π denotes the tissue priors, Θ_μ and Θ_Σ the Gaussian means and covariances, Θ_β the bias field parameters, Θ_a and Θ_u the affine and non-linear deformation parameters respectively.

Recalling that prior probabilities over the latent variables are given by

$$p(\mathbf{z}_{ij}|\Theta_{\pi}^{(n)}, \Theta_{a}^{(n)}, \Theta_{u}^{(n)}) = \prod_{c=1}^{K} \left(\pi_c^{(n)}(\mathbf{y}'_j(\mathbf{a}_i, \mathbf{u}_{ij})) \right)^{z_{ijc}} \ , \tag{3.36}$$

and that the conditional distribution of the observed image intensities, given the hidden labels, is equal to

$$p(\mathbf{x}_{ij}|\mathbf{z}_{ij}, \Theta_{\mu}^{(n)}, \Theta_{\Sigma}^{(n)}, \Theta_{\beta}^{(n)}) = \prod_{c=1}^{K} \left(\det(\mathbf{B}_{ij}) \ \mathcal{N}(\mathbf{B}_{ij}\mathbf{x}_{ij}|\boldsymbol{\mu}_{ic}^{(n)}, \boldsymbol{\Sigma}_{ic}^{(n)}) \right)^{z_{ijc}} \ , \tag{3.37}$$

it is easy to prove that

$$\gamma_{ijk} = \frac{\det(\mathbf{B}_{ij}) \ \pi'_{jk} \ p\left(\mathbf{B}_{ij}\mathbf{x}_{ij}|\boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik}\right)}{\sum_{c=1}^{K} \det(\mathbf{B}_{ij}) \ \pi'_{jc} \, p\left(\mathbf{B}_{ij}\mathbf{x}_{ij}|\boldsymbol{\mu}_{ic}, \boldsymbol{\Sigma} ic\right)} \ , \tag{3.38}$$

where, to unclutter notation, the superscripts indicating iteration number have been omitted and $\pi_k(\mathbf{y}'_j(\mathbf{a}_i, \mathbf{u}_{ij}))$ has been denoted by π'_{jk}.

It should be noted that the prior (3.36) and posterior (3.33) probability distributions of $\mathbf{z}_{ij}$ take the same functional form.

For the entire dataset, the distribution $q(\mathbf{Z})$ is computed by

$$q(\mathbf{Z}) = p(\mathbf{Z}|\, \mathbf{X}, \Theta) = \prod_{i=1}^{M} \prod_{j=1}^{N} \prod_{k=1}^{K} (\gamma_{ijk})^{z_{ijk}} \tag{3.39}$$

In the following M-step the parameters are updated, by maximising the expectation of the complete data log likelihood with respect to the posterior distribution of the latent variables. As already discussed, for the problem addressed here, the M-step can be conveniently broken down into multiple conditional sub-stages. That is to say, first the values of the bias field and deformation parameters are kept fixed to their current estimates and the Gaussian mixture parameters are computed. Secondly, the new Gaussian mixture parameters are retained, as well as the deformations, and the lower bound is maximised with respect to the bias field parameters. Finally, the deformation parameters are updated.

The log likelihood function for the complete data can be easily obtained from (3.35), under the assumption that data points corresponding to different voxels are independent, to give

$$\log p(\mathbf{X}, \mathbf{Z}|\Theta) = \sum_{i=1}^{M} \sum_{j=1}^{N} \sum_{k=1}^{K} z_{ijk} \Big(\log \left(\det(\mathbf{B}_{ij})\ \pi'_{jk}\ p\left(\mathbf{B}_{ij}\mathbf{x}_{ij} | \boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik}\right) \right) \Big) \ . \tag{3.40}$$

The expected value of $\{\mathbf{z}_{ij}\}_{i,j}$ under the estimated variational posterior distribution q, is given by

$$\mathbb{E}_q[z_{ijk}] = 0 \cdot (1 - q(\mathbf{z}_{ij})\big|_{z_{ijk}=1}) + 1 \cdot q(\mathbf{z}_{ij})\big|_{z_{ijk}=1} = \gamma_{ijk}\ , \tag{3.41}$$

with γ_{ijk} often being referred to as responsibility of class k for the observed data $\mathbf{x}_{ij}$.

Finally, the expectation of the complete data log likelihood, which must be maximised in the M-step, can be computed as follows

$$\mathcal{Q}(\Theta) = \sum_{i=1}^{M} \sum_{j=1}^{N} \sum_{k=1}^{K} \gamma_{ijk} \Big(\log \left(\det(\mathbf{B}_{ij})\, \pi'_{jk}\, p\left(\mathbf{B}_{ij}\mathbf{x}_{ij} | \boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik}\right) \right) \Big) \ . \tag{3.42}$$

Differentiating with respect to $\boldsymbol{\mu}_{ik}$

$$\begin{aligned}\frac{\partial \mathcal{Q}}{\partial \boldsymbol{\mu}_{ik}} &= \frac{\partial}{\partial \boldsymbol{\mu}_{ik}} \left(\sum_{j=1}^{N} \gamma_{ijk} \left\{ \log \left[\det(\mathbf{B}_{ij})\, \pi'_{jk}\, p\left(\mathbf{B}_{ij}\mathbf{x}_{ij} | \boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik}\right) \right] \right\} \right) \\ &= \frac{\partial}{\partial \boldsymbol{\mu}_{ik}} \left(\sum_{j=1}^{N} \gamma_{ijk} \left\{ \log \left[\mathcal{N}\left(\mathbf{B}_{ij}\mathbf{x}_{ij} | \boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik}\right) \right] \right\} \right) \\ &= \sum_{j=1}^{N} \gamma_{ijk} \boldsymbol{\Sigma}_{ik}^{-1} \left(\mathbf{B}_{ij}\mathbf{x}_{ij} - \boldsymbol{\mu}_{ik} \right) \ ,\end{aligned} \tag{3.43}$$

and solving

$$\frac{\partial \mathcal{Q}}{\partial \boldsymbol{\mu}_{ik}} = 0 \ , \tag{3.44}$$

gives the following closed form solution

$$\boldsymbol{\mu}_{ik} = \frac{\sum_{j=1}^{N} \gamma_{ijk} (\mathbf{B}_{ij}\mathbf{x}_{ij})}{\sum_{j=1}^{N} \gamma_{ijk}} \ , \tag{3.45}$$

which represents the update rule for the Gaussian means.

Similarly for the covariance matrices, setting

$$\frac{\partial \mathcal{Q}}{\partial \boldsymbol{\Sigma}_{ik}} = 0 \ , \tag{3.46}$$

and solving with respect to $\boldsymbol{\Sigma}_{ik}$, gives

$$\boldsymbol{\Sigma}_{ik} = \frac{\sum_{j=1}^{N} \gamma_{ijk} (\mathbf{B}_{ij}\mathbf{x}_{ij} - \boldsymbol{\mu}_{ik})(\mathbf{B}_{ij}\mathbf{x}_{ij} - \boldsymbol{\mu}_{ik})^T}{\sum_{j=1}^{N} \gamma_{ijk}} \ . \tag{3.47}$$

Computing an update expression for the mixing proportions $\{\boldsymbol{\pi}_k\}_{k=1,\ldots,K}$ requires first the complete data log likelihood (3.42) to be transformed into the atlas coordinate space. Because this is an integral function over the native domains of the M images, the change of variable does not involve just a simple substitution, but also a scaling of the integrand by the determinant of the Jacobian matrix of the coordinate transformation. In particular, having denoted the spatially normalised responsibilities by γ'_{ijk} and making use of a Lagrange multiplier to enforce the constraint

$$\forall j, \ \sum_{k} \pi_{jk} = 1 \ , \tag{3.48}$$

the following update rule can be derived (Ashburner and Friston, 2009; Bishop, 2006)

$$\pi_{jk} = \frac{\sum_{i=1}^{M} \gamma'_{ijk} \det(\mathbf{J}_j^{\boldsymbol{\xi}^{-1}})}{\sum_{i=1}^{M} \sum_{c=1}^{K} \gamma'_{ijc} \det(\mathbf{J}_j^{\boldsymbol{\xi}^{-1}})} \ , \tag{3.49}$$

where

$$\gamma'_{ijk} = \gamma_{ik}(\boldsymbol{\xi}^{-1}(\mathbf{y}_j)) \ , \tag{3.50}$$

and

$$\boldsymbol{\xi}(\boldsymbol{y}) = (\Delta(\mathbf{a}_i) \circ (I_d + \boldsymbol{u}_i)) \ , \tag{3.51}$$

while $\mathbf{J}_j^{\boldsymbol{\xi}^{-1}}$ is the Jacobian tensor field of $\boldsymbol{\xi}^{-1}$ evaluated at voxel j.

In summary in the M-step, the Gaussian mixture parameters can be updated making use of the observed data, through the following sufficient statistics, given by zeroth, first and second order moments, weighted by the responsibilities each tissue class

$$\begin{aligned} N_{ik} &= \sum_{j=1}^{N} \gamma_{ijk} \ , \\ \mathbf{m}_{ik} &= \sum_{j=1}^{N} \gamma_{ijk} \mathbf{B}_{ij} \mathbf{x}_{ij} \ , \\ \mathbf{S}_{ik} &= \sum_{j=1}^{N} \gamma_{ijk} (\mathbf{B}_{ij} \mathbf{x}_{ij})(\mathbf{B}_{ij} \mathbf{x}_{ij})^T \ . \end{aligned} \tag{3.52}$$

3.4.2. Estimating the bias field

Estimation of the bias field parameters can be performed by constraining the Gaussian mixture parameters $\{\Theta_\pi, \Theta_\mu, \Theta_\Sigma\}$ and the deformation parameters $\{\Theta_a, \Theta_u\}$ to remain fixed at their current estimates while the bias field parameters Θ_β get updated.

Within a maximum likelihood formulation, this involves finding, for each $i \in \{1, \ldots, M\}$, an estimator $\hat{\boldsymbol{\beta}}_i^{ML}$ such that

$$\hat{\boldsymbol{\beta}}_i^{ML} = \arg \max_{\boldsymbol{\beta}_i} \mathcal{J}(\mathbf{X}, \Theta) \ , \tag{3.53}$$

with

$$\mathcal{J}(\mathbf{X}, \Theta) = \sum_{i=1}^{M} \sum_{j=1}^{N} \log \left(\det(\mathbf{B}_{ij}) \sum_{k=1}^{K} \pi'_{jk} \, p \left(\mathbf{B}_{ij} \mathbf{x}_{ij} | \boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik} \right) \right) \ . \tag{3.54}$$

Similarly for maximum a posteriori estimation, the aim is to maximise the following

$$\mathcal{F}(\mathbf{X}, \Theta) = \mathcal{J}(\mathbf{X}, \Theta) + \log p(\Theta_\beta) + \text{const} \ , \tag{3.55}$$

with respect to Θ_β.

Rather than computing ML (or MAP) estimators by direct maximisation of $\mathcal{J}$ (or $\mathcal{F}$), the EM approach provides an alternative strategy, which consist in iteratively optimising a lower bound $\mathcal{L}$ on $\mathcal{J}$ (or $\mathcal{F}$), as defined in (3.27), by maximisation of the auxiliary function reported in (3.42), with respect to Θ_β.

Unfortunately, as opposed to the Gaussian mixture parameters, in the case of the bias field, optimising $\mathcal{L}$ instead of $\mathcal{J}$, does not make the problem tractable in closed form, therefore numerical optimisation techniques must be exploited. The Levenberg-Marquardt (LM) algorithm is the scheme that has been mainly explored in this work, as it possesses the interesting property of providing a trade off between the gradient descent method, which is highly convergent but rather slow, and the Gauss-Newton method (Bertsekas, 1999), which converges much faster, but only if the initial estimate is reasonably close to a stationary point and if the Hessian matrix is not ill-conditioned (Marquardt, 1963; Moré, 1978).

It should be noted that, when using gradient based techniques, the conventional ML and EM approaches differ primarily in the form of the Hessians, as the initial gradients are the identical, due to the fact that the lower bound is tangent to the log likelihood function. In practice, for the model presented here, the Hessians of the lower bound were found to be better-behaved computationally. In addition, performing an E-step, that is regenerating the lower bound, every couple of Levenberg-Marquardt iterations ensures fast convergence, even if the optimisation is not performed directly on the log likelihood function.

Having denote by $\boldsymbol{H}$ the Hessian matrix of $\mathcal{L}$ and by $\boldsymbol{g}$ its gradient vector, applying the LM algorithm involves iteratively updating $\boldsymbol{\beta}_i$ by

$$\boldsymbol{\beta}_i^{(n+1)} = \boldsymbol{\beta}_i^{(n)} - (\boldsymbol{H}(\boldsymbol{\beta}_i^{(n)}) - \lambda_\beta I_{\boldsymbol{\beta}})^{-1}\boldsymbol{g}(\boldsymbol{\beta_i}^{(n)}) \,, \tag{3.56}$$

where I_β represents an identity matrix of suitable dimensions and λ_β is a damping parameter, which can be automatically adjusted at every iteration, to modulate the trade off between gradient descent ($\lambda_\beta \to \infty$) and Gauss-Newton ($\lambda_\beta \to 0$).

If the bias fields of the single channels (imaging modalities) are assumed to be independent, the first derivative of $\mathcal{L}$ with respect to the l-th component of the bias is given by

$$\frac{\partial \mathcal{L}}{\partial b_{ijl}} = \frac{1}{b_{ijl}} - \mathrm{x}_{ijl} \sum_{k=1}^{K} \gamma_{ijk} \frac{\mathrm{x}_{ijl}\ \mathrm{b}_{ijl} - \mu_{ikl}}{\sigma_{ikl}^2} \,, \tag{3.57}$$

while the second derivative can be computed as

$$\frac{\partial^2 \mathcal{L}}{\partial \mathrm{b}_{ijl}^2} = -\left(\frac{1}{\mathrm{b}_{ijl}^2} + \mathrm{x}_{ijl}^2 \sum_{k=1}^{K} \frac{\gamma_{ijk}}{\sigma_{ikl}^2} \right) . \tag{3.58}$$

Since b_{ijl} is modelled as

$$\mathrm{b}_{ijl} = \exp\left(\sum_{p=1}^{P} \beta_{ilp}\, \Phi_{jp} \right) , \tag{3.59}$$

with P equal to the number of basis functions and Φ_{jp} indicating the value of the p-th basis function at voxel j, the derivatives of the bias with respect to $\boldsymbol{\beta}_{il}$ are given by

$$\frac{\partial \mathrm{b}_{ijl}}{\partial \boldsymbol{\beta}_{il}} = \mathrm{b}_{ijl}\, \boldsymbol{\Phi}_j , \tag{3.60}$$

$$\frac{\partial^2 \mathrm{b}_{ijl}}{\partial \boldsymbol{\beta}_{il}^2} = \mathrm{b}_{ijl}\, \boldsymbol{\Phi}_j\, (\boldsymbol{\Phi}_j)^T . \tag{3.61}$$

Finally the gradient that must be used in the update rule (3.56) is, for the ML case

$$\begin{aligned} \boldsymbol{g}^{ML}(\boldsymbol{\beta}_{il}) = \frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}_{il}} &= \sum_{j=1}^{N} \frac{\partial \mathcal{L}}{\partial \mathrm{b}_{ijl}} \cdot \frac{\partial \mathrm{b}_{ijl}}{\partial \boldsymbol{\beta}_{il}} \\ &= \sum_{j=1}^{N} \boldsymbol{\Phi}_j \left(1 - \mathrm{x}_{ijl}\, \mathrm{b}_{ijl} \sum_{k=1}^{K} \gamma_{ijk} \frac{\mathrm{x}_{ijl}\, \mathrm{b}_{ijl} - \mu_{ikl}}{\sigma_{ikl}^2} \right) , \end{aligned} \tag{3.62}$$

and for the MAP case

$$\boldsymbol{g}^{MAP}(\boldsymbol{\beta}_{il}) = \boldsymbol{g}^{ML}(\boldsymbol{\beta}_{il}) - \boldsymbol{\Sigma}_{\beta}^{-1} \boldsymbol{\beta}_i . \tag{3.63}$$

While the Hessian can be computed by

$$\boldsymbol{H}^{ML}(\boldsymbol{\beta}_{il}) = \frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\beta}_{il}^2} = \sum_{j=1}^{N} \frac{\partial^2 \mathcal{L}}{\partial \mathrm{b}_{ijl}^2} \cdot \frac{\partial \mathrm{b}_{ijl}}{\partial \boldsymbol{\beta}_{il}} \cdot \left(\frac{\partial \mathrm{b}_{ijl}}{\partial \boldsymbol{\beta}_{il}} \right)^T + \frac{\partial \mathcal{L}}{\partial \mathrm{b}_{ijl}} \cdot \frac{\partial^2 \mathrm{b}_{ijl}}{\partial \boldsymbol{\beta}_{il}^2} . \tag{3.64}$$

or by

$$\boldsymbol{H}^{MAP}(\boldsymbol{\beta}_{il}) = \boldsymbol{H}^{ML}(\boldsymbol{\beta}_{il}) - \boldsymbol{\Sigma}_{\beta}^{-1} , \tag{3.65}$$

for the ML and MAP approaches respectively.

Finally it should be noted that, while for the model presented here, which operates in the native intensity domain, the problem of updating the bias field parameters cannot be solved in closed form, this is not true if the data is log-transformed prior to model fitting (Van Leemput et al., 1999a). In such a case in fact, the dependency of the objective function on the bias field, which becomes additive rather than multiplicative, turns out to be quadratic. The question on which parametrisation is best-suited to represent medical image data is not explicitly explored in this thesis though.

3.4.3. Estimating affine deformations

A similar Levenberg-Marquardt approach can be used to optimise the affine deformation parameters $\{\mathbf{a}_1, \ldots, \mathbf{a}_M\}$ during each M-step. The resulting update rule can be expressed as follows

$$\mathbf{a}_i^{(n+1)} = \mathbf{a}_i^{(n)} - (\boldsymbol{H}(\mathbf{a}_i^{(n)}) - \lambda_a I_a)^{-1} \boldsymbol{g}(\mathbf{a}_i^{(n)}) \ . \tag{3.66}$$

For the MAP problem the gradient vector is equal to

$$\boldsymbol{g}^{MAP}(\mathbf{a}_i) = \boldsymbol{g}^{ML}(\mathbf{a}_i) - \boldsymbol{\Sigma}_a^{-1} \mathbf{a}_i \ , \tag{3.67}$$

where the nth component of $\boldsymbol{g}^{ML}(\mathbf{a}_i)$ is given by

$$g_n^{ML}(\mathbf{a}_i) = \sum_{j=1}^{N} \sum_{k=1}^{K} \frac{\gamma_{ijk}}{\pi'_{jk}} \cdot \frac{\partial \pi'_{jk}}{\partial a_{in}} \ , \tag{3.68}$$

Similarly, for the 12 × 12 Hessian matrix

$$\boldsymbol{H}^{MAP}(\mathbf{a}_i) = \boldsymbol{H}^{ML}(\mathbf{a}_i) - \boldsymbol{\Sigma}_a^{-1} \ , \tag{3.69}$$

where each element $H_{n,m}^{ML}(\mathbf{a}_i)$ is computed making use of the following semidefinite approximation of the second derivatives of $\mathcal{L}$

$$\boldsymbol{H}_{n,m}^{ML}(\mathbf{a}_i) = -\sum_{j=1}^{N} \sum_{k=1}^{K} \frac{\gamma_{ijk}}{\pi'^2_{jk}} \cdot \frac{\partial \pi'_{jk}}{\partial a_{in}} \cdot \frac{\partial \pi'_{jk}}{\partial a_{im}} \ . \tag{3.70}$$

To evaluate expression (3.70), the derivatives of the warped tissue priors π'_k, with respect to the deformation parameters $\mathbf{a}_i$, must be obtained. For this purpose, it is convenient to exploit the chain rule for composed functions, to give

$$\frac{\partial \pi'_{jk}}{\partial \alpha_{im}} = \sum_{r=1}^{3} \sum_{c=1}^{4} \left(\nabla \left[\pi_k(\mathbf{y}'_j)\right]\right)^T \cdot \frac{\partial \mathbf{y}'_j}{\partial \mathrm{T}_{rc}(\mathbf{a}_i)} \cdot \frac{\partial \mathrm{T}_{rc}(\mathbf{a}_i)}{\partial \alpha_{im}} \ , \tag{3.71}$$

where $\mathbf{y}'_j$ is the vector of coordinates mapping from voxel j of image i into the space of the template. The first term on the right hand side of (3.71) represents the gradient of the k-th warped tissue probability map, evaluated at $\mathbf{y}'_j$. The second term is the derivative of the coordinate vector $\mathbf{y}'_j$, with respect to the transformation matrix element T_{rc}. Finally, $\mathrm{T}_{rc}(\mathbf{a}_i)$ must to be derived with respect to the deformation parameter a_{im}, by making use of expression (3.10).

An equivalent matricial representation of equation (3.71) can be obtained, by exploiting the relationship in (3.16), to give

$$\frac{\partial \pi'_{jk}}{\partial a_{im}} = \left(\nabla \left[\pi_k(\mathbf{y}'_j) \right] \otimes \begin{bmatrix} \mathbf{y}_j + \mathbf{u}_{ij} \\ 1 \end{bmatrix} \right)^T \cdot \operatorname{vec}\left(\frac{\partial \mathbf{T}(\mathbf{a}_i)}{\partial a_{im}} \right) , \tag{3.72}$$

where $\otimes$ indicates the Kronecker product. Alternatively, quasi-Newton or pseudo-Newton approaches, which directly compute inverse Hessian approximations, could have been explored to solve this optimisation problem.

3.4.4. Estimating non-linear small deformations

Similarly, for the estimation of the non-linear small displacement fields, gradient-based techniques represent well suited optimisation strategies. For this purpose, the following gradient of the lower bound can be used

$$\boldsymbol{g}^{MAP}(\mathbf{u}_i) = \sum_{k=1}^{K} \boldsymbol{\gamma}_{ik} \otimes \mathbf{g}^{\pi}_{ik} - \mathbf{L}_u \mathbf{u}_i , \tag{3.73}$$

where at each voxel j the vector field $\mathbf{g}^{\pi}_{ik}$ takes a value given by

$$\mathbf{g}^{\pi}_{ijk} = (\mathbf{T}(\boldsymbol{\alpha}_i))^T \, \nabla \left[\log \pi_k(\mathbf{y}'_j) \right] , \tag{3.74}$$

and $\mathbf{L}_u$ is a differential operator used to compute the penalty term.

As for the optimisation problems described previously, the rate of convergence can be greatly increased, compared to a simple gradient descent approach, by taking into account the second derivatives of the objective function (Klein et al., 2007).

In particular, by assuming that

$$\frac{\partial^2 \pi_k}{\partial u_{il} \partial u_{im}} = 0 \, , \forall \, l, m \in \{1, 2, 3\} , \tag{3.75}$$

a positive semidefinite approximation to the Hessian of $\mathcal{L}$ can be computed, as follows

$$\boldsymbol{H}^{MAP}(\mathbf{u}_i) = -\sum_{k=1}^{K} \boldsymbol{\gamma}_{ik} \otimes \mathbf{H}^{\pi}_{ik} - \mathbf{L}_u , \tag{3.76}$$

where $\mathbf{H}^{\pi}_{ik}$ is a tensor field such that, at voxel j

$$\mathbf{H}^{\pi}_{ijk} = (\mathbf{T}(\mathbf{a}_i))^T \, \nabla \left[\log \pi_k(\mathbf{y}'_j) \right] \left(\nabla \left[\log \pi_k(\mathbf{y}'_j) \right] \right)^T \mathbf{T}(\mathbf{a}_i) \, . \tag{3.77}$$

Due to the high dimensionality of the parametrisation of the displacement fields $\{\boldsymbol{u}_i\}_{i=1,\ldots,M}$ it is not possible to solve this optimisation problem by numerical matrix inversion, as indicated in the previous examples, since this would be prohibitively expensive from a computational point of view. One approach is to treat the problem as a partial differential equation problem, which for instance can be solved via multigrid solvers (Ashburner, 2007; Modersitzki, 2004). Alternatively, an approximated inverse of the Hessian can be computed, without having to evaluate second order derivatives, by using only the gradient information (Nocedal, 1980).

3.5. Limitations ML and MAP estimation

The method presented in this chapter relies on maximum likelihood and maximum a posteriori estimation techniques to fit a joint statistical model of shape and intensity to structural imaging data. Both approaches are indeed commonly used for many and diverse data modelling problems. Nevertheless they suffer from a number of limitations, that could, at least potentially, be overcome in a fully Bayesian framework.

A first problem has to do with the fact that, in practice, most log likelihood or log posterior functions are multimodal. As a result, many ML and MAP estimation algorithms, such as gradient-based approaches and the EM algorithm, are quite sensitive to the initialisation of parameters, with a considerable chance of getting trapped in a local optimum, if such an initialisation is not properly tuned (Ueda et al., 2000).

Apart from that, which represents a computational issue that can be mitigated by using specific optimisation techniques, such as simulated annealing (Goffe et al., 1994) or genetic algorithms (Sekhon and Mebane Jr, 1998), there is also a crucial theoretical point that makes ML and MAP methods suboptimal. That is the fact that both of them do not provide the full posterior distribution of the model parameters Θ, but return instead point estimates. In other words, information on the posterior uncertainty in the estimates of Θ is missing, which means that for making predictions on unseen data $\hat{\mathbf{x}}$, given a training data set $\mathbf{X}$, it is necessary to resort to the following approximation

$$p(\hat{\mathbf{x}}|\mathbf{X}) = \int p(\hat{\mathbf{x}}|\mathbf{X}, \Theta) p(\Theta|\mathbf{X}) \mathrm{d}\Theta \approx p(\hat{\mathbf{x}}|\mathbf{X}, \Theta^{MAP}) \, , \tag{3.78}$$

where the true posterior $p(\Theta|\mathbf{X})$ is replaced by a delta Dirac function centred on the mode Θ^{MAP}.

This often results in the occurrence of overfitting, as well as in the difficulty to perform model comparison (Draper, 1995). Overfitting, in particular, is a phenomenon that is very difficult to avoid within the maximum likelihood framework. In fact, when a model that is overly complex (i.e. flexible) is fit to training data via ML estimation, very high values of the likelihood function can typically be obtained. Nevertheless the generalisation performance of such a model to unseen test data, as well as its predictive capability, might be extremely poor because, most likely, noise has also been fit, together with signal. In such a case however, the attained likelihood value is not an indicator of whether the observed data has been overfitted or not.

Moreover, for the particular case of Gaussian mixture models, extreme cases of overfitting can occur because the log likelihood function has a number of singular points. At these points its value goes to infinity, because at least one component of the mixture degenerates to a Dirac delta function centred on one of the observed data points (Bishop, 2006). From a computational point of view this causes the optimisation to fail or become dangerously unstable. In other words, maximisation of the log likelihood function for Gaussian mixture models is an ill-posed problem, because the objective function is unbounded from above (Biernacki and Chrétien, 2003).

Many of the problems associated with ML overfitting can be addressed within a maximum a posteriori framework. In this case the problem is regularised by penalising implausible parameter values and this, in general, also ensures greater computational stability. Anyway the MAP framework does not solve the problem of allowing different models (m) to be compared for the purpose of model selection. In fact, the value taken by the posterior probability at its mode $p(\hat{\Theta}^{MAP}|\mathbf{X}, m)$ cannot be directly used to determine which model possesses optimal complexity, because, in such a way, overly complex models would be favoured. For example, if Gaussian mixture models, with different number of components (K), were trained via MAP estimation, monotonically increasing values of the log posterior should be expected for larger values of K.

In situations of this sort, cross-validation represents a principled way of assessing the optimal structure of the model (Corduneanu and Bishop, 2001). Nevertheless performing an exhaustive cross-validation study can be extremely expensive, for various reasons. First of all, the space of the models, in which the search has to be conducted, is generally too large to be thoroughly explored within a reasonable computational time. Secondly, for big data sets, the amount of computation might become prohibitive, even if the number of compared models is kept rather low.

Model selection criteria, such as the Bayesian information criterion (BIC) or the Akaike information criterion (AIC), provide a much less expensive solution to the problem of model comparison. Since such methods are very easily applicable (e.g. no integrals or inverse matrices have to be computed), they have been widely used for many data modelling problems. Nevertheless they suffer from the limitation of being approximate and, therefore, only valid under a number of assumptions, which regard for example the sample size and distribution of the data (Kuha, 2004). Some authors also question their consistency (Bozdogan, 1987), which, particularly for the problem of determining the optimal order of mixture models, has been shown to be quite poor (Celeux and Soromenho, 1996; Titterington et al., 1985).

In principle, the above mentioned limitations of ML and MAP estimation could be overcome by computing the evidence of the model, also known as marginal likelihood, which is defined as

$$p(\mathbf{X}|m) = \int p(\mathbf{X}|\Theta, m)p(\Theta|m)\,\mathrm{d}\Theta \;=\frac{p(\mathbf{X}|\Theta, m)p(\Theta|m)}{p(\Theta|\mathbf{X}, m)}\;. \tag{3.79}$$

Essentially, this would require integrating the same objective function used in maximum a posteriori estimation, over the entire parameter space. By doing so, models whose complexity exceeds the optimal trade-off between fitting and overfitting the observed data, will attain lower values of the marginal likelihood. Unfortunately the computations that have to be carried out to evaluate the evidence are usually intractable and therefore approximation strategies have to envisaged. One of these strategies, namely Variational Bayes, will be applied in Chapter 5 to solve image segmentation problems. Such an approach relies on analytical approximations and is in contrast with another broad family of approximate inference schemes, which rely on sampling techniques to compute intractable integrals (Rubin, 1976).

3.6. Summary

This chapter has presented a modelling framework for the probabilistic interpretation of structural MR data, using Gaussian mixture latent variable models and deformable, average-shaped tissue priors, which can be learned directly from large imaging data sets. A modified EM algorithm has been presented to estimate maximum likelihood, or maximum a posteriori, solutions. Experiments performed by applying the presented framework to real brain and spinal cord data will be described in Chapter 4, by comparing fully unsupervised versus semisupervised learning strategies.

The deformation model proposed in this chapter combines affine transformations and a small deformation non-parametric approach. Even if such a model allows to capture local shape differences at a small scale level, the results described in Chapter 4 will suggest that this approach is not optimal for encoding large shape variations from the group mean. For this reason a large deformation mapping approach will be discussed and validated in Chapter 6.

Finally, in the last section of the present chapter some limitations associated with ML and MAP estimation techniques have been outlined, such as the difficulty in preventing overfitting and in quantifying the uncertainty relative to point parameter estimates. Full Bayesian inference has already been indicated in this chapter as a possible framework to address such problems. This topic will be further developed and discussed in Chapter 5, where a variational Bayes approach will be adopted to fit generative models to the intensity distributions of MR data sets.

4

Unsupervised vs semisupervised template learning

4.1. Introduction

This chapter will present a series of experimental results, obtained by applying the method introduced in Chapter 3 to both real and synthetic MRI data, for the purpose of assessing its performance, in a fully unsupervised learning framework, as well as in a semisupervised setting. The potential and limitations of the two approaches will be discussed, particularly for the purpose of classifying anatomical tissues, from MRI data.

4.2. Learning with or without supervision?

From a machine learning perspective, generative models, such as the one introduced in the previous chapter, represent a natural framework for unsupervised learning, since they allow inferring the latent structure of the data, without relying on training outputs, or on any other form of feedback from the environment (Ghahramani, 2004). Indeed, from a general statistical perspective, unsupervised learning can be thought of as the process of building a representation of the data, by means of estimating the probability density distribution that the inputs are drawn from. This information can then be used for a number of processing tasks, among which classification and dimensionality reduction are probably the most well studied examples.

Such a framework is antithetic to the notion of supervised learning, which, on the contrary, indicates the process of learning, from a set of training input and output pairs, how to assign correct outputs to new input data, for the purpose of making choices and predictions, in a fully automated manner (Bishop, 1995). A crucial theoretical difference is that these last methods, as opposed to the first ones, do not necessarily need to capture the mechanisms underlying data generation, as long as they can effectively discriminate among different outputs.

A third machine learning strategy is known as reinforcement learning and it includes all those schemes where a machine receives an error, or reward, signal from the environment, in response to a set of output actions. In this case the goal is to learn what choices have to be made in order for the reward to be maximised, but without being explicitly informed about the desired output values (Sutton and Barto, 1998).

The choice on what learning scheme is the best cannot be answered independently from the nature of the problem, which the algorithm is trying to solve. For instance, supervised generative methods have the advantage of providing a direct mapping between inputs and user-defined outputs, which makes their results somewhat more interpretable, especially for real life applications. Unsupervised schemes can instead help to understand the system that is generating the observed data, but given the absence of a predefined output target, classification accuracy might sometimes be sacrificed. The question on interpretability is also a crucial topic in the field of discriminative machine learning (Caruana et al., 2015; Sturm et al., 2016; Vellido et al., 2012), where the validity of the rules learnt during model training is intrinsically harder to assess than in generative machine learning.

While, in principle, fully supervised learning schemes would be attractive for solving medical imaging problems, in practice their applicability is often limited by the amount of available training data. This is indeed a common scenario in many machine learning domains. In fact, since there is often a stronger motivation for collecting data rather than for labelling it, many research fields suffer from the existence of a disproportion between the amount of labelled and unlabelled data (Goldman and Zhou, 2000). In such a scenario, it has been hypothesised that the predictive accuracy of fully supervised algorithms, in the case where only few training examples are available, might be increased by incorporating unlabelled data into the learning framework, thus resorting to a semisupervised approach (Chapelle et al., 2006; Filipovych et al., 2011; Zhu, 2006).

4.3. Unsupervised template learning

The set of experiments presented in this section were conducted by applying the algorithmic scheme introduced in Chapter 3 to cross-sectional data sets of head and neck scans, acquired with different MR imaging modalities. Training of the model was performed in a fully unsupervised manner, that is to say without relying on any labelled data, while different deformation models were compared, namely affine and non-linear small deformations.

4.3.1. Affine tissue templates

The first set of experiments were performed to assess the performance of the presented method in a simple affine registration setting. For the purpose of model training, T1-, T2- and PD-weighted images of fifty subjects were randomly selected from the freely available IXI brain data base (http://brain-development.org/ixi-dataset). The selected subjects were scanned at Guys Hospital in London using a Philips 1.5T system with the scanning protocols detailed in Table 4.1. The resolution of all the scans equals $0.94 \times 0.94 \times 1.2\ mm^3$, with sagittal orientation for the T1-weighted scans and axial orientation for the T2- and PD-weighted images. The average age within the selected sample is 50.8 years and the population consists of twenty-two males and twenty-eight females. No preprocessing of the data was performed, except for resampling the T2- and PD-weighted data in the same coordinate space as the T1-weighted scans, whereas model parameters were initialised as follows

- Twelve classes were included in the Gaussian mixture model, with such a number being chosen purely based on empirical evidence. Initial estimates of the intensity mean vectors and covariance matrices were obtained by performing a K-means clustering analysis on the intensity distribution of a subject, randomly selected from the training database (Biernacki et al., 2003). The tissue probability maps $\{\boldsymbol{\pi}_k\}_{k=1\ldots K}$ instead, were initially assumed to be flat spatial priors ($\pi_{jk} = \pi_k = \frac{1}{K}$).

Table 4.1: MRI acquisition parameters relative to the data from the IXI database used to construct the templates presented in this chapter.

Scanning Parameters	**T1w**	**T2w**	**PDw**
Repetition Time (ms)	9.813	8178.34	8178.34
Echo time (ms)	4.603	100	8
Phase Encoding Steps	192	187	187
Echo Train Length	0	16	16
Reconstruction Diameter (mm)	240	240	240
Flip angle (°)	8	90	90

- To model the bias field, only the first five lower frequency DCT basis functions where used, along each of the three Cartesian axes, to give a total of 5^3 three dimensional basis functions. Initial estimates of the bias field parameters were set to zero.
- The affine transformation parameters were all set equal to zero, thus initialising the coordinate mappings as identity transforms.

Figure 4.1 illustrates an example of intensity non-uniformity correction, performed by the presented algorithm. Coronal, sagittal and axial views of the twelve generated tissue probability maps are illustrated in Figure 4.2 and Figure 4.3. An average-shaped T1-weighted template is also shown in Figure 4.4, which was obtained as an arithmetic average of the intensities of the training data, after having bias corrected, spatially normalised and linearly rescaled the images in the same intensity range (between zero and five hundred and twelve).

Predictive accuracy

In order to quantitatively evaluate the performance of the presented method, the predictive power of the underlying probabilistic model was tested on a set of unseen images, adopting a holdout validation scheme. The aim of such an experiment was to assess the extent to which the population-based templates, generated with the presented method,

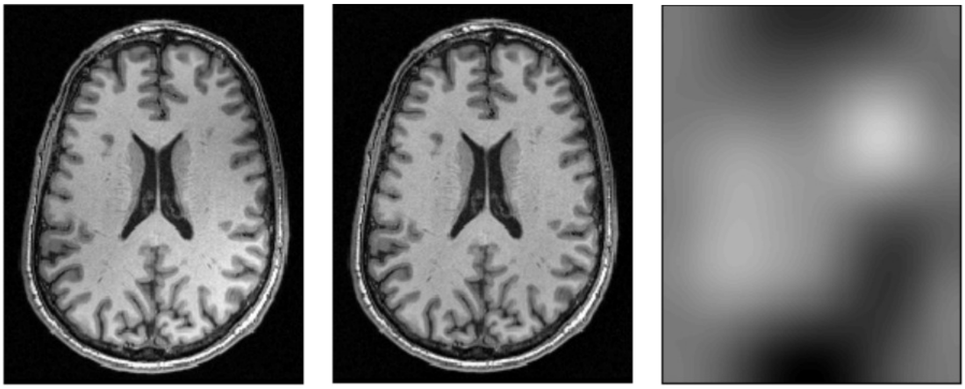


Figure 4.1: Example of bias field correction performed by the presented algorithm. The original scan is depicted in panel (a), while the corrected image and estimated bias field are reported in panels (b) and (c) respectively.

are representative of unseen test data.

For this purpose, the estimated affine tissue probability maps were used as deformable spatial priors within a Gaussian mixture model, to fit the intensity distributions of twenty five test images obtained from the IXI database, after having removed randomly located blocks, of 10 × 10 × 10 voxels, from each image.

Gaussian mixture model parameters $\hat{\Theta} = \{\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}\}$ were estimated making use of the EM algorithm, after having affine registered the tissue probability maps to the test images. Then, the likelihood of the missing intensities $\{\mathbf{X}_i^{(m)}\}_{i=1,\ldots,M}$, given $\hat{\Theta}$, was computed, for each image i, as

$$p(\mathbf{X}_i^{(m)}|\hat{\boldsymbol{\Theta}}, \boldsymbol{\pi}) = \prod_{j=1}^{N^{(m)}} \left(\sum_{k=1}^{K} \pi_{jk} \, \mathcal{N}(\mathbf{x}_{ij}^{(m)}|\hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k) \right) . \tag{4.1}$$

The performance of the presented method was compared against that of a publicly available algorithm for groupwise image registration (Modat et al., 2010) based on mutual information (MI) (Maes et al., 1997). In this case, a comparable measure of predictive accuracy can be derived by computing the likelihood of missing data, from the joint histogram of the T1-weighted template generated with the MI-based algorithm, and the warped test data, by

$$p(\mathbf{X}^{(m)}|\mathbf{Av}) = \prod_{j=1}^{N^{(m)}} \frac{H(\mathbf{a}_j, \mathbf{x}_j^{(m)})}{\sum_{\mathbf{x}} H(\mathbf{a}_j, \mathbf{x}_j^{(m)}) \Delta_{\mathbf{x}} \Delta_{\mathbf{a}} N^{(a)}} = \prod_{j=1}^{N^{(m)}} \frac{h(\mathbf{a}_j, \mathbf{x}_j^{(m)})}{\sum_{\mathbf{x}} h(\mathbf{a}_j, \mathbf{x}_j^{(m)})} , \tag{4.2}$$

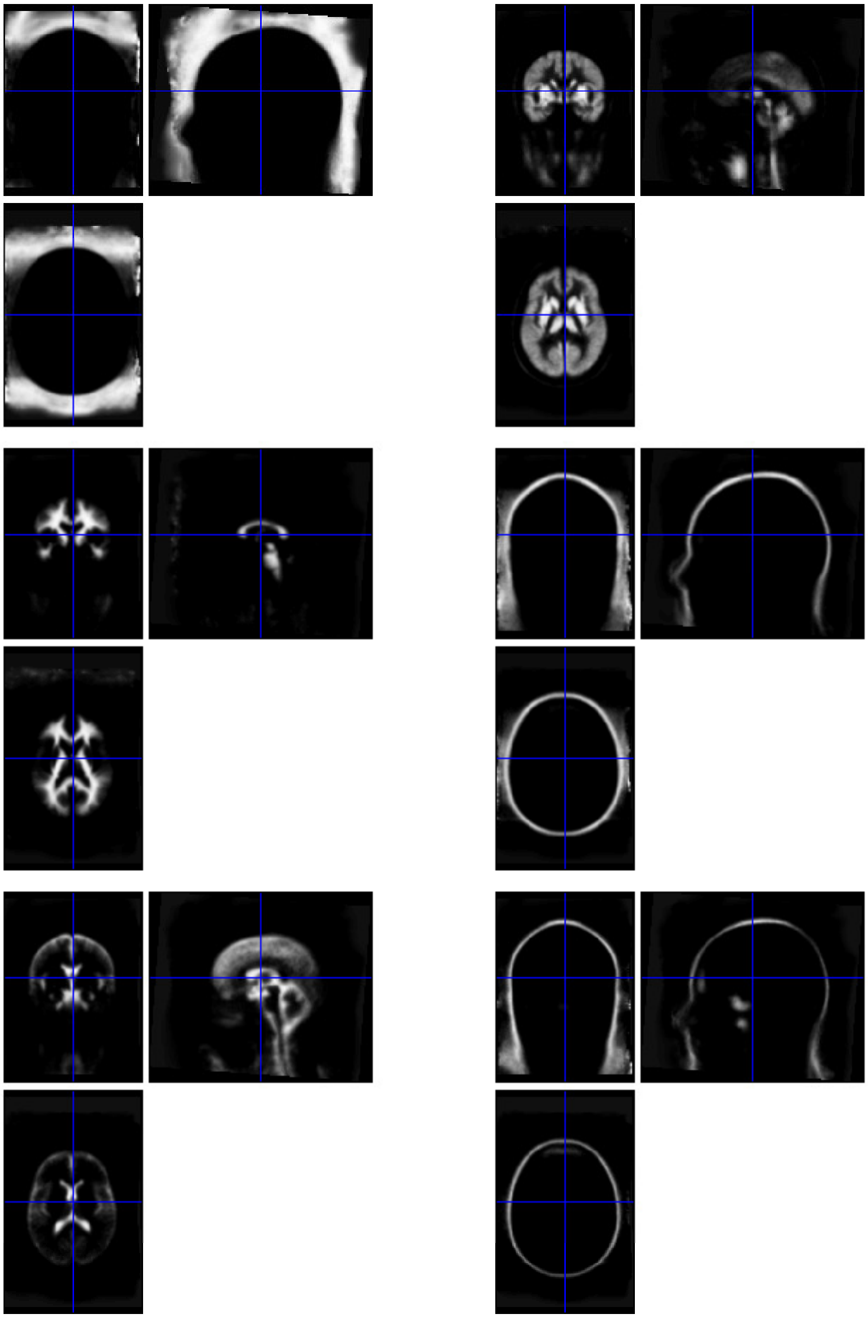


Figure 4.2: Coronal, sagittal and axial views of six out of the twelve tissue probability maps generated with the method described in Chapter 3, using a subset of the freely available IXI dataset and a simple affine deformation model.

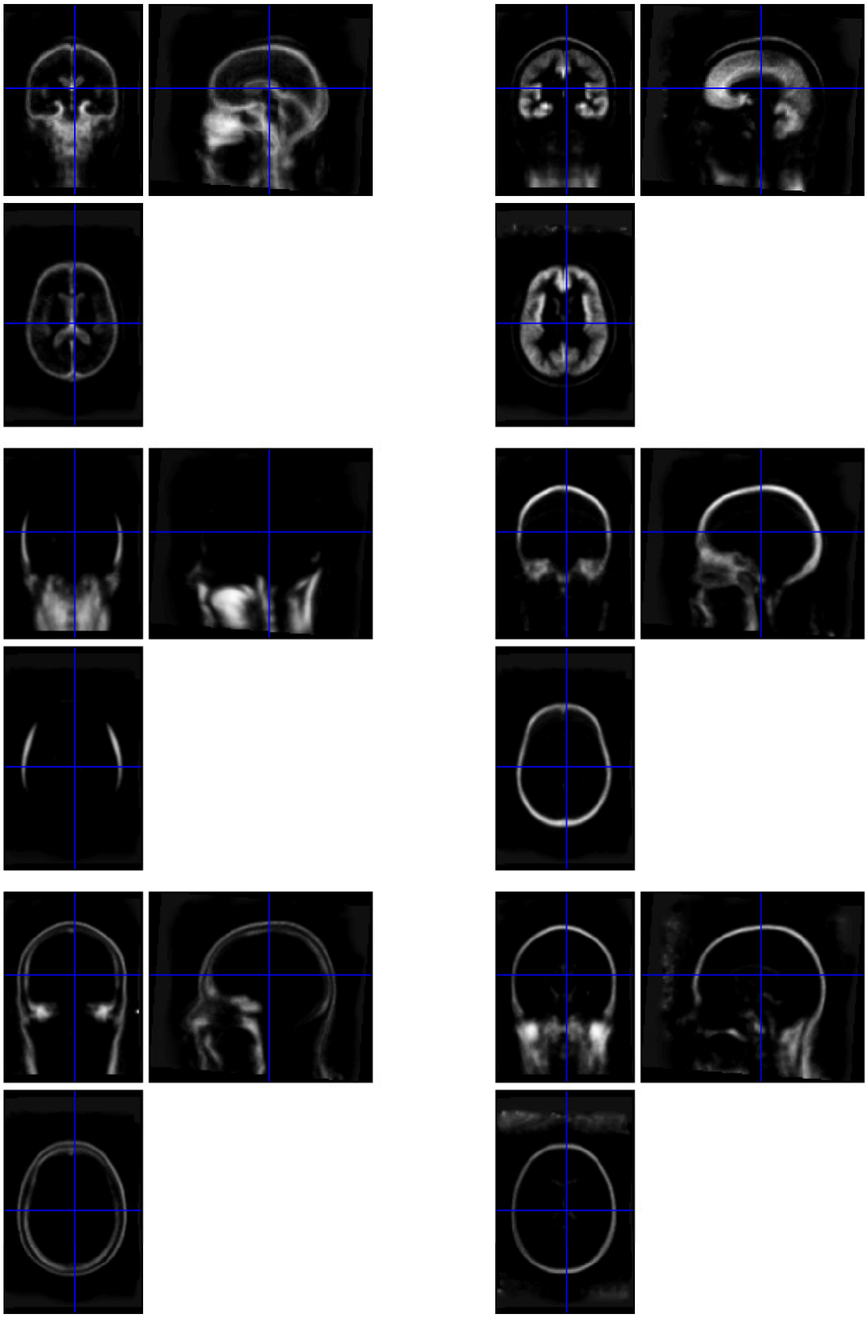


Figure 4.3: Coronal, sagittal and axial views of six out the twelve tissue probability maps generated with the method described in Chapter 3, using a subset of the freely available IXI dataset and a simple affine deformation model.

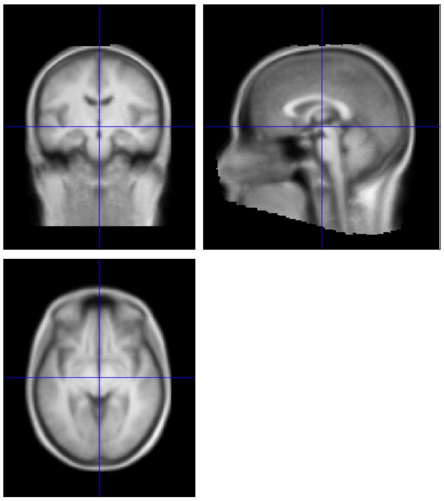


Figure 4.4: Coronal, sagittal and axial views of an average-shaped T1-weighted image generated with the presented groupwise algorithm.

where H and h denote the heights of the unnormalised and normalised histograms respectively, $N^{(a)}$ indicates the number of observed data points and $\Delta_{\mathbf{x}}\Delta_{\mathbf{a}}$ equals the bin area.

It should be noted that, while the two models compared here have fairly different complexity, the fact that predictive accuracy is evaluated on missing (i.e. unseen) test data should implicitly allow to control for the effect of model complexity when assessing the performance of the two approaches. The results of this model cross-validation, which are summarised in Figure 4.5 and table 4.2, seem to indicate that the approach adopted in this work exhibits higher predictive performance, compared to groupwise mutual information modelling. Nonetheless the presented method yields larger variability in test accuracy, which supposedly indicates high sensitivity of the method to registration accuracy. This is indeed a well know limitation of probabilistic atlas-based modelling approaches (Iglesias and Sabuncu, 2015; Yeo et al., 2008).

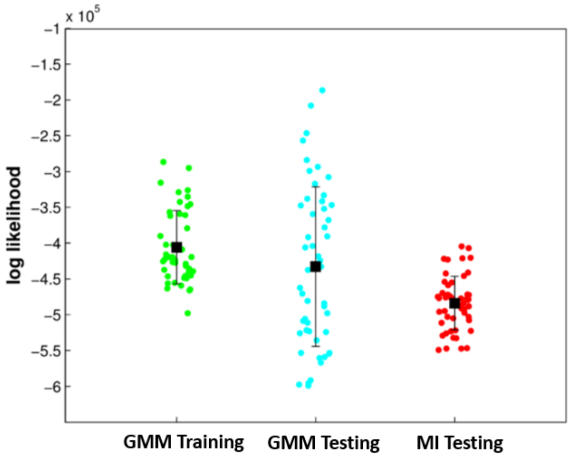


Figure 4.5: Distributions of predictive performance measures obtained for the method presented in Chapter 3 and for a freely available groupwise registration algorithm based on mutual information. Squares indicate the average log likelihood of test data, while error bars represent standard deviations.

Table 4.2: Distributions of predictive accuracy, evaluated as the log likelihood of unseen test data, for the model introduced in Chapter 3, as compared to a mutual information based approach.

	Training	**Testing**	
	GMM	**GMM**	**MI**
Mean	$-4.05 \times 10^{+5}$	$-4.32 \times 10^{+5}$	$-4.83 \times 10^{+5}$
Standard deviation	$5.12 \times 10^{+4}$	$1.11 \times 10^{+5}$	$3.74 \times 10^{+4}$

Segmentation accuracy

Another way of assessing the validity of the presented framework is by means of evaluating the tissue classification accuracy attained when the generated tissue probability maps are used as priors, within atlas-based segmentation methods, to identify tissue types in unseen test data. In fact, even if incorporating such test scans in the training data set used for constructing the templates would improve classification accuracy, the generalisation capability of the proposed framework can only be assessed if the test data is not exploited during training. For this purpose, the tissue probability maps illustrated in Figures 4.2 and 4.3 were tested in combination with the segmentation algorithm implemented in SPM12 (Ashburner and Friston, 2005) to segment synthetic brain data produced by the BrainWeb MR simulator (Cocosco et al., 1997; Collins et al., 1998; Kwan et al., 1999). For this purpose the tissue probability maps containing gray and white matter were manually identified.

While in principle unsupervised learning of the tissue priors might not be the most suitable framework for this type of analyses, due to the difficulty of differentiating tissues with overlapping intensity distributions, in practice, this seems to affect mainly the neck region (see Figure 4.2), at least when using multimodal training data such as in these experiments. Therefore reliable brain tissue classification accuracy measures could be obtained for the Brainweb data, which does not include the neack. In particular, segmentation accuracy was quantified by computing the Dice similarity coefficients[1] (DSC) between the estimated gray and white matter maps and the ground truth provided by the underlying anatomical model of the simulated data.

Results are reported in Figure 4.6 where the DSC are plotted for different noise levels (3%, 5% and 7% of the brightest image intensity). Solid lines correspond to the results obtained when using T1- and T2-weighted simulated scans, while dotted lines refer to the similarity measures attained with a single modality (T1-weighted). Results indicate that high segmentation accuracy can be obtained by using the tissue priors estimated with the presented method, moreover they suggest that the use of multi-spectral data guarantees higher robustness to noise, as opposed employing only T1-weighted data, which yields a linear decrease of accuracy for increasing noise levels. However, in the

[1] The Dice score over two sets A and B is defined as $DSC = 2\frac{|A \cap B|}{|A|+|B|}$.

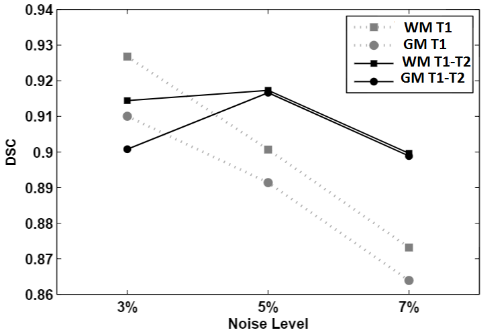


Figure 4.6: Segmentation accuracy obtained on synthetic data, when using the affine templates, generated by the proposed methods, as tissue priors within the segmentation algorithm implemented in SPM12. Dice score coefficients of gray and white matter are reported for different noise levels. The experiments were performed on T1-weighted data, as well as on multispectral data consisting of T1- and T2-weighted images.

presence of little noise corrupting the data (3%) the same multi-modal approach exhibits lower segmentation accuracy for both gray and white matter, compared to the case in which only T1-weighted data is available. This effect might be due to the poorer contrast between gray and white matter in T2-weighted as opposed to T1-weighted scans, in combination with the presence of partial volume effects, which the proposed model does not account for. In fact, this directly affects the accuracy of the estimates of the Gaussian parameters, for instance by inducing overly large eigenvalues of the covariance matrices, in particular for white matter. The same effect however might be concealed at higher noise levels.

4.3.2. Non-linear tissue templates

The experiments reported in the previous section have shown that by means of affine deformations it is possible to capture global shape and size differences across individuals and that templates built with such an approach could be used to perform atlas-based brain segmentation.

Nevertheless, the intrinsic limitation of the affine model, that is its inability to capture higher dimensional shape features, makes it unsuitable to perform morphometric analyses in a domain, such as that of human anatomy, where cross-sectional and longitudinal variability is so large and multidimensional (Bookstein, 1996; Denton et al., 1999; Rueckert et al., 2003).

For this reasons, the framework introduced in Chapter 3 was further tested in a non-linear deformation setting, in particular by means of implementing a small deformation modelling scheme, as described in Section 3.3.3.

Training data, for these experiments, was the same subset of the IXI database selected for generating the affine templates described in the previous sections, so as to facilitate a comparison of the two schemes. Initialisation of the Gaussian, bias, and affine model parameters was as in Section 4.3.1, while the additional non-linear small displacement fields were initialised as zero-valued vector fields.

Figure 4.7 shows some of the tissue probability maps obtained by fully unsupervised training on T1-weighted data. As to be expected, these results indicate that non-linear image registration, in spite of being more computationally expensive, is much more powerful for encoding shape variability, thus yielding sharper tissue probability maps. However, it should also be noted how training on a single imaging modality (e.g. T1-weighted) is confronted with the difficulty of discriminating tissues with overlapping intensity distributions, such as bone and cerebrospinal fluid (CSF) in these examples. With respect to this, acquiring multivariate training data is certainly an effective strategy to enhance classification accuracy. For instance, the results reported in Figure 4.8 show how adding T2- and PD-weighted data to the training set ensures more accurate cortical gray matter and CSF delineation. Unfortunately though, especially in a clinical setting it is not always possible to collect multiple scans for each subject, therefore simply relying on augmenting the dimensionality of the training data might not necessarily be a viable or convenient option.

Segmentation accuracy

As for the affine tissue probability maps described in Section 4.3.1, segmentation accuracy achieved using the non-linear probabilistic templates illustrated in Figure 4.8 was evaluated on synthetic Brainweb data, by providing them as tissue priors within the segmentation algorithm implemented in SPM12.

Results are reported in Figure 4.9 where the Dice score coefficients are plotted for different noise levels (3%, 5% and 7% of the brightest image intensity). Solid lines

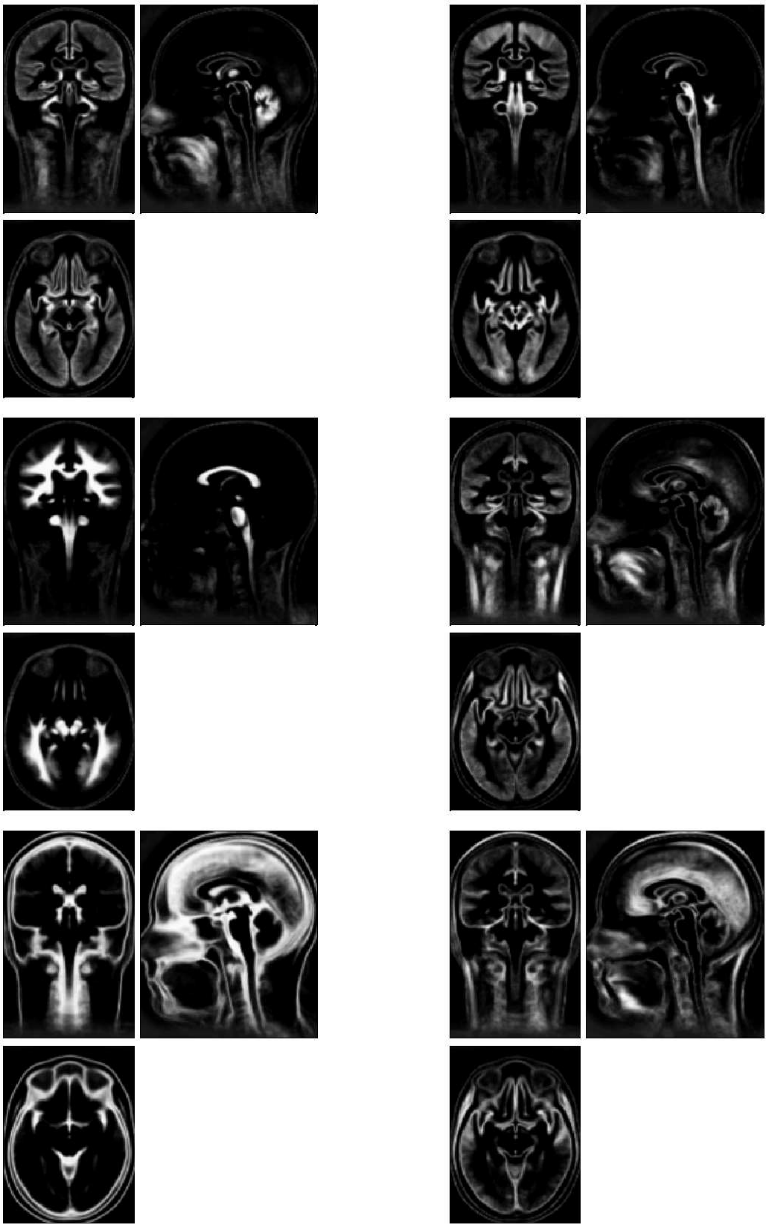


Figure 4.7: Non-linear tissue probability maps obtained by unsupervised training on T1-weighted data.

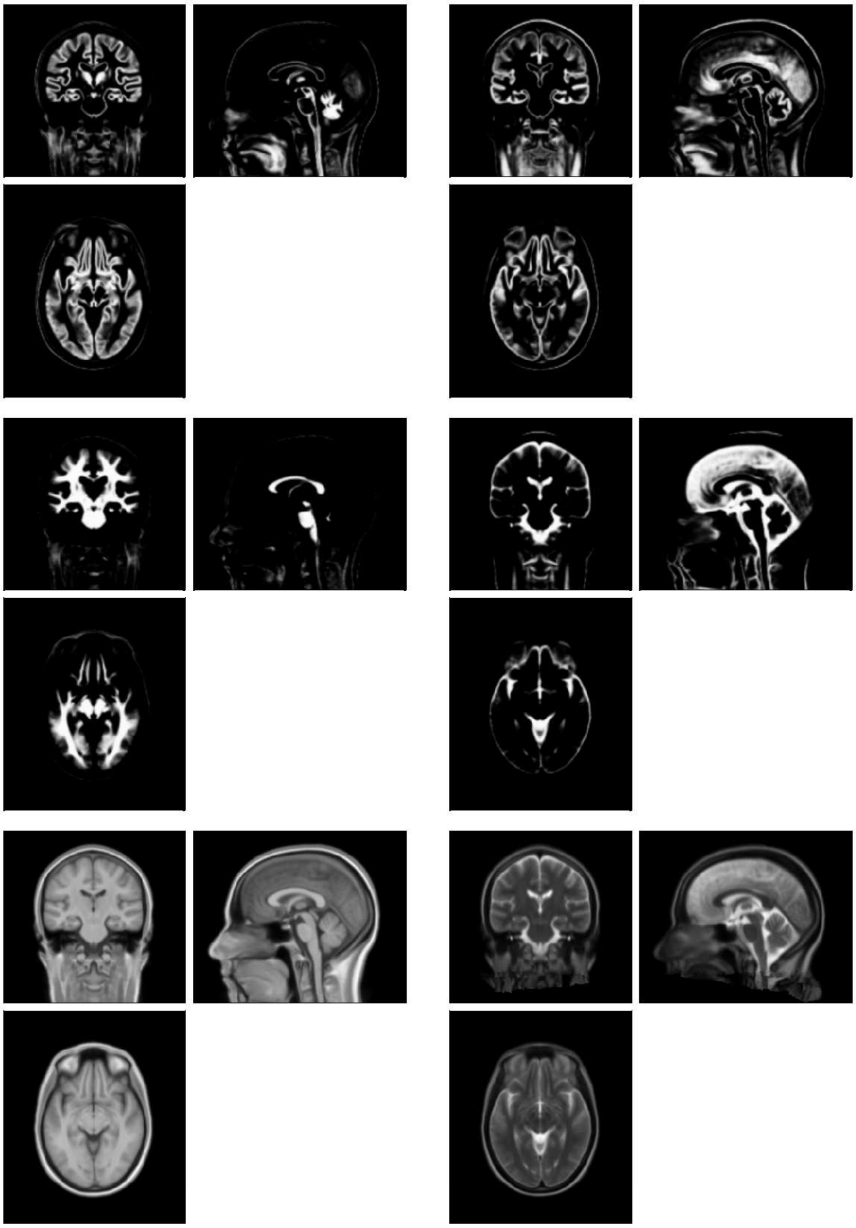


Figure 4.8: Non-linear tissue probability maps obtained by unsupervised training on T1-, T2- and PD-weighted data, together with T1- and T2-weighted average-shaped images (bottom row).

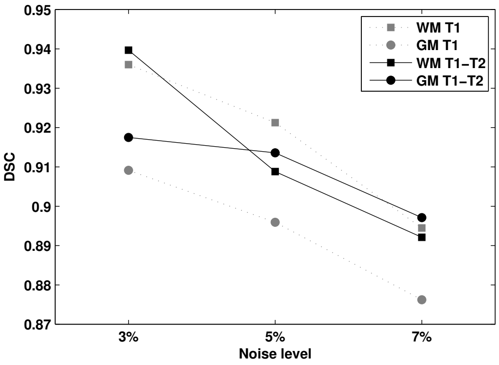


Figure 4.9

correspond to the results obtained when using T1- and T2-weighted simulated scans, while dotted lines refer to the similarity measures attained with a single modality (T1-weighted).

Further evaluation experiments of the presented modelling framework in a non-linear deformation setting will be presented in the remainder of this chapter by exploiting a semi-supervised learning scheme (Zhu, 2006). In fact, by allowing to include few annotated examples in the training data, such an approach provides a much more convenient framework for model evaluation, by allowing direct comparison between the results of model fitting and the available ground truth.

4.4. Semisupervised template learning

The experiments reported in the previous section have outlined, with practical examples, some advantages and intrinsic limitations of unsupervised generative learning from neuroimaging data sets. In particular, it has been shown that discriminating anatomical structures with a solely intensity-driven approach might be a non-trivial task, depending on the available image contrasts. A possible strategy to ameliorate the problem consists in incorporating in the training data a number of labelled examples, so as to implement a semisupervised generative learning scheme. In fact, fully supervised learning is often impractical, due to the expensive cost of defining reliable labelling protocols and generating expert manual annotations (Klein and Tourville, 2012), while semisupervised learning might provide a convenient trade-off solution, to simultaneously exploit the

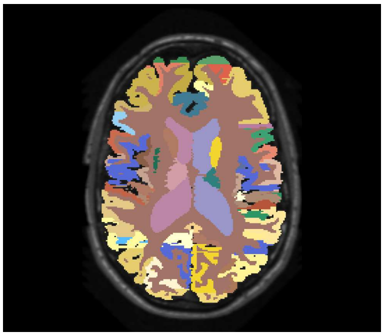


Figure 4.10: Example of manually annotated MR brain data from the OASIS database.

potential of the two approaches, but without having to generate very large volumes of annotated data, as indicated in Koch et al. (2015).

4.4.1. Brain templates

To test the performance of the modelling framework described in Chapter 3 in a semisupervised setting, a series of experiments were performed, making use of data from the OASIS (Open Access Series of Imaging Studies) database, which is publicly available for download from the web site `http://www.oasis-brains.org`.

The OASIS project is aimed at making MRI data sets of the brain freely available to the scientific community and it provides T1-weighted scans of four hundred and sixteen adults, aged between eighteen and ninety six, one hundred of which were diagnosed with very mild to moderate Alzheimer's disease, before or during the time of the study (Marcus et al., 2007). Additionally, for thirty five nondemented subjects, complementary brain labels were generated and made public by Neuromorphometrics, Inc. (`http://Neuromorphometrics.com`) under academic subscription. Such labels provide a fine parcellation of cortical and non-cortical structures, for a total of 139 labels across the brain (see Figure 4.10 for a single slice exemplar and Appendix E for a list of the labels).

In order to perform model fitting in a semisupervised fashion, all the available brain labels were grouped to form three tissue classes, corresponding respectively to cortical gray matter, subcortical gray matter and white matter. More details on how the training labels were used to generate ground truth tissue segmentations are provided in Appendix E. The data set was then split into a training group, including seventeen randomly selected subjects and a test group, consisting of the remaining eighteen subjects. Labels of the training data were provided as known latent variables in the process of model fitting, whereas labels of the test data were not used during training, but only for subsequent cross-validation analyses. The total number of tissue classes was set equal to twelve, with three classes corresponding to the training labels, as defined above. In principle more than one Gaussian could have been used for each training tissue class, however in practice having only one Gaussian per label was found to provide a convenient trade-off between accuracy and computational complexity. For the unlabelled voxels, it was assumed that the data could have been generated from any of the twelve tissue classes.

Figure 4.11 shows the cortical and subcortical gray matter tissue probability maps, resulting from applying the groupwise algorithmic framework introduced in Chapter 3 to the ensemble of training and test data. It should be noted that such a discrimination between cortical and subcortical gray matter would have not been possible, in a fully unsupervised framework, by training the model only on T1-weighted data and without introducing an *a priori* anatomical model.

Segmentation accuracy

To perform a quantitative evaluation of the presented framework, in terms of tissue classification accuracy, the ground truth test labels were compared to the tissue class membership probabilities, as automatically estimated by the algorithm during semisupervised model fitting. In particular, Dice score coefficients were computed, after having applied a threshold of 0.5 to the resulting probabilistic segmentations. In such a manner, as opposed to computing MAP labels, voxels where class membership probabilistic estimates are highly uncertain are not taken into account.

Results are plotted in Figure 4.12, for the three neural tissue types, and they indicate that high tissue classification accuracy can be obtained for both cortical gray matter and white matter, in spite having only one imaging modality available for training, by exploiting the presented modelling framework in a semisupervised learning setting.

However classification accuracy turned out to be lower for subcortical gray matter structures. Most probably, this has to be attributed to the fact that such nuclei are in close proximity to white matter and that some of them (e.g. the lateral nuclei of the

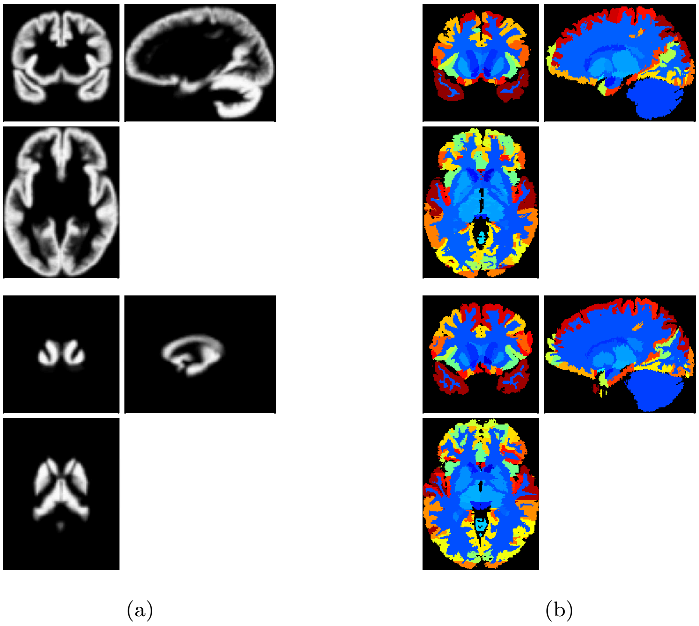


Figure 4.11: Gray matter cortical and subcortical tissue templates (a) obtained with a semisupervised approach and two individual label maps after spatial normalisation (b).

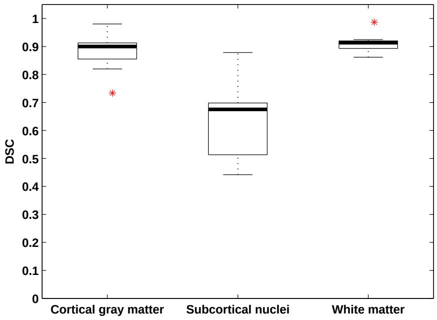


Figure 4.12: Segmentation accuracy obtained with the semisupervised approach presented in this chapter to model test data from the OASIS database. For each boxplot, the central mark indicates the median, the edges of the box are the 25^{th} and 75^{th} percentiles, the whiskers extend to the most extreme data points, while outliers are indicated by red dots.

thalamus) share with white matter very similar intensity distribution. Possibly, getting closer to a fully supervised scheme, by increasing the ratio of training versus test data, together with using multiple Gaussian components to model such a class, might have ensured higher classification accuracy. Indeed, it has been shown that fully supervised Bayesian models of shape and appearance (Patenaude et al., 2011), as well as multi-atlas label fusion techniques (Aljabar et al., 2007; Heckemann et al., 2006), can perform well for segmenting subcortical structures. Additionally, the work of Powell et al. (2008) and Milletari et al. (2016) seems to indicate that some discriminative classification techniques, such as artificial convolutional neural networks (Yegnanarayana, 2009) and support vector machines (Hearst et al., 1998), could also represent suitable strategies.

Registration accuracy

The presented modelling framework was also evaluated in terms of groupwise registration accuracy attained across different brain structures. For this purpose, since no ground truth is available for the unknown average-shaped brain anatomy, overlap measures were computed between each pair of spatially normalised test images, thus resulting in 153 pairwise overlap measures for each of the 139 anatomical labels.

Results are summarised in figures 4.13, 4.14 and 4.15. As to be expected, registration performance is highly dependent on the considered brain region. Larger and less morphologically variable structures, such as the brainstem and the cerebellum exhibit high groupwise overlap, whereas poorer group alignment is obtained for small cortical regions, as a result of significant inter-subject variability. In addition, the presence of negative outliers indicates that the method might not be sufficiently robust, particularly in the presence of significant shape deviations from the estimated average anatomy. Few cases of registration failure were caused by an overly large initial positioning mismatch, which the algorithm failed to compensate for in the absence of a rigid pre-alignment step. These results should be compared against those obtained in Chapter 6, on the same data set, but adopting instead the Large Deformation Diffeomorphic Metric Mapping (LDDMM) framework (Beg et al., 2005). This approach, in fact, in addition to representing morphometric variability in a more mathematically sound fashion, enables to encode larger deformations while incurring lower chances of breaking the underlying topology (Rueckert et al., 2006; Younes, 2010), as opposed to the small deformation setting. This is a desirable property for the sake of estimating average-shaped anatomical models and the results reported in Chapter 6 will demonstrate how this in turn yields significantly higher groupwise label overlap.

4.4.2. Spinal cord templates

Additional experiments were performed on high resolution cervical MR data, to test the applicability of the presented framework to spinal cord imaging data. For this purpose twenty healthy subjects were scanned at Balgrist University Hospital with a 3T Skyra MRI scanner (Siemens Healthcare, Erlangen, Germany). A 3D high-resolution optimised T2*-weighted multi-echo sequence (MEDIC) was applied to acquire five volumes of the cervical cord around the vertebral level of C2/C3. Each volume consisted of twenty contiguous slices acquired in the axial-oblique plane and was obtained with a resolution of $0.25 \times 0.25 \times 2.50\text{mm}^3$. The following parameters were used: field of view (FOV) of $162 \times 192\text{mm}^2$, matrix size of 648×768, repetition time (TR) of 44 ms, echo time (TE) of 19 ms, flip angle $\alpha = 11°$, and readout bandwidth of 260 Hz per pixel. After data acquisition, the five volumes of each subject were averaged in the space domain to increase signal to noise ratio (SNR). Figure 4.16 shows orthogonal sections of two of such averaged volumes.

The spinal cord gray matter was manually segmented in all twenty images by four different expert raters. The segmentations of ten subjects were used as training data, in a semisupervised learning setting with a total number of six Gaussian components, while the remaining ones were used as ground truth for validation. Majority voting label fusion (Heckemann et al., 2006) was performed on the labels provided by the four experts and, for classes that obtained the same number of votes, equal probabilities were assigned.

Figure 4.17 shows an average-shaped T2*-weighted image overlaid with the resulting gray and white matter spinal cord templates. Closeup views of the tissue probability maps are shown in Figure 4.18. When warping such templates to segment individual data a trilinear interpolation scheme was adopted, even if more sophisticated approaches could have been explored to exploit and enforce the cylindrical symmetry of the cord.

The manual labels of the test data were used to compute different accuracy metrics.

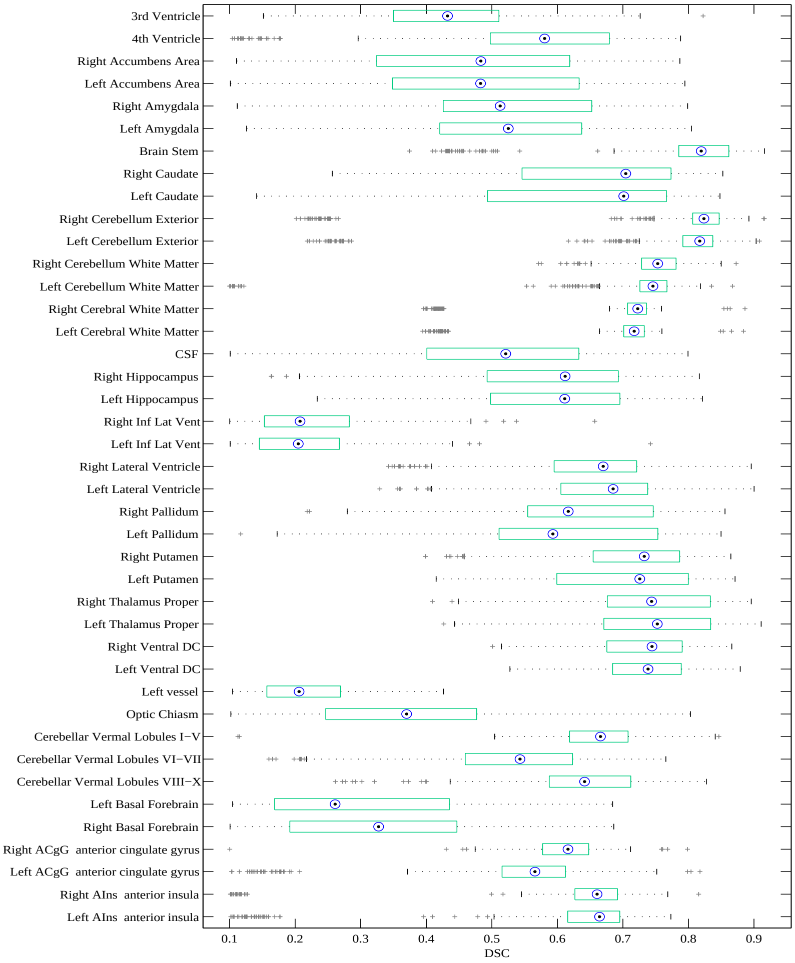


Figure 4.13: Distribution of pairwise overlap measures attained by the presented algorithm across different brain regions. For each boxplot, the central mark indicates the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points, while outliers are indicated by gray crosses.

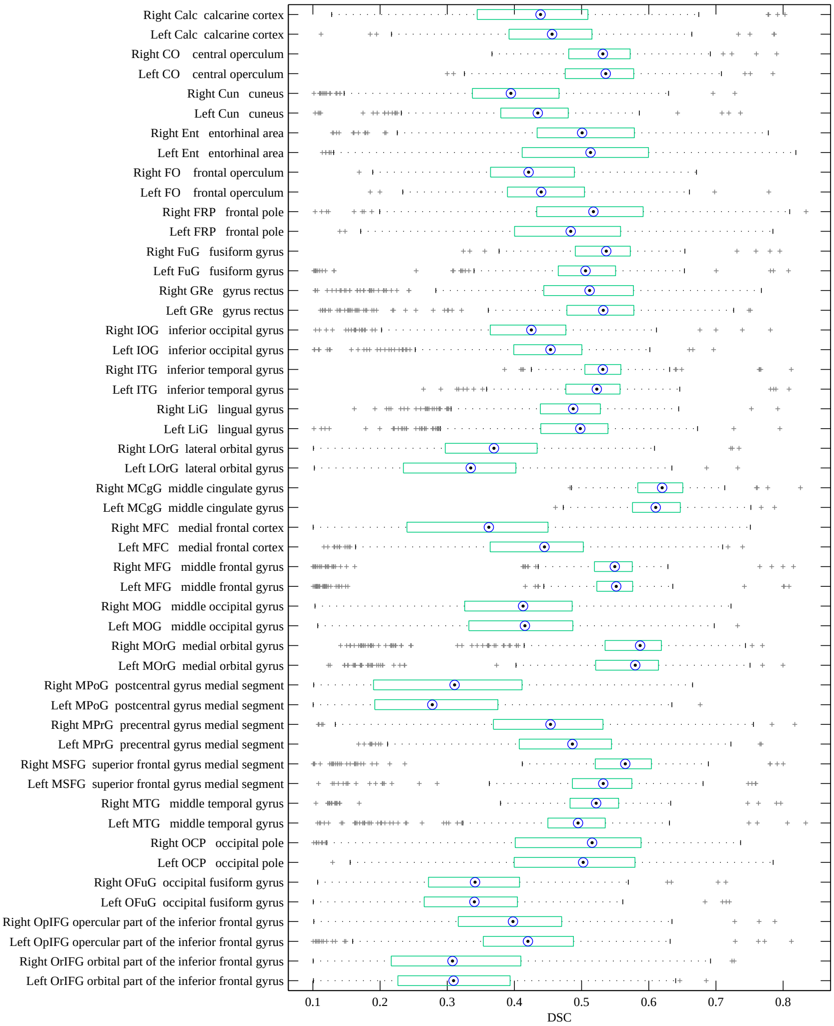


Figure 4.14: Distribution of pairwise overlap measures attained by the presented algorithm across different brain regions. For each boxplot, the central mark indicates the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points, while outliers are indicated by gray crosses.

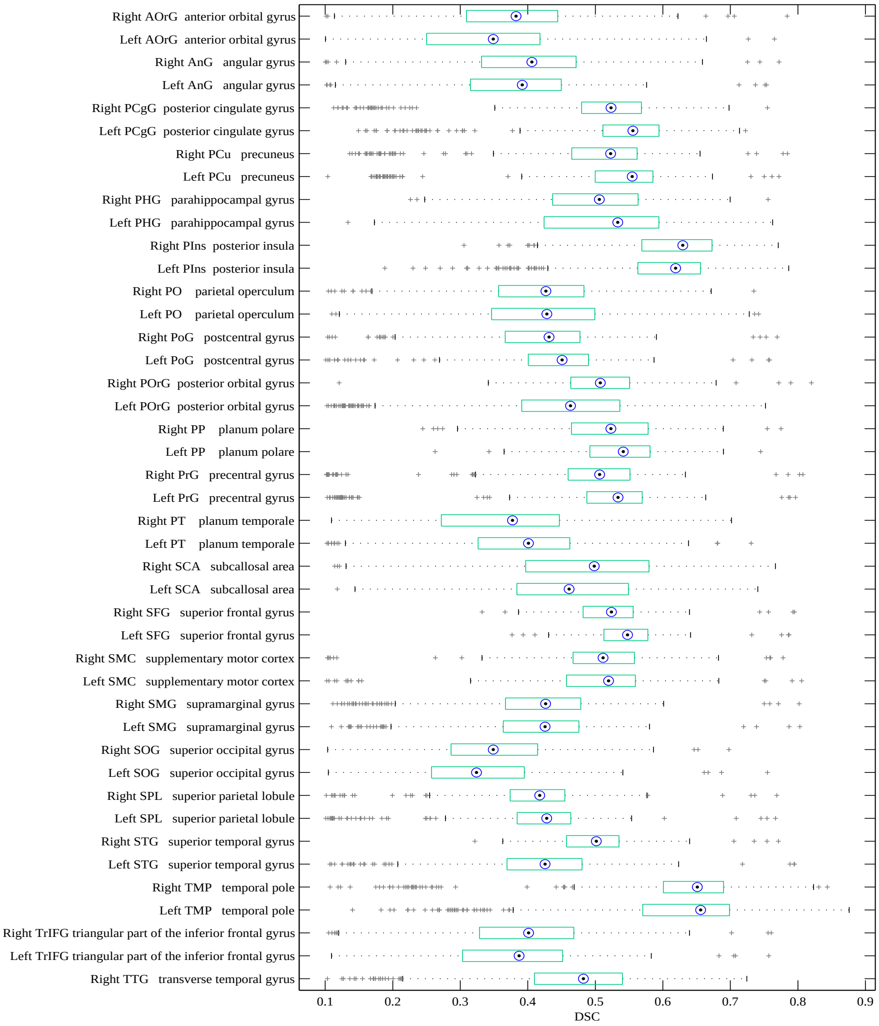


Figure 4.15: Distribution of pairwise overlap measures attained by the presented algorithm across different brain regions. For each boxplot, the central mark indicates the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points, while outliers are indicated by gray crosses.

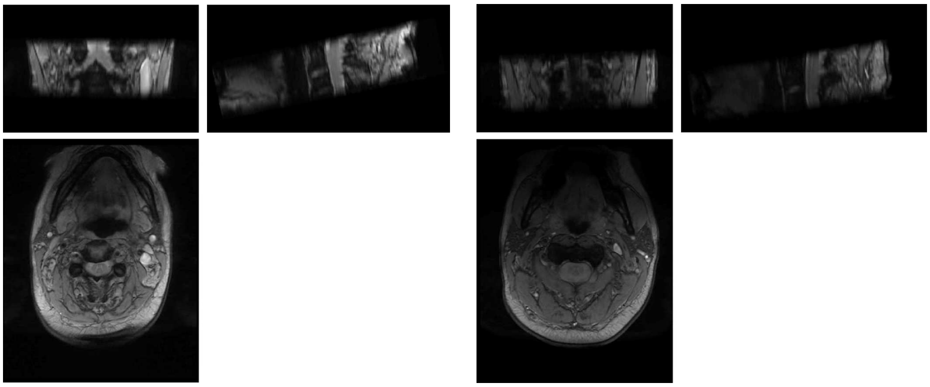
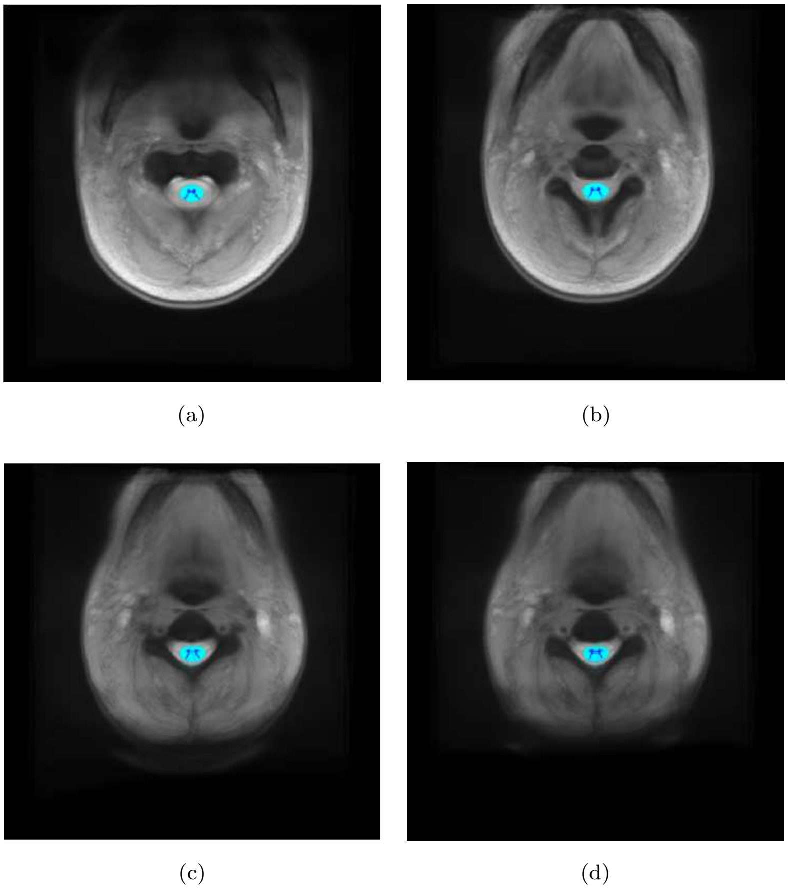


Figure 4.16: Orthogonal sections of two high resolution cervical MR volumes.



(a)

(b)

(c)

(d)

Figure 4.17: Gray and white matter spinal cord templates overlaying an average-shaped T2*-weighted image at different cervical levels between C2 (a) and C3 (d).

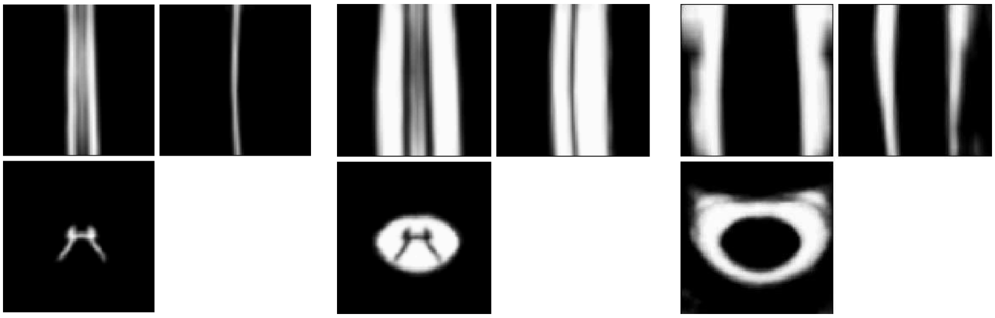


Figure 4.18: Gray matter, white matter and cerebrospinal fluid templates of the cervical cord.

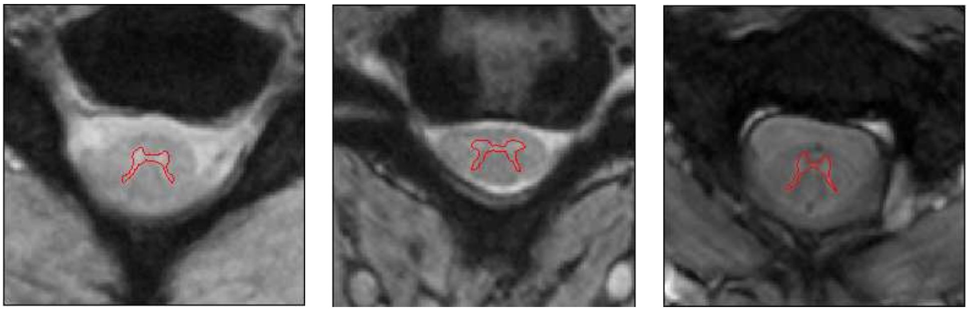


Figure 4.19: Individual gray matter segmentations of the cervical cord produced by the presented algorithm.

First of all, Dice score coefficients were computed to assess the amount of volumetric overlap between the automated and manual segmentations (Crum et al., 2006). The mean Euclidean surface distance was evaluated as an indicator of the contour mismatch between the two set of segmentations. Finally the skeletonized median distance, which compares thinned versions of the segmentations (Zhang and Suen, 1984), was used to assess global shape similarity.

Results are reported in Figure 4.20, for each of the four raters, while examples of individual automated segmentations are illustrated in Figure 4.19. It should be noted that results of these analyses were submitted to the Gray Matter Spinal Cord Segmentation Challenge held in 2016 during the 24th ISMRM annual meeting. Additional information on the challenge and its results can be found in Prados et al. (2017) or at the web page `http://cmictig.cs.ucl.ac.uk/niftyweb`. In particular, the method that achieved the best results in terms of Dice scores is the DEEPSEG method, which is based on the deep 3D convolutional encoder network with shortcut connections proposed by Brosch et al. (2016). However, the performance of the six evaluated algorithms was found to be significantly variable, depending on the selected accuracy metric. For instance, the approach presented here, in spite of having obtained significantly lower Dice scores compared to DEEPSEG, achieved much better results when evaluated in terms of maximal contour distance between manual and automated segmentations (i.e. Hausdorff surface distance). Thus, different methods might be more suitable for different applications. The generative framework proposed in this thesis is particularly convenient to perform statistical volumetric and morphometric analyses, which instead might be harder to implement using some of the competing techniques evalauted in Prados et al. (2017).

4.5. Summary

This chapter has illustrated some of the potential applications of the Bayesian modelling framework introduced in Chapter 3 to analyse neuroimaging data. In particular, the method has been applied to publicly available MR data sets of both the brain and the spinal cord, to construct average-shaped tissue probability maps. Unsupervised and semisupervised learning methods have been tested and compared, so as to evaluate advantages and limitations of the two approaches. The presented results suggest that

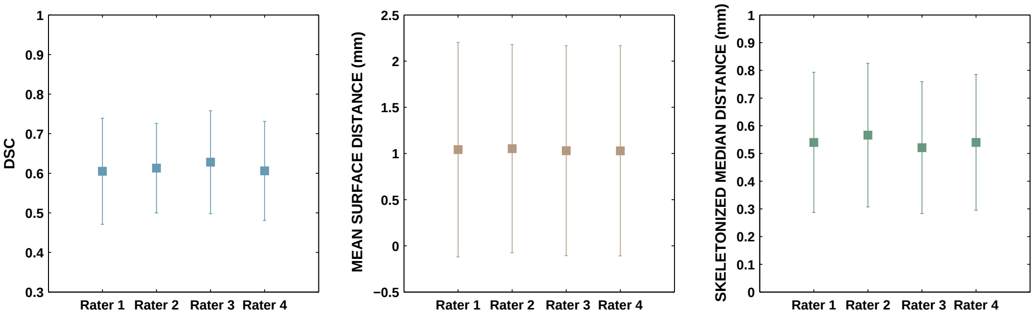


Figure 4.20: Distributions of accuracy metrics obtained by comparing the spinal cord gray matter segmentations produced by our algorithm to the manual labels generated by four trained human raters.

semisupervised learning is effective for differentiating tissue types whose intensity distributions substantially overlap, which is inherently very difficult to achieve in a fully unsupervised generative framework. This property is particularly valuable when multimodal data sets are not available. However, segmentation accuracy for the subcortical nuclei was found to be significantly lower than for cortical gray matter, which indicates that a fully supervised framework might be more suitable for discriminating anatomical structures that are in close proximity but exhibit very low image contrast.

Similarly, the behaviour of different deformation models, namely affine and small non-linear deformations, has been explored. Indeed, both models are somewhat suboptimal for the purpose of encoding anatomical shape variability, the first being too low dimensional, and the second not allowing to model large deviations from the average anatomy without sacrificing smoothness of the transformations and therefore topology preservation. Such limitations will be explicitly addressed in Chapter 6, by exploiting the large deformation diffeomorphic metric mapping (LDDMM) framework, which, in spite of introducing additional mathematical and computational complexity, is a much more powerful framework to represent anatomical shapes.

Chapter 5 instead will introduce a variational scheme to perform Bayesian inference on Gaussian mixture models applied to MRI data and will illustrate some of the advantages of this approach, compared to model fitting by point estimation techniques, such as maximum likelihood or maximum a posteriori estimation, which were both adopted in this chapter.

5

Variational inference for medical image segmentation

5.1. Introduction

In this chapter, the general principles underlying variational Bayesian inference are introduced, together with a computational framework that applies the variational Bayes (VB) approach to fit a generative Gaussian mixture model to the intensities of neuroimaging data. In particular, such a model is used to solve medical image segmentation problems and validated on both simulated and real brain MRI data.

5.2. Advantages and challenges of Bayesian inference

Many widely used image segmentation algorithms rely on probabilistic modelling techniques to fit the intensity distributions of images. These methods commonly operate by means of unsupervised clustering algorithms and assume that the data are drawn from mixture distributions, with different mixture components being associated to different tissue types (Ahmed et al., 2002; Chuang et al., 2006; Lee et al., 2008; Sfikas et al., 2007). In particular, Gaussian mixture models (GMM) have been extensively adopted as they provide a flexible and computationally efficient framework, which can be easily applied to solve the problem of automatically partitioning images into homogeneous regions (Dugas-Phocion et al., 2004; Greenspan et al., 2006; Guillemaud and Brady, 1997; Moon et al., 2002; Noe and Gee, 2001; Van Leemput et al., 1999b; Wells III et al., 1996; Woolrich et al., 2009; Zhang et al., 2001).

Intensity-based segmentation tools of this sort have been developed profusely over the past twenty years. Most of them either rely directly on an explicit Bayesian formulation, or exhibit an implicit probabilistic interpretation. Nevertheless almost all of them are based on maximum likelihood (ML) or maximum a posteriori (MAP) estimation of the model parameters (Ashburner and Friston, 2005; Greenspan et al., 2006; Kovacevic et al., 2002; Liang et al., 1994; Lorenzo-Valdés et al., 2004; Rajapakse and Kruggel, 1998; Van Leemput et al., 2003; Wyatt and Noble, 2003; Xiaohua et al., 2004b; Zhang et al., 2001), without exploiting the full potential of Bayesian inference.

Indeed, ML or MAP techniques often ensure mathematical tractability and sufficient segmentation accuracy for many applications. Nonetheless there is still a crucial theoretical point that makes these methods somewhat suboptimal, regardless of their practical convenience, which is that they just provide point estimates of the model parameters instead of full posterior probability distributions. In other words, information is missing on the posterior uncertainty in estimating unobserved variables, and this often results in the occurrence of overfitting as well as in the inability to perform model comparison (Attias, 1999). In practice, this also means that explicit confidence measures cannot be directly obtained for the estimated parameters, which is a significant drawback for potential clinical applications, where the risk of error or failure needs to be accurately assessed and quantified.

On the other hand, full Bayesian inference has been poorly explored in the field of medical image segmentation, in spite of a promising potential, which was shown for example by Woolrich and Behrens (2006) and Tian et al. (2011). The reason for this is most probably related to the computational challenges that arise when trying to evaluate the model evidence or the posterior probability distributions over the model parameters. In fact, very often and also for relatively simple models, integrating out all the unobserved variables turns out to be intractable in analytical form. On the other hand, numerical integration is generally impractical because either the dimensionality or the complexity of the problem would make the necessary computational resources prohibitive for real world applications.

One approach for dealing with the mathematical difficulties that arise in Bayesian inference is to make use of stochastic techniques to sample from the probability distributions that are of interest (Andrieu et al., 2003). In particular, Markov Chain Monte Carlo (MCMC) methods can provide rather accurate solutions at the expenses of a long processing time. As to be expected, the time required to reach convergence increases with the size of the data set. The result of this being the fact that, for large-scale problems, sampling techniques can become computationally impracticable. For example, the work of Iglesias et al. (2012b) is one among few attempts (da Silva, 2009; Fan et al., 2007; Kato, 2008) to exploit stochastic sampling methods to integrate out model parameters in the context of medical image segmentation. Their atlas-based segmentation approach takes into account the uncertainty in estimating coordinate mappings between individual test images and the reference anatomy. However, they report a running time of the sampling of approximately three hours, for a small anatomical structure like the hippocampus, which indicates that this approach might still be unfeasible outside the context of research.

A second family of approaches is based on introducing analytical approximations (Tierney and Kadane, 1986). For instance, one possibility is to approximate an unknown posterior probability distribution by an unnormalised Gaussian, centred at the mode of the actual posterior, or at one of the modes, if the distribution is multimodal. This is a general mathematical method (Fulks and Sather, 1961), known as Laplace approximation, which, in the context of probabilistic inference, overcomes many of the limitations of sampling techniques, since the number of required computations is much lower in this case. Nevertheless, depending on how different the actual posterior distribution is from a Gaussian, the method might provide a poor approximation. In particular the underlying Gaussian assumption might become inadequate for points that are far from the mode of the probability density function (Geisser et al., 1990).

Variational Bayes (VB) represents an alternative way of obtaining approximate solutions to inference problems. It relies on analytical approximations, as the Laplace method, and likewise it is much less computationally expensive than MCMC. However, the VB framework is more general and flexible than the Laplacian approach because, even if it usually constrains the posterior distributions to have a specific form or factorization (for the sake of computational convenience), such posteriors are not necessarily forced to be Gaussian. In other words, variational Bayesian inference permits finding a trade off between allowing sufficient complexity of the estimated posteriors and ensuring computational tractability. Stochastic variational algorithms have also been proposed (Hoffman et al., 2013).

Even if the estimated posteriors will almost never be exact, variational methods have proved to be more convenient than standard ML or MAP techniques, since, for a similar computational cost, they significantly alleviate the problems related to overfitting, which are intrinsic to the other methods. In other words, variational techniques open up the possibility of learning the optimal model structure (the one with highest generalisation capability) without performing *ad-hoc* cross-validation analyses (Attias, 1999; Bishop, 2006; Corduneanu and Bishop, 2001). Another interesting aspect of working within a VB framework is that it leads to a more general formulation of the EM algorithm, which has the same convergence properties and higher computational stability. For example, one significant limitation of ML estimation for mixture models, which is automatically addressed in a VB setting, is the presence of singular points of the likelihood function, which have to be avoided during optimisation to ensure numerical stability.

So far, very few authors have explored the applicability of the variational Bayes framework to perform medical image segmentation. Among them are Woolrich and Behrens (2006), who exploited variational inference to fit spatial mixture models to medical imaging data, while automatically tuning the parameter controlling regularisation, and Tian et al. (2011), who proposed an algorithm for segmenting brain MR data, which combines variational Bayes and genetic algorithms.

This chapter introduces an extension of the tissue classification algorithm presented by Ashburner and Friston (2005) and publicly distributed as part of the SPM12 software. Specifically, the maximum likelihood approach, adopted in Ashburner and Friston (2005) to estimate the Gaussian mixture parameters, is replaced by a Bayesian inference scheme, relying on variational approximations.

This approach, first of all, increases the robustness of the method, if suitable intensity priors are introduced, thus reducing significantly the chance of the algorithm failing to converge due to a mismatch or misregistration of the tissue probability maps with the individual scans. A second aspect that will be illustrated is how the fundamental problem of determining optimal model complexity, that is, in this case, the number of Gaussian components, can be effectively addressed in a variational setting. Such a framework, in fact, implicitly implements an automatic relevance determination scheme, where redundant mixture components are automatically pruned out of the model (Bishop, 2006). Finally, a parametric empirical Bayes approach will be presented, which can serve to learn informative intensity priors from sufficiently large data sets.

5.3. Background on variational Bayes

Variational Bayesian inference can be formulated as a maximisation problem. Let us consider the marginal log likelihood (i.e. log model evidence), $\log p(\mathbf{X})$, given by

$$\log p(\mathbf{X}) = \log \int p(\mathbf{X}, \boldsymbol{\Upsilon})\, \mathrm{d}\boldsymbol{\Upsilon} \, , \tag{5.1}$$

where $\mathbf{X}$ indicates the observed data and $\boldsymbol{\Upsilon} = \{\mathbf{Z}, \Theta\}$ is a set of unobserved variables (model parameters Θ and latent variables $\mathbf{Z}$).

After introducing a distribution $q(\boldsymbol{\Upsilon})$ over the unobserved variables, the log evidence in (5.1) can be re-expressed as

$$\begin{aligned}
\log p(\mathbf{X}) &= \int q(\boldsymbol{\Upsilon}) \log p(\mathbf{X})\, \mathrm{d}\boldsymbol{\Upsilon} \\
&= \int q(\boldsymbol{\Upsilon}) \log \left\{ \frac{p(\mathbf{X}, \boldsymbol{\Upsilon})}{p(\boldsymbol{\Upsilon}|\mathbf{X})} \right\} \mathrm{d}\boldsymbol{\Upsilon} \\
&= \int q(\boldsymbol{\Upsilon}) \log \left\{ \frac{p(\mathbf{X}, \boldsymbol{\Upsilon})}{q(\boldsymbol{\Upsilon})} \cdot \frac{q(\boldsymbol{\Upsilon})}{p(\boldsymbol{\Upsilon}|\mathbf{X})} \right\} \mathrm{d}\boldsymbol{\Upsilon} \\
&= \int q(\boldsymbol{\Upsilon}) \log \left\{ \frac{p(\mathbf{X}, \boldsymbol{\Upsilon})}{q(\boldsymbol{\Upsilon})} \right\} \mathrm{d}\boldsymbol{\Upsilon} + \int q(\boldsymbol{\Upsilon}) \log \left\{ \frac{q(\boldsymbol{\Upsilon})}{p(\boldsymbol{\Upsilon}|\mathbf{X})} \right\} \mathrm{d}\boldsymbol{\Upsilon} \, .
\end{aligned} \tag{5.2}$$

which is a decomposition of $\log p(\mathbf{X})$ that holds for any $q(\boldsymbol{\Upsilon})$.

The second integral in the last line of (5.2) is the Kullback-Leibler divergence $D_{KL}(q\|p)$ between $q(\boldsymbol{\Upsilon})$, which is a variational approximating posterior, and $p(\boldsymbol{\Upsilon}|\mathbf{X})$, which is the true posterior distribution (Bishop, 2006).

Since $D_{KL}(q\|p) \geq 0$, the first integral in the last line of (5.2) defines a lower bound $\mathcal{L}(q)$ on the logarithm of the model evidence

$$\log p(\mathbf{X}) \geq \mathcal{L}(q) = \int q(\boldsymbol{\Upsilon}) \log \left\{ \frac{p(\mathbf{X}, \boldsymbol{\Upsilon})}{q(\boldsymbol{\Upsilon})} \right\} \mathrm{d}\boldsymbol{\Upsilon} \, . \tag{5.3}$$

The previous statement can also be derived from (5.1) by applying Jensen's inequality.

In summary equation (5.2) can be rewritten as (Tzikas et al., 2008)

$$\log p(\mathbf{X}) = \mathcal{L}(q) + D_{KL}(q\|p) \ . \tag{5.4}$$

$D_{KL}(q\|p)$ is always non-negative and, in particular, it is equal to zero if and only if $q(\boldsymbol{\Upsilon}) = p(\boldsymbol{\Upsilon}|\mathbf{X})$. In such a case the variational posterior is an exact solution and the lower bound is exactly equal to the evidence. In all the other cases, $D_{KL}(q\|p) > 0$ and $\mathcal{L}(q) < \log p(\mathbf{X})$, which means that $q(\boldsymbol{\Upsilon})$ is an approximate posterior.

In summary, the inference problem can be solved by maximising the functional $\mathcal{L}(q)$ with respect to the distribution $q(\boldsymbol{\Upsilon})$, which is equivalent to minimising the Kullback-Leibler divergence between the variational and the true posterior distribution. It should be noted that the approach adopted here constitutes a generalisation of the scheme described in Section 3.4.1, which corresponds to the standard expectation-maximisation (EM) framework. In fact, in the case that is considered here, variational posterior distributions are introduced not only on the latent variables but also on the model parameters.

The lower bound on the model evidence (negative variational free energy) can be further decomposed as

$$\mathcal{L}(q) = \int q(\boldsymbol{\Upsilon}) \log p(\mathbf{X}|\boldsymbol{\Upsilon}) \mathrm{d}\boldsymbol{\Upsilon} + \int q(\boldsymbol{\Upsilon}) \log \left\{ \frac{p(\boldsymbol{\Upsilon})}{q(\boldsymbol{\Upsilon})} \right\} \mathrm{d}\boldsymbol{\Upsilon} \ . \tag{5.5}$$

This shows that the lower bound comprises a likelihood term which is equal to the expected value of the log likelihood $\log p(\mathbf{X}|\boldsymbol{\Upsilon})$ under the variational posterior $q(\boldsymbol{\Upsilon})$

$$\mathcal{L}_1 = \int q(\boldsymbol{\Upsilon}) \log p(\mathbf{X}|\boldsymbol{\Upsilon}) \mathrm{d}\boldsymbol{\Upsilon} = \mathbb{E}_{\boldsymbol{\Upsilon}} \left[\log p(\mathbf{X}|\boldsymbol{\Upsilon}) \right] \ , \tag{5.6}$$

and a regularising term which is the negative Kullback-Leibler divergence beetween the approximating posterior $q(\boldsymbol{\Upsilon})$ and the prior distribution over the unobserved variables $p(\boldsymbol{\Upsilon})$ (Attias, 1999)

$$\mathcal{L}_2 = \int q(\boldsymbol{\Upsilon}) \log \left\{ \frac{p(\boldsymbol{\Upsilon})}{q(\boldsymbol{\Upsilon})} \right\} \mathrm{d}\boldsymbol{\Upsilon} = -D_{KL}(q\|p_0) \ . \tag{5.7}$$

This last term penalizes overly complex or implausible models (Occam factor).

While in principle no constrains are placed on $q(\boldsymbol{\Upsilon})$, a commonly adopted strategy consists in restricting the space of $q(\boldsymbol{\Upsilon})$ so as to ensure mathematical tractability, which also means that $D_{KL}(q \| p) > 0$, or, in other words, that $q(\boldsymbol{\Upsilon}) \neq p(\boldsymbol{\Upsilon}|\mathbf{X})$. In particular, it is often convenient to assume that $q(\boldsymbol{\Upsilon})$ factorizes into a product of terms (Parisi and Zamponi, 2010), each one involving just a subset of $\boldsymbol{\Upsilon}$ (mean field theory)

$$q(\boldsymbol{\Upsilon}) = \prod_{s=1}^{S} q_s(\boldsymbol{\Upsilon}_s) \ . \tag{5.8}$$

In such a case, the lower bound depends on the generic factor $q_{\hat{s}}(\boldsymbol{\Upsilon}_{\hat{s}})$ as follows (Bishop, 2006)

$$\begin{aligned}\mathcal{L}(q_{\hat{s}}) &= \int q_{\hat{s}} \prod_{s \neq \hat{s}} q_s \log \left\{ \frac{p(\mathbf{X}, \boldsymbol{\Upsilon})}{q_{\hat{s}} \prod_{s \neq \hat{s}} q_s} \right\} \mathrm{d}\boldsymbol{\Upsilon} \\ &= \int q_{\hat{s}} \ \mathbb{E}_{s \neq \hat{s}}[\log p(\mathbf{X}, \boldsymbol{\Upsilon})] \mathrm{d}\boldsymbol{\Upsilon}_{\hat{s}} - \int q_{\hat{s}} \log q_{\hat{s}} \mathrm{d}\boldsymbol{\Upsilon}_{\hat{s}} + \text{const} \\ &= -D_{KL}(q_{\hat{s}} \, \| \, \hat{p}(\mathbf{X}, \boldsymbol{\Upsilon}_{\hat{s}})) + \text{const} \ ,\end{aligned} \tag{5.9}$$

with

$$\hat{p}(\mathbf{X}, \boldsymbol{\Upsilon}_{\hat{s}}) \propto \exp(\mathbb{E}_{s \neq \hat{s}}[\log p(\mathbf{X}, \boldsymbol{\Upsilon})]) \ . \tag{5.10}$$

Equation (5.9) shows that the optimal form of the factor $q_{\hat{s}}(\boldsymbol{\Upsilon}_{\hat{s}})$ corresponds to the one that minimises the Kullback-Leibler divergence between $q_{\hat{s}}(\boldsymbol{\Upsilon}_{\hat{s}})$ and $\hat{p}(\mathbf{X}, \boldsymbol{\Upsilon}_{\hat{s}})$ as defined in (5.10). Therefore $q_{\hat{s}}(\boldsymbol{\Upsilon}_{\hat{s}}) = \hat{p}(\mathbf{X}, \boldsymbol{\Upsilon}_{\hat{s}})$.

It should be noted that this solution is not analytical, since the different factors have optimal forms that depend on one another. As a result, the natural approach for solving this variational optimisation problem consist in iteratively updating each factor given the most recent forms of the other ones. This leads to a scheme that turns out to be very similar to the structure of the EM algorithm (Bishop, 2006; Tzikas et al., 2008).

For some complex models, a fully Bayesian treatment of all unobserved variables might still be extremely impractical, if not impossible, even when variational approximations are used. However, one other advantage from adopting a VB approach is that its generality allows it to be combined with standard MAP and ML techniques in a unified and principled framework. If one of the subsets $\{\boldsymbol{\Upsilon}_s\}_{s=1,\dots,S}$ of the unobserved variables cannot be treated in a fully Bayesian manner, it is still possible to obtain MAP point estimates of the corresponding parameters. Such values are computed in a way that is a generalisation of the M-step in the EM algorithm. In particular, the function that needs to be optimised is the expectation of the logarithm of the joint probability of

$\mathbf{X}$ and $\mathbf{\Upsilon}$, $\mathbb{E}[\log p(\mathbf{X}, \mathbf{\Upsilon})]$. The main difference from the EM algorithm for ML, or MAP, estimation is that, in the VBEM case, expectations are computed not only over the latent variables of the model but also over all the model parameters that are described in terms of a full posterior distribution. In such a case however, as opposed to a fully Bayesian approach, the main disadvantage is that, since some unobserved variables cannot be integrated out to compute the evidence, model selection cannot be performed by comparison of the marginal likelihood, especially if the compared models have different complexity.

5.4. Data model

Let $\mathbf{X}$ denote the observed data, that is to say the intensities corresponding to D images of the same subject acquired with different modalities. The signal at voxel j can then be represented by a D-dimensional vector $\mathbf{x}_j \in \mathbb{R}^D$, with $j \in \{1, \ldots, N\}$.

Along the same line adopted in Chapter 3, the distribution of $\mathbf{x}_j$ can be modelled as a multivariate Gaussian mixture, consisting of K clusters, parametrised by mean vectors $\{\boldsymbol{\mu}_k\}_{k=1,\ldots,K}$ and covariance matrices $\{\mathbf{\Sigma}_k\}_{k=1,\ldots,K}$.

Moreover it is assumed here that the K Gaussians are partitioned into T subsets, corresponding to different tissue types. Let $\{C_t\}_{t=1,\ldots,T}$ denote these subsets, with $\bigcup_t^T C_t = \{1, \ldots, K\}$. This means that each tissue $t \in \{1, \ldots, T\}$ is itself represented by a Gaussian mixture, consisting of K_t components, with $\sum_t K_t = K$.

The prior probability of each voxel belonging any of the T tissue types is computed making use of a probabilistic anatomical atlas, indicated by $\{\pi_t(\boldsymbol{y})\}_{t=1,\ldots,T}$, where $\boldsymbol{y}$ is a continuous coordinate vector field. Such an atlas, which is considered as precomputed throughout this chapter rather than being estimated from the data as in Chapter 3, is warped non-linearly, exploiting a coordinate mapping $\phi(\boldsymbol{y})$, to give $\{\pi_t(\phi(\boldsymbol{y}))\}_{t=1,\ldots,T}$. A small deformation model is adopted here, parametrised by a discrete displacement field $\Theta_u = \{\mathbf{u}_j\}_{j=1,\ldots,N}$. In the meanwhile, it is assumed, for simplicity, that the images have already been affinely registered.

The tissue priors are allowed to be rescaled by a set of weights $\{w_t\}_{t=1,\ldots,T}$ to accommodate individual differences in tissue composition. This approach offers some additional flexibility for matching the priors to the individual data, by allowing a small amount of erosion or dilation of the tissue probability maps. Finally, a set of parameters $\{g_k\}_{k=1,\dots,K}$ denotes the normalised weights of the different Gaussians associated with one tissue type, so that

$$\forall t \in \{1, \dots, T\} : \sum_{k \in C_t} g_k = 1. \tag{5.11}$$

As a result, having introduced a set of binary latent variables $\mathbf{Z}$, the probability of $\mathbf{Z}$ given the tissue priors $\Theta_\pi = \{\pi_t\}_{t=1,\dots,T}$, the weights $\Theta_w = \{w_t\}_{t=1,\dots,T}$, the mixing coefficients $\Theta_g = \{g_k\}_{k=1,\dots,K}$ and the deformation parameters Θ_u, is given by

$$\begin{aligned} p(\mathbf{Z}|\Theta_\pi, \Theta_w, \Theta_g, \Theta_u) &= \prod_{j=1}^{N} \prod_{k=1}^{K} \left(g_k \frac{\pi_t(\mathbf{y}_j, \mathbf{u}_j)\, w_t}{\sum_s^T \pi_s(\mathbf{y}_j, \mathbf{u}_j)\, w_s} \right)^{z_{jk}} , \\ &= \prod_{j=1}^{N} \prod_{k=1}^{K} \left(\pi'_{jk}\right)^{z_{jk}} , \end{aligned} \tag{5.12}$$

where $t \in \{1, \dots, T\} : k \in C_t$ and all data points have been assumed independent. It should be noted that Θ_π is known *a priori*, while Θ_w, Θ_g and Θ_u have to be estimated from the observed data $\mathbf{X}$.

To correct for intensity non-uniformity artifacts, a multiplicative D-dimensional bias field, denoted by $\{\mathbf{b}_j(\Theta_\beta)\}_{j=1,\dots,N}$, is introduced, where Θ_β is a vector of parameters. Each of the D components of the bias is modelled as the exponential of a linear combination of discrete cosine transform basis functions (Ashburner and Friston, 2005).

The conditional distribution (i.e. class conditional density) of the observed intensities given the latent variables, the Gaussian parameters $\{\Theta_\mu, \Theta_\Sigma\}$ and the bias field parameters Θ_β, can be expressed as in Chapter 3, to give

$$p(\mathbf{X}|\mathbf{Z}, \Theta_\mu, \Theta_\Sigma, \Theta_\beta) = \prod_{j=1}^{N} \prod_{k=1}^{K} \left(\det(\mathbf{B}_j)\, \mathcal{N}(\mathbf{B}_j \mathbf{x}_j | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right)^{z_{jk}} , \tag{5.13}$$

with $\mathbf{B}_j = \mathrm{diag}(\mathbf{b}_j)$.

The joint probability of all the random variables, conditioned on the mixing proportions, which will serve to compute the variational lower bound, is given by

$$\begin{gathered} p(\mathbf{X}, \mathbf{Z}, \Theta_\mu, \Theta_\Sigma, \Theta_\beta, \Theta_u | \Theta_\pi, \Theta_w, \Theta_g) = \\ p(\mathbf{X}|\mathbf{Z}, \Theta_\mu, \Theta_\Sigma, \Theta_\beta) p(\mathbf{Z}|\Theta_\pi, \Theta_w, \Theta_g, \Theta_u) p(\Theta_\mu, \Theta_\Sigma) p(\Theta_u) p(\Theta_\beta) \, , \end{gathered} \tag{5.14}$$

where, as opposed to the ML approach of Chapter 3, priors on the means and covariances of the different classes have been introduced, which are modelled by conjugate Gaussian-

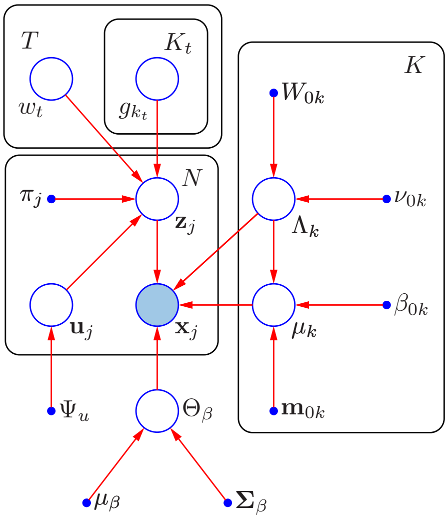


Figure 5.1: Directed acyclic graph representing the generative Gaussian mixture model adopted in this work for the purpose of segmenting neuroimaging data into tissue types. Large filled circles indicate the observed data (image intensities $\mathbf{X}$). Unfilled circles represent unobserved random variables (latent variables $\mathbf{Z}$, which encode class memberships, and model parameters Θ). Solid dots denote fixed hyperparameters. The observed intensities are assumed to be drawn from a Gaussian mixture distribution consisting of K components with means $\{\boldsymbol{\mu}_k\}_{k=1,\ldots,K}$ and covariance matrices $\{\boldsymbol{\Sigma}_k\}_{k=1,\ldots,K}$. Intensity non-uniformities are modelled through a multiplicative bias field parametrised by Θ_β. A smooth anatomical atlas set $\{\pi_t\}_{t=1,\ldots,T}$ is mapped onto the individual data by means of the deformation vector field encoded in $\{\mathbf{u}_j\}_{j=1,\ldots,N}$.

Wishart distributions

$$p(\Theta_\mu, \Theta_\Sigma) = \prod_{k=1}^{K} p(\boldsymbol{\mu}_k | \boldsymbol{\Sigma}_k^{-1}) p(\boldsymbol{\Sigma}_k^{-1}) \; , \tag{5.15}$$

with

$$p(\boldsymbol{\mu}_k | \boldsymbol{\Sigma}_k^{-1}) = \mathcal{N}(\boldsymbol{\mu}_k | \boldsymbol{m}_{0k}, \beta_{0k}^{-1} \boldsymbol{\Sigma}_k) \; , \tag{5.16}$$

$$p(\boldsymbol{\Sigma}_k^{-1}) = \mathcal{W}(\boldsymbol{\Sigma}_k^{-1} | \boldsymbol{W}_{0k}, \nu_{0k}) \; . \tag{5.17}$$

Such a choice is algebraically convenient, as it leads to posterior distributions having the same functional form as the priors (see Appendix A for a more detailed presentation of conjugate Gaussian-Wishart priors). The hyperparameters governing such priors will be indicated, in the remainder of this chapter, as

$$\Phi_0 = \{\beta_{0k}, \boldsymbol{m}_{0k}, \nu_{0k}, \boldsymbol{W}_{0k}\}_{k=1,\dots,K} \; . \tag{5.18}$$

The terms $p(\Theta_u)$ and $p(\Theta_\beta)$ represent prior probability distributions over the deformation and bias field parameters. Their function is to regularise the solution obtained through model fitting by penalising improbable parameters values. In doing so, they ensure greater physical plausibility of the resulting non-uniformity and deformation fields, while also improving numerical stability. Here the same regularisation scheme described in Ashburner and Friston (2005) is adopted. The question of how to determine the optimal amount of regularisation is beyond the scope of this work and therefore is not addressed here. Interestingly, such a problem could also be solved in a variational inference framework, as shown in Loic le Folgoc (2016); Simpson et al. (2015, 2012).

Given the model described above, a variational lower bound on the marginal likelihood $p(\mathbf{X}, \Theta_\beta, \Theta_u | \Theta_\pi, \Theta_w, \Theta_g)$ can be computed as

$$\mathcal{L} = \sum_{\mathbf{Z}} \iint q(\mathbf{Z}, \Theta_\mu, \Theta_\Sigma) \log \left\{ \frac{p(\mathbf{X}, \mathbf{Z}, \Theta_\mu, \Theta_\Sigma, \Theta_\beta, \Theta_u | \Theta_\pi, \Theta_w, \Theta_g)}{q(\mathbf{Z}, \Theta_\mu, \Theta_\Sigma)} \right\} \mathrm{d}\Theta_\mu \mathrm{d}\Theta_\Sigma \; . \tag{5.19}$$

To make the problem tractable, it is convenient to assume that the variational distribution $q(\mathbf{Z}, \Theta_\mu, \Theta_\Sigma)$ factorizes as $q(\mathbf{Z}, \Theta_\mu, \Theta_\Sigma) = q(\mathbf{Z})q(\Theta_\mu, \Theta_\Sigma)$, so that

$$\begin{aligned} \mathcal{L} = & \sum_{\mathbf{Z}} \iint q(\mathbf{Z}) q(\Theta_\mu, \Theta_\Sigma) \log p(\mathbf{X} | \mathbf{Z}, \Theta_\mu, \Theta_\Sigma, \Theta_\beta) \mathrm{d}\Theta_\mu \mathrm{d}\Theta_\Sigma \\ & + \sum_{\mathbf{Z}} \iint q(\mathbf{Z}) q(\Theta_\mu, \Theta_\Sigma) \log \left\{ \frac{p(\mathbf{Z} | \Theta_\pi, \Theta_w, \Theta_g, \Theta_u) p(\Theta_\mu, \Theta_\Sigma)}{q(\mathbf{Z}) q(\Theta_\mu, \Theta_\Sigma)} \right\} \mathrm{d}\Theta_\mu \mathrm{d}\Theta_\Sigma \\ & + p(\Theta_\beta) + p(\Theta_u) \; . \end{aligned} \tag{5.20}$$

The described probabilistic model can be represented by a directed acyclic graph, as shown in Figure 5.1.

5.5. Model learning

The statistical model described in the previous section can be fit to neuroimaging data by means of an iterative learning scheme, which constitutes a generalisation of the expectation-maximisation (EM) algorithm for maximum likelihood estimation.

In this instance, the aim is to obtain a variational posterior distribution $q(\mathbf{Z})q(\Theta_\mu, \Theta_\Sigma)$, maximum a posteriori estimates of $\{\Theta_u, \Theta_\beta\}$ and maximum likelihood estimates of $\{g_k\}_{k=1,\ldots,K}$ and $\{w_t\}_{t=1,\ldots,T}$.

5.5.1. Variational E-step

Similarly to the EM algorithm, its variational generalisation, namely variational Bayes expectation maximisation (VBEM), can be decomposed into two main steps, a variational E-step (VE) and a variational M-step (VM). In the first VE-step, the functional $\mathcal{L}$ of equation (5.19) is maximised with respect to the posterior factor $q(\mathbf{Z})$ over the latent variables (Bishop, 2006). Making use of (5.10) it is possible to derive

$$\begin{aligned} q(\mathbf{Z}) \propto \exp\big(& \log p(\mathbf{Z}|\Theta_\pi, \Theta_w, \Theta_g, \Theta_u) \\ & + \mathbb{E}_{\Theta_\mu,\Theta_\Sigma}\left[\log p(\mathbf{X}|\mathbf{Z}, \Theta_\mu, \Theta_\Sigma, \Theta_\beta)\right]\big) \ , \end{aligned} \tag{5.21}$$

and, having defined

$$\log \rho_{jk} = \log p(\mathbf{Z}|\Theta_\pi, \Theta_w, \Theta_g, \Theta_u) + \mathbb{E}_{\Theta_\mu,\Theta_\Sigma}\left[\log p(\mathbf{X}|\mathbf{Z}, \Theta_\mu, \Theta_\Sigma, \Theta_\beta)\right] \ , \tag{5.22}$$

it follows that

$$q(\mathbf{Z}) \propto \prod_{j=1}^{N}\prod_{k=1}^{K} (\rho_{jk})^{z_{jk}} \ . \tag{5.23}$$

Normalising of this variational posterior distribution gives

$$q(\mathbf{Z}) = \prod_{j=1}^{N}\prod_{k=1}^{K} \left(\frac{\rho_{jk}}{\sum_{c=1}^{K}\rho_{jc}}\right)^{z_{jk}} = \prod_{j=1}^{N}\prod_{k=1}^{K} (\gamma_{jk})^{z_{jk}} \ . \tag{5.24}$$

The quantity ρ_{jk} can be computed from (5.22) to give

$$\begin{aligned}\rho_{jk} =& \exp\big(\log \pi'_{jk} - \frac{D}{2}\log(2\pi) + \frac{1}{2}\mathbb{E}_{\boldsymbol{\Sigma}_k}\big[\log|(\boldsymbol{\Sigma}_k)^{-1}|\big] \\ &-\frac{1}{2}\mathbb{E}_{\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k}\big[(\mathbf{B}_j\mathbf{x}_j - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\mathbf{B}_j\mathbf{x}_j - \boldsymbol{\mu}_k)\big]\big) \ .\end{aligned} \tag{5.25}$$

The expectations that appear in (5.25) have to be computed with respect to the current estimate of the variational posterior distribution on $\{\boldsymbol{\mu}_k\}_{k=1,\dots,K}$ and $\{\boldsymbol{\Sigma}_k\}_{k=1,\dots,K}$, which will in turn be updated during the subsequent VM-step (see Appendix A for further details on how to compute these expected values under a Gaussian-Wishart distribution).

The class probability vectors $\{\boldsymbol{\gamma}_j\}_{j=1,\dots,N}$, which are evaluated during the VE-step represent expectations of the latent variables, with respect to their posterior variational distribution (i.e. responsibilities). They can be used to compute the following sufficient statistics of the observed data (Bishop, 2006), which will serve during the VM-step, as explained in the following section

$$\begin{aligned} s_{0k} &= \sum_{j=1}^{N}\gamma_{jk} \ , \\ \boldsymbol{s}_{1k} &= \sum_{j=1}^{N}\gamma_{jk}\mathbf{B}_j\mathbf{x}_j \ , \\ \boldsymbol{S}_{2k} &= \sum_{j=1}^{N}\gamma_{jk}(\mathbf{B}_j\mathbf{x}_j)(\mathbf{B}_j\mathbf{x}_j)^T \ . \end{aligned} \tag{5.26}$$

It should be noted that the computational complexity of this VE-step is identical to that of the E-step in the standard EM algorithm for Gaussian mixture model fitting, as derived in Chapter 3.

5.5.2. Variational M-step

During the VM-step, an approximate solution for the posterior distribution $q(\Theta_\mu, \Theta_\Sigma)$ is derived. Making again use of equation (5.10) gives

$$q(\Theta_\mu,\Theta_\Sigma) \propto \exp\left\{\sum_{j=1}^{N}\sum_{k=1}^{K}\gamma_{jk}\log\mathcal{N}(\mathbf{B}_j\mathbf{x}_j|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) + \sum_{k=1}^{K}\log p(\Theta_\mu,\Theta_\Sigma)\right\} . \tag{5.27}$$

It can be proved (see Appendix B) that the posterior distribution on the means and covariances of the different classes takes the same form as the corresponding prior

(Bishop, 2006), that is

$$q(\Theta_\mu, \Theta_\Sigma) = \prod_{k=1}^{K} q(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k^{-1}) q(\boldsymbol{\Sigma}_k^{-1}) \ , \tag{5.28}$$

with

$$q(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k^{-1}) = \mathcal{N}(\boldsymbol{\mu}_k|\boldsymbol{m}_k, \beta_k^{-1}\boldsymbol{\Sigma}_k) \ , \tag{5.29}$$

$$q(\boldsymbol{\Sigma}_k^{-1}) = \mathcal{W}(\boldsymbol{\Sigma}_k^{-1}|\boldsymbol{W}_k, \nu_k) \ . \tag{5.30}$$

The hyperparameters that govern these posterior distribution are

$$\Phi = \{\beta_k, \boldsymbol{m}_k, \nu_k, \boldsymbol{W}_k\}_{k=1,\dots,K} \ , \tag{5.31}$$

and they can be computed as a function of the prior hyperparameters and the sufficient statistics, obtained in the previous VE-step, as follows (see Appendix B for the detailed mathematical derivation)

$$\begin{aligned}
\beta_k &= \beta_{0k} + s_{0k} \ , \\
\boldsymbol{m}_k &= \frac{\beta_{0k}\boldsymbol{m}_{0k} + \boldsymbol{s}_{1k}}{\beta_{0k} + s_{0k}} \ , \\
\boldsymbol{W}_k^{-1} &= \boldsymbol{W}_{0k}^{-1} + \boldsymbol{S}_{2k} + \frac{\beta_{0k} s_{0k} \boldsymbol{m}_{0k}\boldsymbol{m}_{0k}^T}{\beta_{0k} + s_{0k}} - \frac{\boldsymbol{s}_{1k}\boldsymbol{s}_{1k}^T}{\beta_{0k} + s_{0k}} \\
&\quad - \frac{\beta_{0k}\boldsymbol{s}_{1k}\boldsymbol{m}_{0k}^T}{\beta_{0k} + s_{0k}} - \frac{\beta_{0k}\boldsymbol{m}_{0k}\boldsymbol{s}_{1k}^T}{\beta_{0k} + s_{0k}} \ , \\
\nu_k &= \nu_{0k} + s_{0k} \ .
\end{aligned} \tag{5.32}$$

The point estimates of the mixing proportions $\{g_k\}_{k=1,\dots,K}$ within each tissue type and of the tissue weights $\{w_t\}_{t=1,\dots,T}$ can instead be updated by means of the following ML estimators

$$g_k = \frac{s_{0k}}{\sum_{c \in C_t} s_{0c}} \ , \tag{5.33}$$

$$w_t = \frac{\sum_{k \in C_t} s_{0k}}{\displaystyle\sum_{j=1}^{N} \frac{\pi_t(\mathbf{y}_j, \mathbf{u}_j)}{\sum_{s=1}^{T} \pi_s(\mathbf{y}_j, \mathbf{u}_j) w_s}} \ . \tag{5.34}$$

5.5.3. Computing the lower bound

The lower bound of equation 5.19 can be easily evaluated, once the sufficient statistics and the variational posterior distributions have been computed (Bishop, 2006), by

$$\begin{aligned}\mathcal{L} = \; & \mathbb{E}_{\mathbf{Z},\Theta_\mu,\Theta_\Sigma}[\log p(\mathbf{X}|\mathbf{Z},\Theta_\mu,\Theta_\Sigma,\Theta_\beta)] + \mathbb{E}_{\mathbf{Z}}[\log p(\mathbf{Z}|\Theta_\pi,\Theta_w,\Theta_g,\Theta_u)] \\ & + \mathbb{E}_{\Theta_\mu,\Theta_\Sigma}[\log p(\Theta_\mu,\Theta_\Sigma)] + \log p(\Theta_u) + \log p(\Theta_\beta) \\ & - \mathbb{E}_{\mathbf{Z}}[\log q(\mathbf{Z})] - \mathbb{E}_{\Theta_\mu,\Theta_\Sigma}[\log q(\Theta_\mu,\Theta_\Sigma)] \;,\end{aligned} \tag{5.35}$$

with

$$\begin{aligned} & \mathbb{E}_{\mathbf{Z},\Theta_\mu,\Theta_\Sigma}\big[\log p(\mathbf{X}|\mathbf{Z},\Theta_\mu,\Theta_\Sigma,\Theta_\beta)\big] = \\ & \frac{1}{2}\sum_{k=1}^{K} s_{0k}\,\mathbb{E}\big[\log|\mathbf{\Sigma}_k^{-1}|\big] - D\log(2\pi) - \frac{D}{\beta_k} \\ & - \frac{1}{2}\sum_{k=1}^{K} s_{0k}\nu_k \boldsymbol{m}_k^T \boldsymbol{W}_k \boldsymbol{m}_k \\ & - \frac{1}{2}\sum_{k=1}^{K} \nu_k \operatorname{Tr}(\boldsymbol{W}_k \boldsymbol{S}_{2k} - 2\boldsymbol{s}_{1k}\boldsymbol{m}_k^T \boldsymbol{W}_k) \\ & + \sum_{j=1}^{N}\sum_{k=1}^{K} \gamma_{jk} \log|\mathbf{B}_j| \;.\end{aligned} \tag{5.36}$$

$$\mathbb{E}_{\mathbf{Z}}[\log p(\mathbf{Z}|\Theta_\pi,\Theta_w,\Theta_g,\Theta_u)] = \sum_{j=1}^{N}\sum_{k=1}^{K} \gamma_{jk} \log \pi'_{jk} \;. \tag{5.37}$$

$$\begin{aligned} & \mathbb{E}_{\Theta_\mu,\Theta_\Sigma}[\log p(\Theta_\mu,\Theta_\Sigma)] = \\ & + \frac{1}{2}\sum_{k=1}^{K}\left\{ D\log\frac{\beta_{0k}}{2\pi} - D\frac{\beta_{0k}}{\beta_k}\right\} + 2K\log B_W(\boldsymbol{W}_{0k},\nu_{0k}) \\ & - \sum_{k=1}^{K}\left\{ \frac{\nu_k}{2}\operatorname{Tr}\big((\boldsymbol{W}_{0k}^{-1} + \beta_{0k}(\boldsymbol{m}_k - \boldsymbol{m}_{0k})(\boldsymbol{m}_k - \boldsymbol{m}_{0k})^T)\boldsymbol{W}_k\big)\right. \\ & \left. + \mathbb{E}\big[\log|\mathbf{\Sigma}_k^{-1}|\big](\nu_{0k} - D)\right\} \;.\end{aligned} \tag{5.38}$$

$$\mathbb{E}_{\mathbf{Z}}[\log q(\mathbf{Z})] = \sum_{j=1}^{N}\sum_{k=1}^{K} \gamma_{jk} \log \gamma_{jk} \;. \tag{5.39}$$

$$\begin{aligned}
&\mathbb{E}_{\Theta_\mu,\Theta_\Sigma}[\log q(\Theta_\mu,\Theta_\Sigma)] = \\
&\sum_{k=1}^{K}\left\{\frac{1}{2}D\left(\log\frac{\beta_k}{2\pi}-1-\nu_k\right)+\log B_W(\boldsymbol{W}_k,\nu_k)\right. \\
&\left.+\,\mathbb{E}\big[\log|\boldsymbol{\Sigma}_k^{-1}|\big]\Big(\frac{1}{2}\nu_k-D\Big)\right\}. \qquad (5.40)
\end{aligned}$$

The term $B_W(\boldsymbol{W},\nu)$ in equations (5.38) and (5.40) indicates the normalising constant of a Wishart distribution parametrised by $\boldsymbol{W}$ and ν.

5.5.4. Estimating the bias field and the deformations

In order to estimate optimal parameters to represent the bias and the deformation fields, the lower bound of equation (5.35) has to be maximised, at each iteration of the algorithm, with respect to the parameters Θ_β and Θ_u, respectively. A closed form solution does not exist in this case, so recourse to numerical optimisation techniques cannot be avoided.

The two optimisation problems can be formalised as follows

$$\begin{aligned}
\hat{\Theta}_\beta &= \arg\max_{\Theta_\beta}\ \Big\{\mathbb{E}_{\mathbf{Z},\Theta_\mu,\Theta_\Sigma}\big[\log p(\mathbf{X}|\mathbf{Z},\Theta_\mu,\Theta_\Sigma,\Theta_\beta)\big]+\log p(\Theta_\beta)\Big\}\,, \\
\hat{\Theta}_u &= \arg\max_{\Theta_u}\ \Big\{\mathbb{E}_{\mathbf{Z}}\big[\log p(\mathbf{Z}|\Theta_\pi,\Theta_w,\Theta_g,\Theta_u)\big]+\log p(\Theta_u)\Big\}\,.
\end{aligned} \qquad (5.41)$$

Along the same line of Chapter 3, the problem is addressed here by making use of gradient-based optimisation techniques, such as the Gauss-Newton method (Bertsekas, 1999), or the Levenberg-Marquardt method (Moré, 1978), which are robust and fast converging strategies. This involves computing the first and second derivatives of $\mathcal{L}$ with respect to Θ_β and Θ_u. The resulting update rules are not significantly different from the ones reported in Chapter 3, except for having to compute additional expectations with respect to the Gaussian posteriors, therefore further mathematical details are omitted here.

Additionally, since the registration problem is formulated by means of a very high dimensional parametrisation, a multigrid scheme, with the same implementation described in Ashburner (2007), is used to solve numerically the Gauss-Newton update of Θ_u.

5.5.5. Empirical Bayes learning of intensity priors

The hyperparameters Φ_0 should reflect prior beliefs on how signal intensities are likely to be distributed within each tissue type. Having adopted a Gaussian-Wishart parametrisation, the following hyperparameter setting ensures minimally informative and yet proper (i.e. integrable) priors

$$\beta_{0k} = 0.01 \wedge \nu_{0k} = D - 0.99 \implies p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \simeq \text{const} . \tag{5.42}$$

With such a choice, the posterior distributions of the Gaussian parameters $\{\Theta_\mu, \Theta_\Sigma\}$ would essentially be determined by fitting of the data, similarly to the maximum likelihood framework, and the regularising term of the lower bound (see equation (5.7)) would reduce to the entropy of the posterior distributions.

On the contrary, choosing more informative priors can potentially increase the robustness of the algorithm by enforcing plausibility of the estimated posteriors and, at the same time, ensure faster convergence. However, defining pertinent priors is a non-trivial task, as ideally such priors should summarise information inferred from previously acquired data, rather than simple subjective beliefs. In other words, an appropriate hyperparameter configuration should be learned directly from large data sets, rather than arbitrarily set *a priori* (Lawrence and Platt, 2004; Raina et al., 2006; Seeger, 2002)

Interestingly, the hierarchical model described so far defines a natural framework for estimating empirical priors. In fact, supposing that posteriors $\{q_i(\Theta_\mu, \Theta_\Sigma)\}_{i=1,\ldots,M}$ have been estimated for a population of M subjects, the following lower bound on the marginal likelihood can be maximised with respect to the prior distribution $p(\Theta_\mu, \Theta_\Sigma)$

$$\begin{aligned} \mathcal{L} = \sum_{i=1}^{M} \sum_{\mathbf{Z}} \iint & q_i(\mathbf{Z}_i, \Theta_\mu, \Theta_\Sigma) \\ \times \log & \left\{ \frac{p_i(\mathbf{X}_i, \mathbf{Z}_i, \Theta_\mu, \Theta_\Sigma, \Theta_\beta, \Theta_u | \Theta_\pi, \Theta_w, \Theta_g)}{q_i(\mathbf{Z}_i, \Theta_\mu, \Theta_\Sigma)} \right\} \mathrm{d}\Theta_\mu \mathrm{d}\Theta_\Sigma . \end{aligned} \tag{5.43}$$

Additionally, since the functional form of this distribution is parametric and known (Gaussian-Wishart), standard non-linear optimisation techniques can be exploited to find maximum likelihood estimates of the hyperparameters Φ_0.

Indeed, the lower bound of equation (5.43) can be expressed as a function of Φ_0, as follows

$$
\begin{aligned}
\mathcal{L}(\Phi_0) &= \sum_{i=1}^{m} \int\int q_i(\Theta_\mu, \Theta_\Sigma) \log p(\Theta_\mu, \Theta_\Sigma)\ \mathrm{d}\Theta_\mu \mathrm{d}\Theta_\Sigma + \text{const} \\
&= \frac{1}{2} \sum_{i=1}^{M} \sum_{k=1}^{K} \Bigg\{ \mathbb{E}\big[\log |\boldsymbol{\Sigma}_{ik}^{-1}|\big] (\nu_{0k} - D) \\
&- \nu_k \operatorname{Tr}(\boldsymbol{W}_{0k}^{-1} \boldsymbol{W}_{ik} + \beta_{0k} (\boldsymbol{m}_{ik} - \boldsymbol{m}_{0k})(\boldsymbol{m}_{ik} - \boldsymbol{m}_{0k})^T \boldsymbol{W}_k) \Bigg\} \\
&+ \frac{M}{2} \sum_{k=1}^{K} D \log \frac{\beta_{0k}}{2\pi} - D \sum_{i=1}^{M} \sum_{k=1}^{K} \frac{\beta_{0k}}{\beta_{ik}} \\
&+ 2M \sum_{k=1}^{K} \log B_W(\boldsymbol{W}_{0k}, \nu_{0k}) + \text{const}\ ,
\end{aligned}
\tag{5.44}
$$

where B_W indicates the normalising constant of a Wishart distribution. The first and second derivatives of $\mathcal{L}(\Phi_0)$, which are useful to solve this optimisation problem using gradient-based techniques, are reported in Appendix C.

In practice, a convenient strategy for learning intensity priors consists in, first, initialising the hyperparameters so as to obtain weak priors, secondly, estimating the posterior distributions for a population of M subjects, finally, optimising $\mathcal{L}$ with respect to Φ_0. The estimates of the hyperparameters Φ_0 can then be further refined by using these empirical priors to re-estimate the posteriors and so on, thus leading to an iterative learning scheme, as illustrated by Algorithm 2.

5.6. Experimental results

This section will present a series of experiments that were performed to assess the validity of the proposed approach and to explore some of its properties and potential applications. The results presented in 5.6.1 were produced making use of synthetic data while the ones described in 5.6.2 were obtained on real, publicly available, MRI data.

5.6.1. Experiments on synthetic data

The performance of the variational algorithm presented in the previous section was first evaluated making use of simulated data produced by the Brainweb MRI simulator (Cocosco et al., 1997; Collins et al., 1998; Kwan et al., 1999). To assess the accuracy of

```
Input: a data set consisting of MR image intensities $\mathbf{X}$ of $M$ subjects
Output: posterior estimates of hyperparameters $\{\Phi_i\}_{i=1}^{M}$; MAP estimates of
        parameters $\{\mathbf{u}_i, \boldsymbol{\beta}_i\}_{i=1}^{M}$; ML estimates of parameters $\{\boldsymbol{g}_i, \boldsymbol{w}_i\}_{i=1}^{M}$
begin
    initialise $\{\Phi_0, \{\boldsymbol{\beta}_i, \mathbf{u}_i, \boldsymbol{g}_i, \boldsymbol{w}_i\}_{i=1}^{M}\}$;
    for $it = 1, \ldots, I_n$ do
        for each subject $i$ do
            for $subit = 1, \ldots, I_m$ do
                VE-step:
                evaluate $q_i(\mathbf{Z})$, (equation 5.25);
                VM-steps:
                (1) update $\Phi_i$, (equation 5.32);
                (2) update $\{\boldsymbol{g}_i, \boldsymbol{w}_i\}$, (equations 5.33 and 5.34);
                Bias update
                for $it_\beta = 1, \ldots, I_\beta$ do
                    update $\boldsymbol{\beta}_i$, (Section 3.4.2);
                end
                Deformations update
                for $it_u = 1, \ldots, I_u$ do
                    update $\mathbf{u}_i$, (Section 3.4.4);
                end
            end
        end
        Update intensity priors
        for $it_\Phi = 1, \ldots, I_\Phi$ do
            update $\Phi_0$, (Appendix C);
        end
    end
end
```

Algorithm 2: optimisation algorithm for joint segmentation and intensity prior learning from cross-sectional MR data sets.

brain tissue classification, twenty synthetic T1-weighted scans of healthy adult subjects (Aubert-Broche et al., 2006) were generated with the following MR simulation parameters: SFLASH (spoiled FLASH) sequence with TR=22 ms, TE=9.2 ms, flip angle=30 deg and 1 mm isotropic voxel size. Noise in these simulated scans has a standard deviation equal to 3% of the brightest image intensity, while no intensity inhomogeneities are present.

Such volumes were segmented using the algorithm presented in the previous section, after having set the following hyperparameter values, so as to obtain weakly informative intensity priors (WIP). This choice in fact permits quantifying the accuracy of the proposed method in the most general case, that is to say when no reliable information is available on the distribution of tissue intensities.

$$\begin{aligned}
\beta_{0k} &= 0.1, \\
\boldsymbol{m}_{0k} &= \frac{1}{N}\sum_{j=1}^{N}\mathbf{x}_j, \\
\nu_{0k} &= D - 0.9, \\
\boldsymbol{W}_{0k}^{-1} &= \frac{1}{N}\sum_{j=1}^{N}(\mathbf{x}_j - \boldsymbol{m}_{0k})(\mathbf{x}_j - \boldsymbol{m}_{0k})^T.
\end{aligned} \tag{5.45}$$

The tissue probability maps distributed with the SPM12 software were used as priors, with a number of Gaussian components for each tissue type equal to that used in SPM12 with the default settings.

The resulting segmentations were compared to the anatomical models used to generate the data by computing Dice similarity coefficients (DSC). Results, which are reported in Figure 5.2, indicate that the presented method can segment gray and white matter with an accuracy that is at least equal to that of some widely used, state-of the-art segmentation tools, such as the ones provided with SPM (Ashburner and Friston, 2005), FSL (Zhang et al., 2001) and Freesurfer (Fischl et al., 2004), whose performance was assessed in (Klauschen et al., 2009). In addition, Dice score coefficients attained by SPM12 on the same data were computed and reported in Figure 5.2 for comparison. For these experiments, both the proposed algorithm and the segmentation method implemented in SPM12 were applied after having down-sampled the data every 3 mm to reduce the run time.

The Brainweb database also provides multi-modality MR data, even if, in this case,

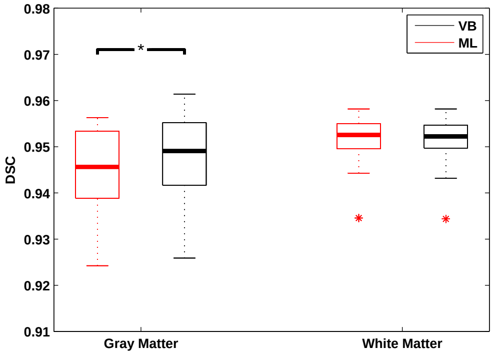


Figure 5.2: Dice similarity coefficients (DSC) between the gray and white matter segmentations produced by the presented algorithm (VB) and the underlying ground truth, for twenty simulated T1-weighted scans. DSC obtained with the ML algorithm provided with SPM12 are also reported for comparison. For each boxplot, the central mark indicates the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points, while outliers are indicated by red stars. Asterisks indicate statistically significant differences, assessed by means of paired t-tests with a significance threshold of 0.05.

only one anatomical model is available. To test the performance of the proposed algorithm in segmenting multispectral data, T1-weighted and T2-weighted volumes were simulated from the available anatomical model, with pulse sequence parameters reported in table 5.1 and then segmented with the same hyperparameter setting used for the previous experiment. To examine the behaviour of the algorithm with respect to noise, the analyses were repeated with three different levels of noise in the data (3%, 5% and 9% of the brightest intensity).

Results, which are summarised in Table 5.2, indicate that the presented method can successfully handle multi-modality data sets and that, even if the use of a single modality (e.g. T1-weighted) already ensures accurate segmentations, the availability of scans with different contrast can provide additional robustness to noise. A similar behaviour is exhibited by the ML algorithm provided with the SPM software (Table 5.2). However, comparison of the accuracy attained by the two methods indicates that the variational approach provides significantly better results, as assessed by means of a paired t-test performed on the entire set of scores, with a significance threshold of 0.05.

This simulated data was also used to assess the validity of bias field correction, as

Table 5.1: Simulation parameters selected to generate the synthetic data which was used to evaluate the accuracy of the presented VB algorithm in segmenting multispectral data. SFLASH and DSE indicate respectively a spoiled fast low angle shot and a dual spin echo sequence.

		Sequence	**TR (ms)**	**Flip angle (deg)**	**TE (ms)**	**Bias field**
Modality	T1w	SFLASH	18	30	10	20%
	T2w	DSE_LATE	3300	90	35,120	20%

Table 5.2: Dice similarity coefficients between the ground truth tissue labels and the segmentations produced by the presented algorithm (VB) and by the ML implementation provided with the SPM software. The experiments were performed on simulated normal brain scans (T1- and T2-weighted) for three different noise levels.

		Maximum Likelihood (ML)					
	Noise level	**3%**		**5%**		**9%**	
	Modality	**T1w**	**T1w and T2w**	**T1w**	**T1w and T2w**	**T1w**	**T1w and T2w**
Tissue	GM	0.93	0.90	0.91	0.90	0.87	0.88
	WM	0.95	0.95	0.93	0.94	0.88	0.89
		Variational Bayes (VB)					
	Noise level	**3%**		**5%**		**9%**	
	Modality	**T1w**	**T1w and T2w**	**T1w**	**T1w and T2w**	**T1w**	**T1w and T2w**
Tissue	GM	0.92	0.92	0.92	0.92	0.87	0.89
	WM	0.95	0.96	0.94	0.94	0.89	0.90

performed by the proposed method. To do so, Pearson's correlation coefficients were computed between the estimated non-uniformity fields and the ground truth. Results are shown in table 5.3, where the correlation coefficients attained by SPM ML-based segmentation algorithm are also reported. As to be expected the two methods perform quite similarly in estimating the non-uniformity field. In fact, they rely on the same parametrisation and optimisation of the bias. Nevertheless, because the accuracy in correcting intensity inhomogeneities depends heavily on how reliable the estimates of the Gaussian parameters are, the proposed algorithm, which takes into account the posterior uncertainty of such estimates, can outperform the maximum likelihood approach when noise in the data increases.

The presented method was implemented in MATLAB and, when subsampling the data every 3 mm, required a run time, for each individual segmentation, of approxi-

Table 5.3: Pearson's correlation coefficients between estimated and ground truth bias fields for the presented VB method and for SPM ML method.

	Noise level	**3%**		**5%**		**9%**	
	Algorithm	**VB**	**ML**	**VB**	**ML**	**VB**	**ML**
Modality	T1w	0.83	0.83	0.84	0.82	0.68	0.61
	T2w	0.88	0.88	0.89	0.89	0.88	0.70

mately 3 min 30 s, on a Quad-Core PC at 3.19 GHz with 12 GB RAM.

Learning GMM priors

Among the advantages of the variational framework, which is exploited here, is the fact that, like MAP estimation, it allows incorporating priors on the parameters modelling the intensity distribution of brain (and potentially non-brain) tissues. This form of *a priori* knowledge acts conjointly with the shape information carried by the tissue probability maps, thus ensuring additional robustness. The use of different intensity priors leads to differences in the estimated posteriors and segmentations, in the sense that the algorithm will try to simultaneously maximise the model fit, that is the likelihood of the data, while minimising the divergence between the prior and posterior probability distributions.

Determining suitable priors for each application, that is to say tissue or imaging modality, is a fundamental question. However, it should also be noted that the need to define priors does not limit the applicability of the method if compared to standard maximum likelihood techniques. In fact, whenever no information is available on what priors it is most convenient or correct to use, it is always possible to resort to minimally informative priors, which would simply let the algorithm determine the posterior distributions that explain the data best, given the assumption that all parameter settings, within the admissible parameter space, are equally (or almost equally) probable *a priori*.

As explained in Section 5.5.5 the variational framework presented in this chapter can be exploited to learn empirical priors on the Gaussian mixture parameters from large cross-sectional data sets. The efficacy of this procedure is demonstrated here using the same set of simulated T1-weighted scans employed for the previous experiments.

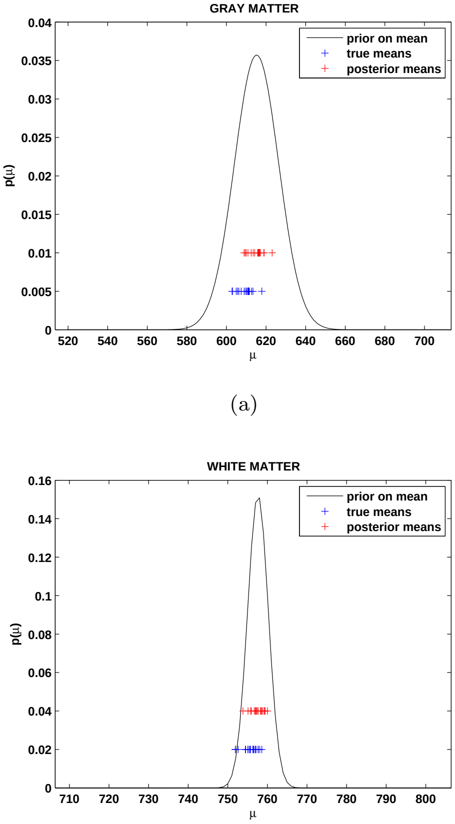
In particular, priors were learned relative to the intensities of gray and white matter (with one Gaussian component per tissue type), by first collecting posterior probability distributions for all of the subjects in the data set and then maximising the functional of equation (5.44) with respect to Φ_0. This optimisation problem was solved making use of a Gauss-Newton scheme, by iterating over optimising the priors and updating the posteriors so as reduce the chance of finding suboptimal solutions.

Results are depicted in Figure 5.3, which reports the estimated Gaussian priors on the mean intensity of gray (5.3a) and white (5.3b) matter. These should be compared to the modes of the corresponding posteriors, which are marked in the same figure by red crosses. The proposed empirical Bayes learning scheme captures very precisely the information encoded in the variational posteriors. In particular the more the posteriors are peaked and the more they overlap, the more informative the priors will be. If one or more posteriors have higher variance, this uncertainty will be directly reflected in the empirical priors, which will become less informative. This is the reason why the priors shown in Figure 5.3 are broader for gray than for white matter (in spite of a similar distribution of the modes), as the gray matter posteriors have higher variance compared to white matter.

The true means are also shown in Figure 5.3, marked by blue crosses. For white matter, they are extremely consistent with the estimated posterior means. In fact, for this data set, the presented algorithm exhibits higher accuracy in segmenting white matter than gray matter (see Figure 5.2). A slightly higher discrepancy emerges between the true and estimated gray matter mean intensities, which also explains the relatively lower accuracy in classifying gray matter tissue.

Robustness to misregistration and atlas-free segmentation

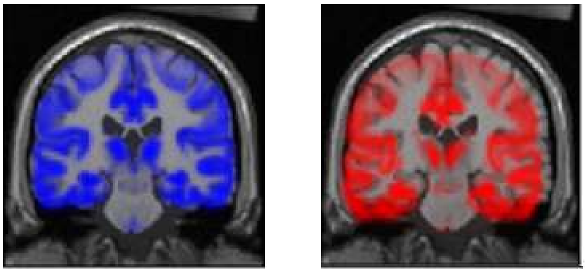
All atlas-based segmentation methods rely heavily on the accuracy in estimating the deformations mapping from the atlas to the individual volumes. Solving the segmentation and registration problems within a single modelling and computational framework has been widely accepted as a powerful and effective strategy in order to ensure the success of both processing tasks, additionally to being a theoretically principled approach (Ashburner and Friston, 2005; DAgostino et al., 2006; Pohl et al., 2006; Xiaohua et al., 2004b; Yezzi et al., 2001). Nonetheless, it is possible to encounter cases in which



(a)

(b)

Figure 5.3: Priors over the mean intensities of gray (a) and white (b) matter. The priors were learned from a synthetic data set consisting of 20 T1-weighted scans generated with the Brainweb MR simulator. The Gaussian curves show the estimated priors, while crosses represent the true (blue) and estimated (red) tissue means. The estimated means correspond to the modes of the posterior distributions computed by the proposed VB algorithm.



(a) Aligned template (b) Misaligned template

Figure 5.4: To test the robustness of the algorithm to misregistration, the tissue probability maps were deliberately shifted from their optimal positioning (a) by imposing a 7.5 mm translation in each direction, as illustrated in (b).

aligning the template to an individual scan turns out to be particularly difficult, due for example to a poor initialisation of the deformations or to the presence of anatomical features, for instance pathological ones, which the atlas does not capture. In such cases segmentation accuracy can be strongly affected by misregistration errors.

Introducing priors over the intensity distribution parameters is a convenient and reliable solution to cope with these difficulties. In fact, it can help to prevent implausible parameter estimates, whenever registration errors are misleading the model fitting process. To demonstrate this property, the synthetic data set consisting of twenty T1-weighted scans was split into a training and a test subset, of ten volumes each. The first ten images were processed by the proposed variational algorithm to learn empirical intensity priors, as explained in 5.6.1. Secondly, the remaining test images were segmented making use of these priors, while registration failure was simulated by imposing a 7.5 mm shift of the atlas from its optimal alignment configuration in each of the three Cartesian directions (Figure 5.4).

The accuracy of the resulting segmentations was finally assessed by computing Dice overlap coefficients. Results are illustrated in Figure 5.5. Here the performance of the presented method, used in combination with the empirical priors, is compared to that of the same algorithm with uninformative priors, as well as to that of a maximum likelihood method, as implemented in SPM12.

As to be expected the maximum likelihood method and the variational method with uninformative priors do not perform very differently, except for the fact that the

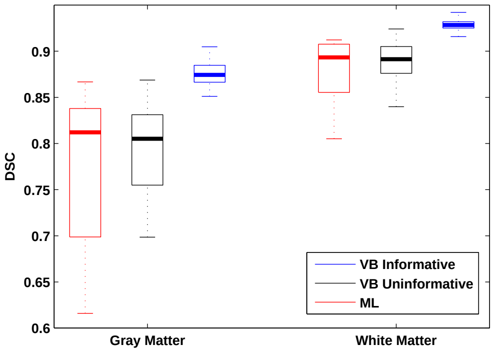


Figure 5.5: Accuracy of the presented variational algorithm obtained on synthetic data in the presence of registration errors. The performance of the VB algorithm with empirical informative priors (blue) is compared to that of the same algorithm with uninformative priors (red) and to the ML approach, as implemented in SPM12 (black).

ML algorithm shows higher variance of the results. On the contrary, when using the priors learned from the training data, the accuracy in segmenting gray and white matter increases significantly, yielding simultaneously lower variance of the overlap measures. Examples of gray matter segmentations obtained with the ML approach and with the VB method using informative priors are shown in Figure 5.6. These results confirms that variational Bayesian inference can augment the robustness of standard maximum likelihood algorithms, while providing a general and flexible computational framework, which could be applied to many real world problems, by learning appropriate priors from available training data.

As an additional proof of validity, an atlas free version of the presented algorithm was also implemented and tested on the same synthetic data. This purely intensity-based framework in not expected to achieve segmentation accuracy, or reliability, comparable to that of the full, atlas driven method. However, the fact that, even in the absence of tissue probability maps, fairly accurate segmentations can be obtained (see Figure 5.7), demonstrates again the soundness of the presented algorithm.

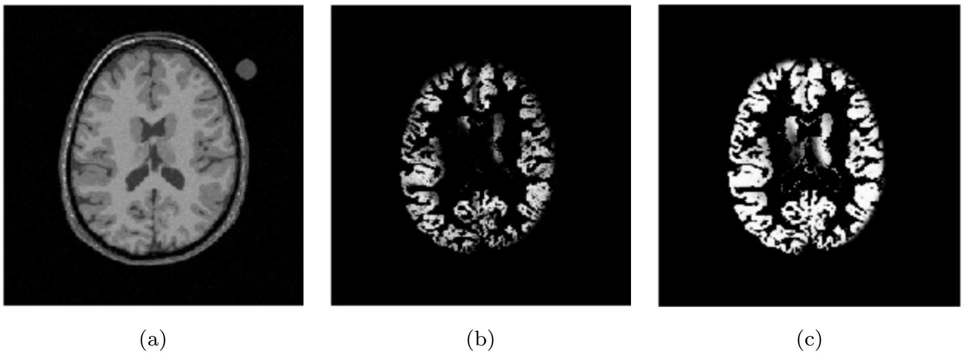


Figure 5.6: Example of gray matter segmentation obtained on a simulated T1-weighted scan (a) in the presence of misregistration between the data and the template, using a ML approach (b) and a VB approach with informative intensity priors (c).

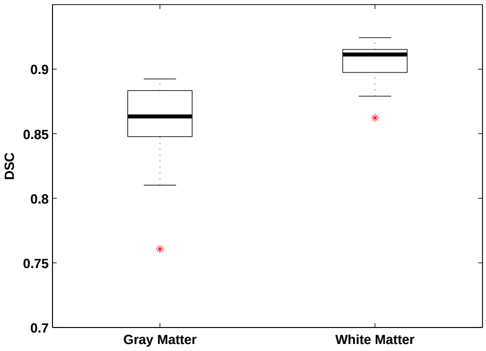


Figure 5.7: Dice similarity coefficients between the gray and white matter segmentations produced by the presented algorithm in an atlas free setting and the underlying ground truth, for twenty simulated T1-weighted scans.

5.6.2. Experiments on real data

The previous experiments, performed on simulated data, have demonstrated and quantified the accuracy of the presented method for segmenting brain tissues from MRI volumes. In fact, due to the availability of the underlying ground truth, working with synthetic data is especially convenient for the objective of testing new techniques and for the comparison of their performance to that of the methods that have become established as current state-of-the-art. Nonetheless, simulated data is intrinsically less complex than the data encoded in any real scan, from a biological point of view, as well as in terms of signal and noise properties. Therefore it is important to asses the behaviour of image processing tools also on real data.

For this reason, this section presents a series of experiments performed on real MRI data from two publicly available data sets: the OASIS (http://www.oasis-brains.org) and the IXI (http://brain-development.org/ixi-dataset) databases. Such experiments provide further evidence regarding the accuracy of the proposed method for segmenting brain tissues and illustrate some of its distinctive properties, which derive from adopting a variational inference scheme.

Assessing segmentation accuracy

The performance of the proposed segmentation algorithm was assessed on real data, making use of T1-weighted scans from the cross-sectional OASIS database (Marcus et al., 2007). In fact, manual labels, provided by Neuromorphometrics, Inc. (http://Neuromorphometrics.com) under academic subscription, are available for a small subset of this data set consisting of 35 subjects.

The data was processed by the presented segmentation algorithm, whose performance was compared to the SPM12 segmentation software. Figure 5.8 summarises the distributions of Dice coefficients for gray and white matter, which were obtained by comparing the manual labels with the segmentations produced by the proposed VB method using minimally informative priors and by SPM ML algorithm. For both tissue types, the presented variational approach yields a statistically significant increase in segmentation accuracy, compared to the maximum likelihood framework.

As to be expected, the Dice scores are generally lower, compared to the experi-

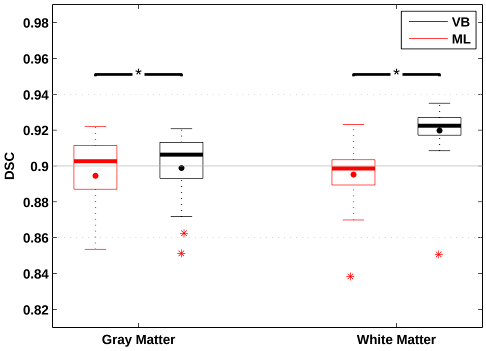


Figure 5.8: Dice scores computed between the manual labels provided by Neuromorphometrics for a subset of the OASIS data set and the gray and white matter segmentations obtained with the proposed VB method, using minimally informative priors, and with SPM ML algorithm.

ments performed on synthetic data. This is due to the more complex nature of real MRI signals. Additionally, the subset of the OASIS database that was used for this experiment comprises few scans of elderly subjects with severe atrophy and abnormal signal intensities, which explains the presence of negative outliers in the distribution of accuracy scores. Finally, when evaluated directly against manual labels, the accuracy attained by automated segmentation techniques depends quite heavily on the protocol adopted for manually annotating the data. For instance, in the data labelled by Neuromorphometrics, gray matter labels often include also a few CSF voxels.

Additional validation experiments were performed using data from the freely available IXI brain database (http://brain-development.org/ixi-dataset/), which, as opposed to the OASIS database, includes multiple modalities, in particular T1-, T2- and PD-weighted images of healthy adult subjects, acquired in three different sites, with different scanning systems. Ground truth segmentations are not available for such a data set. However, in this case, as opposed to the previous experiments, the aim is to illustrate some of the properties and advantages of the proposed method, rather than providing explicit accuracy measures.

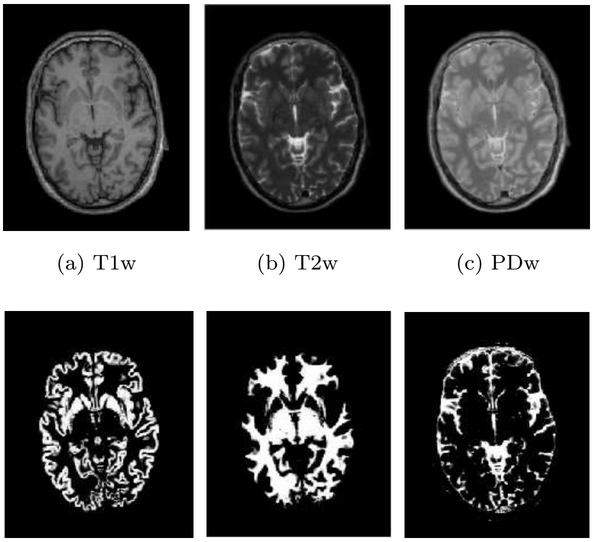
Determining model complexity

One of the most significant advantages of variational inference over maximum likelihood estimation is its intrinsic capability of containing the effects of overfitting (Attias, 1999; Bishop, 2006). In the case of mixture models this allows, for instance, determining the optimal number of components (K) without performing cross-validation, which is usually rather demanding for the amount of computation, as well as for the amount of data, that it requires (Bishop, 2006).

Indeed, the question of selecting model complexity has often been overlooked in the framework of medical image segmentation: throughout the literature, the most common way of handling the choice on the number of classes, is to manually tune K, based on visual inspection of the segmentations and/or intensity histograms. Clearly, this is too arbitrary and subjective for even being considered as a model selection strategy.

Instead, the proposed method implements an implicit automated relevance determination (ARD) scheme, where, if the number of Gaussians is set to a value that is higher than the optimal one, the redundant components will be automatically pruned out of the model (Corduneanu and Bishop, 2001; Tzikas et al., 2008), as their responsibilities $\{\gamma_{jk}\}_{j=1,\ldots,N}$ are quickly driven to zero by the algorithm. This follows from adopting a variational lower bound to approximate the marginal likelihood, which causes overly complex models, that is to say models with additional clusters that do not significantly help to explain the observed data, to be implicitly penalised (Attias, 1999). A similar behaviour is inherently impossible to reproduce within model fitting strategies that do not take into account estimation uncertainty, such as the maximum likelihood framework.

This property is illustrated making use of the scans of one subject included in the IXI database. In particular, the data depicted in figures 5.9a, 5.9b and 5.9c was processed by the presented VB algorithm, after having set five Gaussians for each of the tissue types of interest. At convergence, only two components survived for gray matter, one for white matter, three for CSF, two for bone and four for soft tissues, as shown in Figure 5.10. The plots reported in Figure 5.11 illustrate how the posterior densities over the mean intensity of white matter evolve during model learning and, in particular, how four irrelevant components are reverted to their prior distributions, which in this case are



(a) T1w (b) T2w (c) PDw

(d) GM (e) WM (f) CSF

Figure 5.9: Axial slices of T1-weighted (a), T2-weighted (b) and PD-weighted (c) scans and resulting gray matter (d), white matter (e) and cerebrospinal fluid (f) segmentations obtained with the variational algorithm described in this chapter.

uninformative. In a similar setting, ML or MAP algorithms would have simply found the best fit to the data, making use of all the available components, but the optimal number of Gaussians would have had to be determined *a priori*, through some form of model comparison.

Learning informative GMM priors via intensity normalisation

One of the difficulties of working with conventional (i.e. non-quantitative) MRI data is the lack of a standardised intensity scale (Nyúl et al., 2000). With respect to the work presented here, this makes it difficult to define, or learn, intensity priors that can effectively generalise to unseen data. Indeed, even for images of a single data set, comprising volumes acquired with the same scanner and protocol, the distribution of intensities across subjects might be poorly consistent.

Unsurprisingly, when trying to learn intensity priors using real MR data, one is directly confronted with the problem of normalising signal intensities. For instance, if fifty randomly selected T1-weighted scans from the IXI data set, acquired in the same

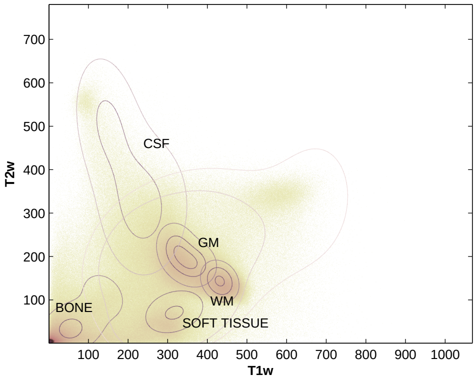
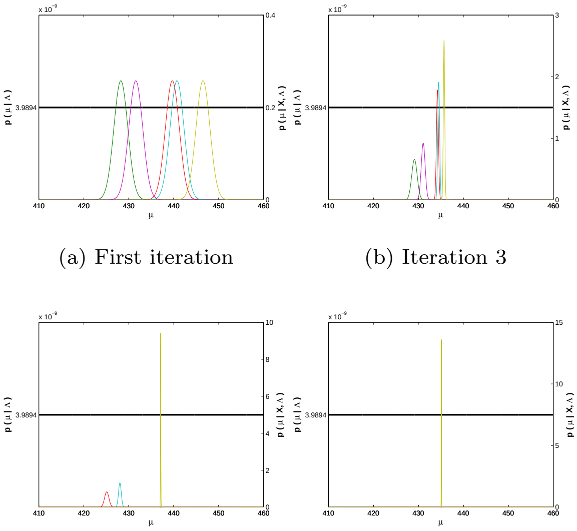


Figure 5.10: Contour plot of the intensity distributions of gray matter, white matter, cerebrospinal fluid, bone and soft tissue obtained for one subject included in the IXI data set, overlaid on the joint histogram of the T1- and T2-weighted images. The optimal number of components is determined automatically by the presented VB algorithm.



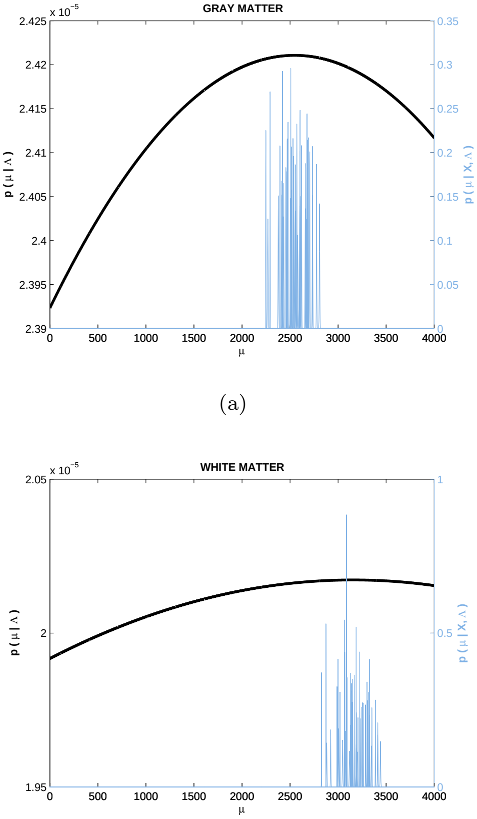
(a) First iteration

(b) Iteration 3

(c) Iteration 7

(d) Last iteration

Figure 5.11: Posterior densities over the mean intensity of white matter, at different iterations of the presented algorithm, showing non-relevant components being reverted to their prior distributions.



(a)

(b)

Figure 5.12: Collection of individual posteriors on the mean T1-weighted intensity of gray (a) and white (b) matter, obtained from 50 subjects included in the IXI database. Without performing any intensity normalisation the resulting empirical priors (black curves) are poorly informative.

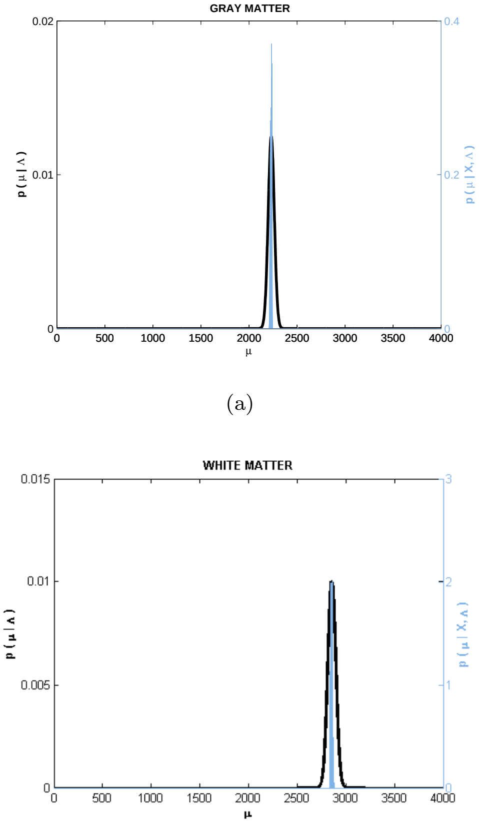
site and with the same scanner, are processed with the proposed variational algorithm to estimate intensity priors, as described in Section 5.5.5, the empirical priors turn out to be weakly informative, as they properly reflect the uncertainty due to the variability of the intensity scales (see Figure 5.12). The situation would be even worse if the MR volumes were acquired with different scanners or sequences (only quantitative imaging techniques would virtually be immune from such a problem).

Nonetheless, the generative model presented here can also be exploited to address the problems associated with the non-standardised nature of MRI signals. In fact, assuming that the parameter controlling the constant component of the bias field is not heavily penalised by the regularisation, the zeroeth order DCT basis function can serve to compensate for the variability in intensity scaling, as long as informative intensity priors are introduced in the model. Furthermore, here the effect of having an additional global scaling parameter is explored. Such a parameter can be optimised within the same learning scheme presented earlier in this chapter and, in particular, this can be formalised as a maximisation problem, where the aim is to maximise the following term ($\mathcal{L}_2$) contributing to the lower bound

$$\begin{aligned}\mathcal{L}_2 &= \iint q(\Theta_\mu, \Theta_\Sigma|\Theta_{gs}) \log \left\{ \frac{p(\Theta_\mu, \Theta_\Sigma)}{q(\Theta_\mu, \Theta_\Sigma|\Theta_{gs})} \right\} \mathrm{d}\Theta_\mu \mathrm{d}\Theta_\mu \\ &= -D_{KL}(q(\Theta_\mu, \Theta_\Sigma|\Theta_{gs}) \| p(\Theta_\mu, \Theta_\Sigma)) \ ,\end{aligned} \tag{5.46}$$

which corresponds to minimising the KL divergence between the intensity priors and the approximating posteriors. The problem can be solved using non-linear, gradient-based optimisation techniques, by computing the first and second derivatives of $\mathcal{L}_2(\Theta_{gs})$ with respect to the global scaling parameters Θ_{gs}. By iterating over updating the empirical priors and estimating the scaling factors for the individual scans, it is possible learn informative intensity priors, as illustrated in Figure 5.13, while automatically compensating for the inconsistency of MRI signal intensities.

Naturally, such a procedure requires accurate estimates of the intensity distribution, bias and deformation parameters for each individual, that is to say, the problems of learning priors and estimating individual posteriors are inherently related in a circular manner. As a result, for particularly critical data sets, e.g. pathological data, which often exhibit larger anatomical variability, the Bayesian framework described in this chapter might not be able to provide informative priors, due to the lack of a sufficient number of samples or to poor initial estimates of the model parameters. Nonetheless, in such cases, the presented computational framework, which represents a coherent generalisation of some state-of-the-art segmentation algorithms that rely on ML model fitting, could be applied with minimally informative intensity priors and yet it would outperform ML estimation, as indicated by the experiments presented in Section 5.6.2.



(a)

(b)

Figure 5.13: Collection of individual posteriors on the mean T1-weighted intensity of gray (a) and white (b) matter, after including a global rescaling parameter, which is optimised as part of the same generative modelling framework presented in the previous sections. The estimated priors (black curves) are now much more informative than the ones depicted in Figure 5.12.

5.7. Summary

This chapter has shown that variational Bayes represents a viable and effective framework for performing atlas-based medical image segmentation, in spite of not having been thoroughly exploited so far in such a field. In fact, the variational Gaussian mixture model presented in this chapter, which is an extension and a generalisation of the model adopted in Chapter 3, was tested on both synthetic and real MRI data to demonstrate, first of all, how the proposed framework can provide accurate segmentation results at an equivalent computational cost compared to ML or MAP implementations. In addition, some of advantages deriving from adopting a fully Bayesian formulation, such as the possibility of automatically determining optimal model complexity and quantifying the uncertainty of model parameter estimates, have been illustrated using neuroimaging data. Finally, an empirical Bayes learning scheme has been presented, which can serve to estimate informative intensity priors, thus ensuring even greater robustness of the proposed modelling approach, for instance in the presence of misregistration between individual data and the warped templates.

However, as opposed to the line defined both in Chapter 3 and Chapter 4, the approach adopted in this chapter consists in treating the tissue probability maps as fixed hyperparameters, rather than as random variables to be inferred from the observed data. This was a design choice, which was made so as to enable a direct comparison between ML and VB approaches, by validating the proposed variational algorithm against the widely used ML implementation distributed with the SPM software (Ashburner and Friston, 2005). In the next chapter, the VB scheme introduced here will be combined with the groupwise generative perspective of Chapter 3. Furthermore, a diffeomorphic modelling framework will be exploited, which is more suitable than the small deformation approach adopted in this chapter for the purpose of capturing modes of anatomical variability (Cootes et al., 2004; Fletcher et al., 2004).

6

Generative diffeomorphic atlas construction from brain and spinal cord MRI data

6.1. Introduction

This chapter will focus on the potential and on the challenges associated with the development of an integrated brain and spinal cord modelling framework for the processing of MR neuroimaging data.

The aim of the work presented here is to demonstrate how a hierarchical generative model of imaging data, which captures simultaneously the distribution of signal intensities and the variability of anatomical shapes across a large population of subjects, can serve to quantitatively investigate, *in vivo*, the morphology of the central nervous system (CNS). This can be achieved by processing simultaneously information related to the different compartments of the CNS, such as the brain and the spinal cord, without having to resort to organ specific solutions (e.g. tools optimised only for the brain, or only for the spine), which are inevitably harder to integrate.

6.2. Data model

Along the line already delineated in Chapter 3, the work presented here investigates the potential of a general and comprehensive modelling framework whose aim is to interpret large data sets of MRI scans from a Bayesian generative perspective. This is achieved by building on the modelling elements introduced in the previous chapters, which are further developed here and integrated in one single algorithmic framework. Specifically, the aim is to demonstrate the validity of such a generative approach for the purpose of performing simultaneous brain and spinal cord morphometric analyses from MRI data sets. In doing so, a strategy is outlined on how to overcome some of the limitations of most currently available image processing tools for neuroimaging, whose performance has been optimised on the brain at the expense of the spinal cord (indeed the spinal cord is frequently neglected *tout court* by such tools). For this reason, the imaging data that will be used for validation in this chapter, consist of a large set of multimodal head and neck MRI scans, acquired at different sites, with different imaging systems and scanning protocols.

Let us consider a population of M subjects belonging to a homogeneous group, from an anatomical point of view, and let us assume that D image volumes of different contrast are available for each subject.

As seen in the previous chapters, from a generative perspective, the image intensities $\overline{\mathbf{X}} = \{\mathbf{X}_i\}_{i=1,\dots,M}$, which constitute the observed data, can be thought of as being generated by sampling from D-dimensional Gaussian mixture probability distributions, after non-linear warping of a probabilistic anatomical atlas. Such an atlas carries *a priori* anatomical knowledge, in the form of average-shaped tissue probability maps, while from a mathematical modelling point of view, it encodes local (i.e. spatially varying) mixing proportions $\Theta_\pi = \{\boldsymbol{\pi}_j\}_{j=1,\dots,N}$ of the mixture model, with j being an index set over the N template voxels.

6.2.1. Tissue priors

Since each image voxel $j \in \{1, \dots, N_i\}$, for each subject $i \in \{1, \dots, M\}$ is considered as being drawn from K possible tissue classes, the following prior latent variable model defines the probability of finding tissue type k, at a specific location j (i.e. centre of voxel j), in image i, prior to observing the corresponding image intensity signal

$$p(z_{ijk} = 1|\Theta_\pi, \Theta_w, \Theta_u) = \frac{w_{ik}\, \pi_k(\boldsymbol{\xi}_i(\mathbf{y}_j))}{\sum_{c=1}^K w_{ic}\, \pi_c(\boldsymbol{\xi}_i(\mathbf{y}_j))} \ , \tag{6.1}$$

or equivalently

$$p(\mathbf{z}_{ij}|\Theta_\pi, \Theta_w, \Theta_u) = \prod_{k=1}^K \left(\frac{w_{ik}\, \pi_k(\boldsymbol{\xi}_i(\mathbf{y}_j))}{\sum_{c=1}^K w_{ic}\, \pi_c(\boldsymbol{\xi}_i(\mathbf{y}_j))} \right)^{z_{ijk}} \ . \tag{6.2}$$

Class memberships, for each subject and each voxel, are encoded in the latent variable $\mathbf{z}_{ij}$, which is a K-dimensional binary vector. $\{\pi_k\}_{k=1,\ldots,K}$ are scalar functions of space $\pi_k : \Omega_\pi \to \mathbb{R}$, common across the entire polulation, which satisfy the constrain

$$\sum_{k=1}^K \pi_k(\boldsymbol{y}) = 1 \ , \ \forall \boldsymbol{y} \in \Omega_\pi \subset \mathbb{R}^3 \ , \tag{6.3}$$

with $\boldsymbol{y}$ being a continuous coordinate vector field, as opposed to $\mathbf{y}_j$, which indicates discrete coordinates sampled at the centre of voxel j. Global weights $\Theta_w = \{\boldsymbol{w}_i\}_{i=1,\ldots,M}$ are introduced to further compensate for individual differences in tissue composition.

In equation (6.1), $\boldsymbol{\xi}_i$ denotes a generic spatial transformation, parametrised by Θ_u, which allows projecting prior information onto individual data, with $\boldsymbol{\xi}_i : \Omega_i \to \Omega_\pi$ being a continuous mapping from the domain $\Omega_i \subset \mathbb{R}^3$ of image i, into the space of the tissue priors $\Omega_\pi \subset \mathbb{R}^3$. Since digital image data for subject i is a discrete signal, defined on a tridimensional grid of N_i voxels, the mapping $\boldsymbol{\xi}_i$ needs to be discretised as well, on the same grid, by sampling it at the centre of each voxel $j = \{1, \ldots, N_i\}$, to give the discrete mapping $\{\boldsymbol{\xi}_i(\mathbf{y}_j)\}_{j=1,\ldots,N}$ that appears in (6.1).

As opposed to the modelling approach described in Chapter 5, where the tissue priors were considered as fixed and known *a priori* quantities, here the tissue probability maps are treated as random variables, whose point estimates or full posteriors can be inferred via model fitting, along the same line of Chapter 3.

For this purpose, a finite dimensional parametrisation needs to be defined. Typically, whenever a continuous function needs to be reconstructed from a finite sequence, it is possible to formulate the problem as an interpolation that makes use of a finite set of coefficients and continuous basis functions. Since the priors $\{\pi_k\}_{k=1,\ldots,K}$ are bounded to take values in the interval [0, 1] on the entire domain Ω_π (see equation (6.3)), not all basis functions are well suited here. Linear basis functions, besides being quite a computationally efficient choice, have the convenient property of preserving the values of $\{\pi_k\}_{k=1,\ldots,K}$ in the interval [0, 1], as long as the coefficients are also in the same interval. Such coefficients belong to the discrete set $\Theta_\pi = \{\boldsymbol{\pi}_j\}_{j=1,\ldots,N}$ of K-dimensional vectors, with

$$\sum_{k=1}^{K} \pi_{jk} = 1, \quad \forall j \in \{1, \ldots, N\} \; . \tag{6.4}$$

They can be learned directly from the data, as it will be shown in the following section.

Additionally, prior distributions on the parameters $\{\boldsymbol{\pi}_j\}_{j=1,\ldots,N}$ can be introduced (Bishop, 2006), which are particularly useful both to ensure computational stability, by preventing the logarithm of the tissue priors from diverging to infinity, and to obtain smoother templates around the edges of the field of view, where less observed data is available. Dirichlet priors are the most convenient choice here, since they are conjugate to multinomial forms of the type in (6.2)

$$p(\boldsymbol{\pi}_j) = \mathrm{Dir}(\boldsymbol{\pi}_j|\boldsymbol{\alpha}_0) = C(\boldsymbol{\alpha}_0) \prod_{k=1}^{K} \pi_{jk}^{\alpha_k - 1} \; , \tag{6.5}$$

where the normalising constant is given by

$$C(\boldsymbol{\alpha}_0) = \frac{\Gamma(\bar{\alpha})}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_k)} \; , \tag{6.6}$$

with $\Gamma(\cdot)$ being the gamma function and

$$\bar{\alpha} = \sum_{k=1}^{K} \alpha_k \; . \tag{6.7}$$

6.2.2. Diffeomorphic image registration

The generative interpretation of imaging data that this thesis relies on involves warping an average-shaped atlas to match a series of individual scans. Such a problem, that is to say template matching via non-rigid registration, has been largely explored in medical imaging, mainly for solving image segmentation or structural labelling problems, in an automated fashion (Ashburner and Friston, 2005; Bajcsy et al., 1983; Bowden et al., 1998; Christensen, 1999; Chui et al., 2001; Iglesias et al., 2012a; Joshi et al., 2004; Khan et al., 2008; Pluta et al., 2009; Shen and Davatzikos, 2004; Warfield et al., 1999).

Indeed, the modelling of spatial mappings between different anatomies can be approached in a variety of manners, depending on the adopted model of shape and on the objective function (i.e. similarity metric and regularisation) that the optimisation is based on, thus leading to a variety of algorithms with remarkably different properties (Denton et al., 1999; Klein et al., 2009; Penney et al., 1998).

The work presented in this chapter is formulated according to the Large Deformation Diffeomorphic Metric Mapping (LDDMM) framework (Younes, 2010), as opposed to the material included in the previous chapters, which is based on affine and small non-linear deformations.

In the so called small deformation setting, a mapping $\boldsymbol{\phi} : \Omega \rightarrow \Omega$ is defined as

$$\boldsymbol{\phi}(\mathbf{y}) = \boldsymbol{y} + \boldsymbol{u} \, , \forall \boldsymbol{y} \in \Omega \subset \mathbb{R}^3 \, , \tag{6.8}$$

where $\boldsymbol{u}$ is a displacement vector field, belonging to an adequate Hilbert space[1] $\mathcal{H}$ of smooth, compactly supported vector fields on Ω , equipped with a scalar product $\langle \, \cdot, \, \cdot \rangle_H$.

The inverse map $\boldsymbol{\phi}^{-1}$ is approximated by

$$\boldsymbol{\phi}^{-1}(\boldsymbol{y}) = \boldsymbol{y} - \boldsymbol{u} \, . \tag{6.12}$$

Such a first order (linear) approximation can be acceptable for small displacements $\boldsymbol{u}$, but as the norm $||\boldsymbol{u}||_H = \langle \, \boldsymbol{u} \, , \boldsymbol{u} \, \rangle_H^{1/2}$ grows larger, the invertibility of $\boldsymbol{\phi}$ is no longer guaranteed. For this reason, a more convenient way of parametrising large deformations $\boldsymbol{\phi}$ is by means of composing a series of sufficiently small deformations (ideally infinitesimally small) of the type in equation (6.8) (Trouvé, 1998).

In the LDDMM framework the transformations mapping between the source images

[1] A Hilbert space $\mathcal{H}$ is a complete inner product space, where an inner product is a map $\langle \cdot, \cdot \rangle$: $\mathcal{H} \times \mathcal{H} \rightarrow \mathbb{C}$, which associates each pair of vectors in the space with a scalar quantity. In particular given $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathcal{H}$ and $a, b \in \mathbb{C}$

$$\langle a\boldsymbol{x} + b\boldsymbol{y}, \boldsymbol{z} \rangle = a \langle \boldsymbol{x}, \boldsymbol{z} \rangle + b \langle \boldsymbol{y}, \boldsymbol{z} \rangle \, , \tag{6.9}$$

$$\langle \boldsymbol{x}, \boldsymbol{x} \rangle \geq 0, \text{ and } \langle \boldsymbol{x}, \boldsymbol{x} \rangle = 0 \Leftrightarrow \boldsymbol{x} = 0 \, , \tag{6.10}$$

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \langle \boldsymbol{y}, \boldsymbol{x} \rangle \, . \tag{6.11}$$

An inner product naturally induces a norm by $||\boldsymbol{x}|| = \langle \boldsymbol{x}, \boldsymbol{x} \rangle^{1/2}$, therefore every inner product space is also a normed vector space (Dieudonné, 2013).

and the target image are assumed to belong to a Riemannian manifold [2] of diffeomorphisms. A diffeomorphism $\boldsymbol{\phi} : \Omega \rightarrow \Omega$ is a smooth differentiable map (with a smooth differentiable inverse $\boldsymbol{\phi}^{-1}$) defined on a compact, simply connected domain $\Omega \subset \mathbb{R}^3$.

One way of constructing transformations belonging to the diffeomorphic group Diff(Ω) is to solve the following non-stationary transport equation (Joshi and Miller, 2000)

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\phi}(\boldsymbol{y}, t) = \boldsymbol{u}(\boldsymbol{\phi}(\boldsymbol{y}, t), t),\ \boldsymbol{\phi}(\boldsymbol{y}, 0) = \boldsymbol{y},\ t \in [0, 1]\ , \tag{6.13}$$

where $\boldsymbol{u}(\boldsymbol{\phi}(\boldsymbol{y}, t), t) \in \mathcal{H}$ is a time dependent, smooth velocity vector field, in the Hilbert space $\mathcal{H}$.

The initial map, at $t = 0$, is equal to the identity transform $\boldsymbol{\phi}(\boldsymbol{y}, 0) = \boldsymbol{y}$, while the final map, endpoint of the flow of the velocity field $\boldsymbol{u}$, can be computed by integration on the unitary time interval $t \in [0, 1]$ (Beg et al., 2005).

$$\boldsymbol{\phi}(\boldsymbol{y}, 1) = \int_0^1 \boldsymbol{u}(\boldsymbol{\phi}(\boldsymbol{y}, t))\mathrm{d}t + \boldsymbol{\phi}(\boldsymbol{y}, 0)\ . \tag{6.14}$$

Following from the theorems of existence and uniqueness of the solution of partial differential equations (*p.d.e.*), the solution of (6.13) is uniquely determined by the velocity field $\boldsymbol{u}(\boldsymbol{\phi}(\boldsymbol{y}, t))$ and by the initial condition $\boldsymbol{\phi}(\boldsymbol{y}, 0)$.

A diffeomorphic path $\boldsymbol{\phi}$ is not only differentiable, but also guaranteed to be a one-to-one mapping. Such qualities are highly desirable for finding morphological and functional correspondences between different anatomies without introducing tears or foldings, which would violate the conditions for topology preservation (Christensen, 1999). Additionally, the diffeomorphic framework provides metrics to quantitatively evaluate distances between anatomies or shapes. It should also be noted that diffeomorphisms are locally analogous to affine transformations (Avants et al., 2006).

In practice, finding an optimal diffeomorphic transformation to align a pair, or a group, of images involves optimising an objective function (e.g. minimising a cost function), in the space $\mathcal{H}$ of smooth velocity vector fields defined on the domain Ω. The

[2] A Riemannian manifold, in differential geometry, is a smooth manifold M equipped with a Riemannian metric (inner product). In particular, the Riemannian metric G_p on the n-dimensional manifold M^n defines, for every point $p \in M$, the scalar product of vectors in the tangent space T_pM, in such a way that given two vectors $\boldsymbol{x}, \boldsymbol{y} \in M$, the inner product $G_p(\boldsymbol{x}, \boldsymbol{y})$ depends smoothly on the point p. The tangent space represents the nearest approximation of the manifold by a vector space (Warner, 2013).

required smoothness is enforced by constructing the norm on the space $\mathcal{H}$ through a differential operator $\mathbf{L_u}$ (Beg et al., 2005), such that a quantitative measure of smoothness can be obtained via

$$\mathcal{R}(\mathbf{u}) = ||\mathbf{L_u}\mathbf{u}||^2_{L^2} \ , \tag{6.15}$$

where $\mathbf{u}$ is a discretised version of $\boldsymbol{u}$.

The form of the cost function will depend on how the observed data is modelled. For the work presented here, groupwise alignment is achieved via maximisation of the following variational objective function

$$\begin{aligned}\mathcal{E}(\Theta_u) =& \mathbb{E}_{\mathbf{Z}}[\log p(\ \overline{\mathbf{Z}} \ |\Theta_\pi, \Theta_w, \Theta_u)] + \log p(\Theta_u) + \text{const} \\ =& \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{K} \gamma_{ijk} \log\left(\frac{w_{ik}\pi_k(\boldsymbol{\phi}_i(\mathbf{y}_j))}{\sum_{c=1}^{K} w_{ic}\,\pi_c(\boldsymbol{\phi}_i(\mathbf{y}_j))}\right) - \frac{1}{2}\sum_{i=1}^{M} ||\mathbf{L_u}\mathbf{u}_i||^2_{L^2} + \text{const} \ ,\end{aligned} \tag{6.16}$$

where $\overline{\mathbf{Z}} \ = \{\mathbf{Z}_i\}_{i=1,\ldots,M}$ is the set of latent variables across the entire population, $\{\boldsymbol{\gamma}_{ij}\}_{i,j} = \{\mathbb{E}[\mathbf{z}_{ij}]\}_{i,j}$ are K-dimensional vectors of posterior belonging probabilities, Θ_π indicates the coefficients used to parametrise the tissue priors $\{\pi_k\}_{k=1,\ldots,K}$ and Θ_w denotes a set of individual tissue weights $\{\boldsymbol{w}_i\}_{i=1,\ldots,M}$ for rescaling the tissue probability maps. The coordinate mappings $\{\boldsymbol{\phi}_i\}_{i=1,\ldots,M}$ are encoded in the parameter set Θ_u, which consists of M vectors of coefficients $\{\mathbf{u}_i\}_{i=1,\ldots,M}$, containing $3 \times N_i$ elements each. Such coefficients can be used to construct continuous initial velocity fields via trilinear, or higher order, interpolation.

A procedure known as geodesic shooting (Allassonnière et al., 2005; Ashburner and Friston, 2011; Beg and Khan, 2006; Miller et al., 2006; Vialard et al., 2012) is applied, within the work presented here, to compute diffeomorphic deformation fields from corresponding initial velocity fields. Such a procedures exploits the principle of conservation of momentum (Younes et al., 2009), which is given by $\mathbf{m}_t = \mathbf{L}^\dagger_{\mathbf{u}}\mathbf{L_u}\mathbf{u}_t$, with $\mathbf{L}^\dagger_{\mathbf{u}}$ being the adjoint of the differential operator $\mathbf{L_u}$, to integrate the dynamical system governed by (6.13) without having to store an entire time series of velocity fields. The implementation adopted here relies on the work presented in Ashburner and Friston (2011).

The posterior membership probabilities $\{\boldsymbol{\gamma}_{ij}\}_{i,j}$ that appear in (6.16) can be computed by combining the prior latent variable model introduced in 6.2.1 with a likelihood model of image intensities, which will be described in subsection 6.2.4, thus leading to a fully unsupervised learning scheme.

Alternatively, when manual labels are available, binary posterior class probabilities can be derived directly from such categorical annotations, without performing inference from the observed image intensity data. In particular, if all input data has been manually labelled, then the resulting algorithm would implement a fully supervised learning strategy, while, if only some of the data has associated training labels, a hybrid approach can be adopted, which would fall into the category of semisupervised learning, as discussed in Chapter 4.

6.2.3. Combining diffeomorphic with affine registration

Anatomical shapes are very high dimensional objects. The diffeomorphic model described in the previous subsection can account for a significant amount of shape variability in the observed data.

Nevertheless, it is still convenient, mainly for computational reasons, to combine such a local, high dimensional shape model with global, lower dimensional transformations, such as rigid body or affine transforms. In fact, by beginning to solve the registration problem from the coarsest deformation components (e.g. rigid body or affine), it is possible to ensure that the subsequent diffeomorphic registration starts from a good initial estimate of image alignment, that is to say closer to the desired global optimum.

This makes the optimisation problem faster to solve and at the same time it reduces significantly the rate of registration failure (Modersitzki, 2004). Indeed, it is relatively common for non-linear registration algorithms to fail in the presence of a large translational or size mismatch between the reference and the target images (Jenkinson and Smith, 2001).

A possible parametrisation that combines affine and diffeomorphic transformations is

$$\boldsymbol{\xi}_i(\boldsymbol{y}) = \mathbf{T}_i\, \boldsymbol{\phi}_i(\boldsymbol{y}) + \mathbf{t}_i, \quad \forall \boldsymbol{y} \in \Omega_i \ , \tag{6.17}$$

where $\boldsymbol{\xi}_i(\boldsymbol{y})$ is the resulting mapping from image of subject i into the template space. Such a mapping is obtained by affine transforming the diffeomorphic deformation field

$\boldsymbol{\phi}_i$. The transformation matrix $\mathbf{T}_i$ encodes nine degrees of freedom (rotation, zooming and shearing) and, like in Chapter 3, is computed via an exponential map $\mathbf{T}_i = \exp(\mathbf{Q}_i(\mathbf{a}_i))$ with $\mathbf{Q}_i(\mathbf{a}_i) \in \mathfrak{ga}(3)$, where $\mathfrak{ga}(3)$ is the Lie algebra for the affine group in three dimension $GA(3)$ and $\mathbf{a}_i$ is a vector of nine parameters. Translations are modelled by the vector $\mathbf{t}_i \in \mathbb{R}^3$. The entire set of affine parameters is denoted as $\Theta_a = \{\mathbf{a}_i, \mathbf{t}_i\}_{i=1,\ldots,M}$.

6.2.4. Intensity model

From a general probabilistic perspective, classification of tissue types based on MR signal intensities requires a model of the observed data that is capable of capturing the probability of occurrence of each signal sample value $\mathbf{x}_{ij}$, provided that the true labels are known. In other words, the problem breaks down into defining suitable conditional probabilities $p(\mathbf{x}_{ij}|z_{ijk} = 1)$, for each $k = \{1, \ldots, K\}$ and then applying Bayes' rule to infer the posterior class probabilities.

The model adopted here is the same employed throughout this thesis, where image intensity distributions are represented as Gaussian mixtures. As in Chapter 5, the unknown mean $\boldsymbol{\mu}_{ik}$ and covariance matrix $\boldsymbol{\Sigma}_{ik}$ of each Gaussian component k, for subject i, are governed by Gaussian-Wishart priors.

Correction of intensity inhomogeneities is also performed within the same modelling framework and it involves multiplying the uncorrected intensities of each image volume by a bias field, which is modelled as the exponential of a weighted sum of discrete cosine transform basis functions. Such an approach is conceptually equivalent to scaling the probability distributions of all Gaussian components by a local scale parameter, which is the bias itself, such that

$$\begin{aligned} p(\mathbf{x}_{ij}|z_{ijk} = 1, \boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik}, \Theta_\beta) &= \det(\operatorname{diag}(\mathbf{b}_{ij}))\, \mathcal{N}(\operatorname{diag}(\mathbf{b}_{ij})\, \mathbf{x}_{ij}|\boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik}) && (6.18) \\ &= \mathcal{N}(\mathbf{x}_{ij}|\hat{\boldsymbol{\mu}}_{ik}, \hat{\boldsymbol{\Sigma}}_{ik}) \ , && (6.19) \end{aligned}$$

with

$$\begin{aligned} \hat{\boldsymbol{\mu}}_{ik} &= (\operatorname{diag}(\mathbf{b}_{ij}))^{-1}\, \boldsymbol{\mu}_{ik} \ , \\ \hat{\boldsymbol{\Sigma}}_{ik} &= (\operatorname{diag}(\mathbf{b}_{ij}))^{-1}\, \boldsymbol{\Sigma}_{ik}\, (\operatorname{diag}(\mathbf{b}_{ij}))^{-1} \ , \end{aligned} \qquad (6.20)$$

where Θ_β denotes the set of bias field parameters and $\mathbf{b}_{ij}$ is a D-dimensional vector representing the bias for subject i at voxel j.

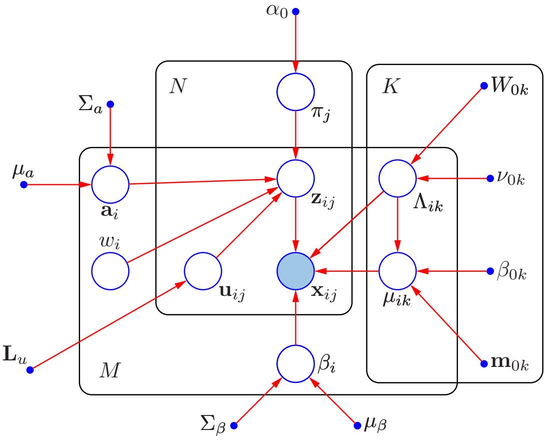


Figure 6.1: Graphical representation of the model adopted for the work presented in this chapter. Observed variables $\{\mathbf{x}_{ij}\}$ are represented by a filled circle. Latent variables $\{\mathbf{z}_{ij}\}$ as well as model parameters are depicted as unfilled circles. Blue solid dots correspond to hyperparameters. The so called plate notation is adopted to indicated repeated variables. Symbols referring to all variables and parameters are listed in table 6.1.

6.2.5. Graphical model

A graphical representation of the model adopted in this chapter is depicted in Figure 6.1, while a legend of the symbols used to indicate the different variables can be found in table 6.1.

Given such a model, it is possible to define the following variational objective function $\mathcal{L}$, which constitutes a lower bound on the logarithm of the marginal joint probability $p(\ \overline{\mathbf{X}}\ , \Theta_\beta, \Theta_a, \Theta_u, \Theta_\pi | \Theta_w)$, such that

$$\log p(\ \overline{\mathbf{X}}\ , \Theta_\beta, \Theta_a, \Theta_u, \Theta_\pi | \Theta_w) \geq \mathcal{L} \tag{6.21}$$

Symbol	**Meaning**
$\mathbf{x}_{ij}$	Observed image intensity at voxel j of image i.
$\mathbf{z}_{ij}$	Vector of latent class membership probabilities.
$\boldsymbol{\pi}_j$	Tissue priors at voxel j.
$\boldsymbol{\mu}_{ik}$	Mean intensity of class k for subject i.
$\mathbf{\Sigma}_{ik}$	Covariance of intensities for class k and subject i.
$\boldsymbol{W}_{0k}$	Scale matrix of Wishart prior distribution on $\mathbf{\Lambda}_k = (\mathbf{\Sigma}_k)^{-1}$.
ν_{0k}	Degrees of freedom of Wishart prior distribution on $\mathbf{\Lambda}_k$.
$\boldsymbol{m}_{0k}$	Mean of Gaussian prior distribution over $\boldsymbol{\mu}_k$
β_{0k}	Scaling hyperparameter of Gaussian prior distribution over $\boldsymbol{\mu}_k$
α_0	Hyperparameter governing the Dirichlet prior on $\boldsymbol{\pi}$.
Θ_β	Bias field parameters.
$\boldsymbol{\mu}_\beta$	Prior mean of bias parameters.
$\mathbf{\Sigma}_\beta$	Prior covariance matrix of bias parameters.
Θ_a	Affine transformation parameters.
$\boldsymbol{\mu}_a$	Prior mean of affine transformation parameters.
$\mathbf{\Sigma}_a$	Prior covariance matrix of affine transformation parameters.
$\boldsymbol{w}_i$	Weights for rescaling the tissue priors.
$\mathbf{u}_{ij}$	Initial velocity at voxel j for subject i.
$\boldsymbol{L}_u$	Differential operator to compute penalty on $\mathbf{u}_i$.
N	Number of image voxels.
K	Number of Gaussian mixture components.
M	Number of subjects.

Table 6.1: List of mathematical symbols used in this chapter.

and

$$
\begin{aligned}
\mathcal{L} =& \sum_{\mathbf{Z}} \iint \; q(\;\overline{\mathbf{Z}}\;, \Theta_\mu, \Theta_\Sigma) \log \left\{ \frac{p(\;\overline{\mathbf{X}}\;,\;\overline{\mathbf{Z}}\;, \Theta_\mu, \Theta_\Sigma, \Theta_\pi, \Theta_\beta, \Theta_a, \Theta_u | \Theta_w)}{q(\;\overline{\mathbf{Z}}\;, \Theta_\mu, \Theta_\Sigma)} \right\} \mathrm{d}\Theta_\mu \mathrm{d}\Theta_\Sigma \\
=& \mathbb{E}_{\mathbf{Z}, \Theta_\mu, \Theta_\Sigma} [\log p(\;\overline{\mathbf{X}}\;|\;\overline{\mathbf{Z}}\;, \Theta_\mu, \Theta_\Sigma, \Theta_\beta)] + \mathbb{E}_{\mathbf{Z}} [\log p(\;\overline{\mathbf{Z}}\;| \Theta_\pi, \Theta_w, \Theta_u, \Theta_a)] \\
&+ \mathbb{E}_{\Theta_\mu, \Theta_\Sigma} [\log p(\Theta_\mu, \Theta_\Sigma)] + \log p(\Theta_\pi) + \log p(\Theta_\beta) + \log p(\Theta_a) + \log p(\Theta_u) \\
&- \mathbb{E}_{\mathbf{Z}} [\log q(\;\overline{\mathbf{Z}}\;)] - \mathbb{E}_{\Theta_\mu, \Theta_\Sigma} [\log q(\Theta_\mu, \Theta_\Sigma)] \; ,
\end{aligned}
\tag{6.22}
$$

where the expectations indicated as $\mathbb{E}_{\mathbf{Z}}$ and $\mathbb{E}_{\Theta_\mu, \Theta_\Sigma}$ are computed with respect to variational posterior distributions $q(\cdot)$ on the latent variables $\overline{\mathbf{Z}}$ and on the Gaussian means and covariances $\{\Theta_\mu, \Theta_\Sigma\}$, respectively. Optimisation of $\mathcal{L}$, which provides optimal parameter and hyperparameter estimates, will be discussed in the following section.

6.3. Model fitting

The model described in the previous section can be fit to data sets of MR images by combining a variational expectation-maximisation (VBEM) algorithm with gradient-based numerical optimisation techniques.

Indeed, the VBEM algorithm described in Chapter 5 is well-suited for solving the problem discussed here since it allows estimating variational posterior distributions on the Gaussian mixture parameters, under the assumption that $q(\;\overline{\mathbf{Z}}\;, \Theta_\mu, \Theta_\Sigma)$ factorizes as $q(\;\overline{\mathbf{Z}}\;)q(\Theta_\mu, \Theta_\Sigma)$ (Bishop, 2006).

Optimisation of the bias field parameters Θ_β can be performed via non-linear numerical techniques. Here the problem is solved using the Gauss-Newton method (Bertsekas, 1999), so as to maximise the objective function in (6.22) with respect to Θ_β. The resulting implementation is very similar to the one described in Chapter 3 therefore further details are omitted here.

The following subsections instead will provide a more detailed description of the algorithmic scheme and the relative computations useful for learning the average-shaped tissue templates $\Theta_\pi = \{\boldsymbol{\pi}_j\}_{j=1,\ldots,N}$ and for estimating the set of initial velocity fields $\Theta_u = \{\mathbf{u}_i\}_{i=1,\ldots,M}$, as well as the set of affine parameters $\Theta_a = \{\mathbf{a}_i\}_{i=1,\ldots,M}$, for the entire population.

6.3.1. Updating the tissue priors

At each iteration of the algorithm the tissue priors $\Theta_\pi = \{\boldsymbol{\pi}_j\}_{j=1,\ldots,N}$ need to be updated, given the current estimates of all the other parameters, which are kept fixed for each individual in the population.

Considering only the terms in (6.22) that depend on Θ_π gives the following objective function, which has to be maximised with respect to Θ_π

$$\begin{aligned}\mathcal{L}_\pi &= \mathbb{E}_{\mathbf{Z}}[\log p(\ \overline{\mathbf{Z}}\ |\Theta_\pi, \Theta_w, \Theta_u, \Theta_a)] + \log p(\Theta_\pi) + \text{const} \\ &= \sum_{i=1}^{M} \int\limits_{\boldsymbol{y} \in \Omega_i} \sum_{k=1}^{K} \gamma_{ik}(\boldsymbol{y}) \log \left(\frac{w_{ik} \pi_k(\boldsymbol{\xi}_i(\boldsymbol{y}))}{\sum_{c=1}^{K} w_{ic}\, \pi_c(\boldsymbol{\xi}_i(\boldsymbol{y}))} \right) \mathrm{d}\boldsymbol{y} + \log p(\Theta_\pi) + \text{const} \ .\end{aligned} \tag{6.23}$$

It should be noted that the parameters Θ_π that need to be estimated are defined on the domain of the template Ω_π, rather than on the individual spaces $\{\Omega_i\}_{i=1,\ldots,M}$. For this reason equation (6.23), which is a sum of integrals on the native domains, needs to be mapped to Ω_π, by inverting the warps $\{\boldsymbol{\xi}_i\}_{i=1,\ldots,M}$, to give

$$\mathcal{L'}_\pi = \sum_{i=1}^{M} \int\limits_{\boldsymbol{y} \in \Omega_\pi} \sum_{k=1}^{K} \det\left(\frac{\partial \boldsymbol{\xi}_i^{-1}}{\partial \boldsymbol{y}} \right) \gamma_{ik}(\boldsymbol{\xi}_i^{-1}(\boldsymbol{y})) \log \left(\frac{w_{ik} \pi_k(\boldsymbol{y})}{\sum_{c=1}^{K} w_{ic}\, \pi_c(\boldsymbol{y})} \right) \mathrm{d}\boldsymbol{y} + \log p(\Theta_\pi) + \text{const} \ , \tag{6.24}$$

where the determinants of the Jacobian matrices of the deformations are included to preserve volumes after the change of variables.

Finally equation (6.24) needs to be discretised on a regular voxel grid, whose centres have coordinates $\{\mathbf{y}_j\}_{j=1,\ldots,N}$, to give

$$\mathcal{L'}_\pi = \sum_{i=1}^{M} \sum_{j=1}^{N} \sum_{k=1}^{K} \det(\mathbf{J}_{ij}^{\boldsymbol{\xi}^{-1}})\, \gamma_{ik}(\boldsymbol{\xi}_{ij}^{-1})\ \log \left(\frac{w_{ik} \pi_{jk}}{\sum_{c=1}^{K} w_{ic}\, \pi_{jc}} \right) + \log p(\Theta_\pi) + \text{const} \ , \tag{6.25}$$

where $\mathbf{J}_{ij}^{\boldsymbol{\xi}^{-1}}$ is obtained by sampling from the corresponding continuous Jacobian determinant field

$$\boldsymbol{\xi}_{ij}^{-1} = \boldsymbol{\xi}_i^{-1}(\boldsymbol{y})|_{\boldsymbol{y}=\mathbf{y}_j} \ , \tag{6.26}$$

$$\det(\mathbf{J}_{ij}^{\boldsymbol{\xi}^{-1}}) = \det\left(\frac{\partial \boldsymbol{\xi}_i^{-1}(\boldsymbol{y})}{\partial \boldsymbol{y}} \right)\Bigg|_{\boldsymbol{y}=\mathbf{y}_j} \ , \tag{6.27}$$

$$\pi_{jk} = \pi_k(\boldsymbol{y})|_{\boldsymbol{y}=\mathbf{y}_j} \ . \tag{6.28}$$

The prior term $p(\Theta_\pi)$ is given by the following Dirichlet distribution

$$p(\Theta_\pi) = \prod_{j=1}^{N} \mathrm{Dir}(\boldsymbol{\pi}_j|\boldsymbol{\alpha}_0) = C(\boldsymbol{\alpha}_0) \prod_{j=1}^{N} \prod_{k=1}^{K} \pi_{jk}^{\alpha_{0k}-1} \ . \tag{6.29}$$

Maximising equation (6.25) is a constrained optimisation problem, subject to

$$\sum_{k=1}^{K} \pi_{jk} = 1 \ , \ \forall j \in \{1, \ldots, N\} \tag{6.30}$$

A closed form solution could be easily found if the rescaling weights w were all equal to one. In such a case

$$\mathcal{L}'_\pi = \sum_{i=1}^{M} \sum_{j=1}^{N} \sum_{k=1}^{K} \det(\boldsymbol{J}_{ij})\, \gamma_{ik}(\boldsymbol{\xi}_{ij}^{-1})\, \log{(\pi_{jk})} + \sum_{j=1}^{N} \sum_{k=1}^{K} (\alpha_{0k} - 1) \log p(\pi_{jk}) + \mathrm{const} \ , \tag{6.31}$$

which could be maximised under the constraint (6.30) making use of Lagrange multipliers (Falk, 1967), to give

$$\pi_{jk} = \frac{N_{jk} + \alpha_{0k} - 1}{\sum_{k=1}^{K} (N_{jk} + \alpha_{0k}) - K} \ , \tag{6.32}$$

with $N_{jk} = \sum_{i=1}^{M} \det(\boldsymbol{J}_{ij})\, \gamma_{ik}(\boldsymbol{\xi}_{ij}^{-1})$.

This solution would provide maximum a posteriori point estimates of $\Theta_\pi = \{\boldsymbol{\pi}_j\}_{j=1,\ldots,N}$. However for this problem, it would also be possible to derive a full variational posterior distribution, which, like its prior, would take a Dirichlet form, with parameters $\boldsymbol{\alpha}_j = \boldsymbol{\alpha}_{\mathbf{0}} + \boldsymbol{N}_j$.

Unfortunately, when rescaling of the tissue priors by $\{\boldsymbol{w}_i\}_{i=1,\ldots,M}$ is allowed the optimisation problem becomes more complex. The strategy adopted here consists in finding an approximate solution to the unconstrained optimisation problem by setting the derivatives of the objective function in (6.25) to zero

$$\frac{\partial \mathcal{L}_\pi'}{\partial \pi_{jk}} = \sum_{i=1}^{M} \left(\det(\mathbf{J}_{ij}^{\boldsymbol{\xi}^{-1}})\, \gamma_{ik}(\boldsymbol{\xi}_{ij}^{-1}) \left(\frac{1}{\pi_{jk}} - \frac{w_{ik}}{\sum_{c=1}^{K} w_{ic} \pi_{jc}} \right) \right) + \frac{\alpha_{0k} - 1}{\pi_{jk}} = 0 \ . \tag{6.33}$$

Solving with respect to π_{jk}, under the simplifying assumption that the term $\sum_{c=1}^{K} w_{ic}\pi_{jc}$ can be treated as a constant, which is valid if the weights are sufficiently close to one, gives

$$\bar{\pi}_{jk} = \frac{N_{jk} + \alpha_{0k} - 1}{\sum_{i=1}^{M} \frac{\det(\mathbf{J}_{ij}^{\boldsymbol{\xi}^{-1}})\, \gamma_{ik}(\boldsymbol{\phi}_{ij}^{-1}) w_{ik}}{\sum_{c=1}^{K} w_{ic}\pi_{jc}}} \ . \tag{6.34}$$

Such a solution is then projected onto the constraining hyperplane, by preserving the same tissue proportions at each voxel

$$\pi_{jk} = \frac{\bar{\pi}_{jk}}{\sum_{c=1}^{K} \bar{\pi}_{jc}} \ . \tag{6.35}$$

Experimental testing of this strategy indicated that it gave a constant improvement of the objective function at a relatively cheap computational cost. Alternatively, numerical constrained optimisation techniques (Powell, 1978) could have been exploited to solve the template update problem, at the expenses of a slightly longer processing time.

6.3.2. Computing the deformation fields

Groupwise image alignment is achieved by optimisation of the variational objective function defined in (6.22), with respect to the parameters used to compute the deformations. This is equivalent to adopting the following image matching, or similarity, term

$$\begin{aligned} \mathcal{D} &= \mathbb{E}_{\mathbf{Z}}[\log p(\ \overline{\mathbf{Z}} \ | \Theta_\pi, \Theta_w, \Theta_u, \Theta_a)] \\ &= \sum_{i=1}^{M} \int_{\boldsymbol{y} \in \Omega_i} \sum_{k=1}^{K} \gamma_{ik}(\boldsymbol{y}) \log \left(\frac{w_{ik} \pi_k(\boldsymbol{\xi}_i(\boldsymbol{y}))}{\sum_{c=1}^{K} w_{ic} \, \pi_c(\boldsymbol{\xi}_i(\boldsymbol{y}))} \right) \mathrm{d}\boldsymbol{y} \ , \end{aligned} \tag{6.36}$$

which, working on discretised image grids, becomes

$$\mathcal{D} = \sum_{i=1}^{M} \sum_{j=1}^{N_i} \sum_{k=1}^{K} \gamma_{ijk} \log \frac{w_{ik} \pi_k(\boldsymbol{\xi}_{ij})}{\sum_{c=1}^{K} w_{ic} \pi_c(\boldsymbol{\xi}_{ij})} = \sum_{i=1}^{M} \sum_{j=1}^{N_i} \sum_{k=1}^{K} \gamma_{ijk} \log \frac{w_{ik} \pi'_{jk}}{\sum_{c=1}^{K} w_{ic} \pi'_{jc}} \ , \tag{6.37}$$

with

$$\boldsymbol{\xi}_{ij} = \boldsymbol{\xi}_i(\boldsymbol{y})|_{\boldsymbol{y}=\mathbf{y}_{ij}} \ , \tag{6.38}$$

$$\pi'_{jk} = \pi_k(\boldsymbol{\xi}_i(\boldsymbol{y}))|_{\boldsymbol{y}=\mathbf{y}_{ij}} \ . \tag{6.39}$$

The penalty term for this groupwise image registration problem is instead given by

$$\mathcal{R} = \mathcal{R}_{dif} + \mathcal{R}_{af} = \log p(\Theta_u) + \log p(\Theta_a) = -\frac{1}{2} \sum_{i=1}^{M} \left(||\mathbf{L}_{\mathbf{u}} \mathbf{u}_i||_{L^2}^2 + \mathbf{a}_i^T \boldsymbol{\Sigma}_{\boldsymbol{a}}^{-1} \mathbf{a}_i \right) + \text{const} \ , \tag{6.40}$$

with $\mathbf{u}_i$ being a $3 \times N_i$ dimensional vector of parameters used for representing the initial velocity field of image i and $\mathbf{a}_i$ encoding twelve affine deformation parameters used to compute the transformation in (6.17).

Updating the initial velocities

For each image i in the data set, updating the corresponding initial velocity field, given the current estimates of the templates, involves optimising the following objective function

$$\mathcal{E}_{dif}^{(i)} = \mathcal{D}^{(i)} + \mathcal{R}_{dif}^{(i)} = \sum_{j=1}^{N_i} \sum_{k=1}^{K} \gamma_{ijk} \log \frac{w_{ik} \pi_k(\boldsymbol{\xi}_{ij})}{\sum_{c=1}^{K} w_{ic} \pi_c(\boldsymbol{\xi}_{ij})} - \frac{1}{2} ||\mathbf{L}_{\mathbf{u}} \mathbf{u}_i||_{L^2}^2 \ , \tag{6.41}$$

with respect to $\mathbf{u}_i$, under the following deformation model

$$\boldsymbol{\xi}_{ij} = \boldsymbol{\xi}_i(\mathbf{y}_{ij}) = \mathbf{T}_i \, \boldsymbol{\phi}_i(\mathbf{y}_{ij}) + \mathbf{t}_i \ , \tag{6.42}$$

where $\boldsymbol{\phi}_i$ is a diffeomorphism computed via geodesic shooting (Ashburner and Friston, 2011) from the corresponding initial velocity field $\mathbf{u}_i$.

Here image registration is solved via Gauss-Newton optimisation, which requires computing both the first and second derivatives of the objective function (Hernandez and Olmos, 2008). A line search scheme is used to determine the optimal step size, which turned out to ensure faster convergence compared to the Levenberg-Marquardt scheme adopted in Chapter 3. This leads to a very high dimensional inverse problem, which unfortunately cannot be solved via numerical matrix inversion, since this would be prohibitively expensive from a computational point of view. The approach adopted in this work consists in treating this optimisation as a partial differential equation problem, which can efficiently be solved using multigrid methods (Modersitzki, 2004). The same full multigrid implementation as in Ashburner (2007) is adopted.

In particular, the gradient of the matching term $\mathcal{D}$ with respect to $\mathbf{u}_i$ is given by

$$\begin{aligned} \frac{\partial \mathcal{D}^{(i)}}{\partial \mathbf{u}_i} &= \sum_{k=1}^{K} \gamma_{ijk} \frac{\partial}{\partial \mathbf{u}_i} \left(\log \frac{w_{ik} \, \pi_k(\boldsymbol{\xi}_i)}{\sum_{c=1}^{K} w_{ic} \, \pi_c(\boldsymbol{\xi}_i)} \right) \\ &= \sum_{k=1}^{K} \gamma_{ijk} \left(\boldsymbol{g}_k^{\pi} - \sum_{c=1}^{K} \frac{w_{ic} \, \pi_c(\boldsymbol{\xi}_i)}{\sum_{c=1}^{K} w_{ic} \, \pi_c(\boldsymbol{\xi}_i)} \boldsymbol{g}_c^{\pi} \right) \ , \end{aligned} \tag{6.43}$$

which, making use of $\sum_{k=1}^{K} \gamma_{ijk} = 1$, can be rewritten as

$$\frac{\partial \mathcal{D}^{(i)}}{\partial \mathbf{u}_i} = \sum_{k=1}^{K} \left(\gamma_{ik} - \frac{w_{ik} \, \pi_k(\boldsymbol{\xi}_i)}{\sum_{c=1}^{K} w_{ic} \, \pi_c(\boldsymbol{\xi}_i)} \right) \boldsymbol{g}_k^{\pi} \ , \tag{6.44}$$

where $\boldsymbol{g}_k^{\pi}$ is computed, at each voxel j, by

$$\boldsymbol{g}_{jk}^{\pi} = \left(\mathbf{T}_i, \mathbf{J}_{ij}^{\boldsymbol{\xi}} \right)^T \nabla \left[\log \left(\pi_k(\boldsymbol{\xi}_{ij}) \right) \right] \ , \tag{6.45}$$

and $\mathbf{J}_i^{\boldsymbol{\xi}}$ indicates the Jacobian matrix of $\boldsymbol{\xi}_{ij}$.

An approximate positive semidefinite Hessian of $\mathcal{D}$ can instead be computed by discarding the second derivatives of the logarithm of tissue priors

$$\frac{\partial^2}{\partial \boldsymbol{y}^2} \log \left(\frac{w_{ik}\left(\pi_k(\boldsymbol{\xi}_i(\boldsymbol{y}))\right)}{\sum_{c=1}^{K} w_{ic}\left(\pi_c(\boldsymbol{\xi}_i(\boldsymbol{y}))\right)} \right) = 0 \text{ , } \forall \boldsymbol{y} \in \Omega_i \tag{6.46}$$

to give

$$\begin{aligned} \frac{\partial^2 \mathcal{D}^{(i)}}{\partial {\mathbf{u}_i}^2} = & \left(\sum_{k=1}^{K} \frac{w_{ik}\, \pi_k(\boldsymbol{\xi}_i)}{\sum_{c=1}^{K} w_{ic}\, \pi_c(\boldsymbol{\xi}_i)} \boldsymbol{g}_k^{\pi} \right) \left(\sum_{k=1}^{K} \frac{w_{ik}\left(\pi_k(\boldsymbol{\xi}_i)\right)}{\sum_{c=1}^{K} w_{ic}\left(\pi_k(\boldsymbol{\xi}_i)\right)} \boldsymbol{g}_k^{\pi} \right)^T \\ & - \sum_{k=1}^{K} \frac{w_{ik}\, \pi_k(\boldsymbol{\xi}_i)}{\sum_{c=1}^{K} w_{ic}\, \pi_c(\boldsymbol{\xi}_i)} \boldsymbol{g}_k^{\pi} \left(\boldsymbol{g}_k^{\pi}\right)^T \ . \end{aligned} \tag{6.47}$$

This ensures that each Gauss-Newton step is taken in the correct direction.

The first and second derivatives of the penalty term $\mathcal{R}$ are also required to solve this optimisation problem

$$\frac{\partial \mathcal{R}_{dif}^{(i)}}{\partial \mathbf{u}_i} = -{\mathbf{L}_{\mathbf{u}}}^{\dagger} \mathbf{L}_{\mathbf{u}} \mathbf{u}_i \ , \tag{6.48}$$

$$\frac{\partial^2 \mathcal{R}_{dif}^{(i)}}{\partial {\mathbf{u}_i}^2} = -{\mathbf{L}_{\mathbf{u}}}^{\dagger} \mathbf{L}_{\mathbf{u}} \ . \tag{6.49}$$

Finally, all the gradients and Hessians reported above are used within a Gauss-Newton optimisation scheme, to update the estimates of the initial velocity fields, as follows

$$\mathbf{u}_i^{iter} = \mathbf{u}_i^{iter-1} - (\boldsymbol{H})^{-1} \boldsymbol{g} \ , \tag{6.50}$$

where

$$\boldsymbol{g} = \frac{\partial \mathcal{D}^{(i)}}{\partial \mathbf{u}_i} + \frac{\partial \mathcal{R}_{dif}^{(i)}}{\partial \mathbf{u}_i} \ , \tag{6.51}$$

and

$$\boldsymbol{H} = \frac{\partial^2 \mathcal{D}^{(i)}}{\partial {\mathbf{u}_i}^2} + \frac{\partial^2 \mathcal{R}_{dif}^{(i)}}{\partial {\mathbf{u}_i}^2} \ . \tag{6.52}$$

Updating the affine parameters

Similarly to the strategy outlined above for the diffeomorphisms, the affine parameters, for each image i, can also be updated (i.e. optimised), so as to maximise of the following objective function

$$\mathcal{E}_{af}^{(i)} = \mathcal{D}^{(i)} + \mathcal{R}_{af}^{(i)} = \sum_{j=1}^{N_i} \sum_{k=1}^{K} \gamma_{ijk} \log \frac{w_{ik} \pi_k(\boldsymbol{\xi}_{ij})}{\sum_{c=1}^{K} w_{ic} \pi_c(\boldsymbol{\xi}_{ij})} - \frac{1}{2} \mathbf{a}_i^T \boldsymbol{\Sigma}_{\boldsymbol{a}}^{-1} \mathbf{a}_i \ , \tag{6.53}$$

with respect to $\mathbf{a}_i$.

The gradients and Hessians, which are useful in this case are reported below.

In particular, for the matching term the following derivatives need to be computed

$$\frac{\partial \mathcal{D}^{(i)}}{\partial \mathbf{a}_i} = \sum_{j=1}^{N_i} \sum_{k=1}^{K} \left(\gamma_{ijk} - \frac{w_{ik}\, \pi_k(\boldsymbol{\xi}_{ij})}{\sum_{c=1}^{K} w_{ic}\, \pi_c(\boldsymbol{\xi}_{ij})} \right) \boldsymbol{g}^{\pi}_{jk} \ , \tag{6.54}$$

where $\boldsymbol{g}^{\pi}_{jk}$ is defined as

$$\boldsymbol{g}^{\pi}_{jk} = \mathbf{B}_i^T \left([\boldsymbol{\phi}_{ij}\, , 1] \otimes \nabla \left[\log \left(\pi_k(\boldsymbol{\xi}_{ij}) \right) \right] \right) \ , \tag{6.55}$$

with

$$\mathbf{B}_i^T = \frac{\partial \mathbf{S}_i}{\partial \mathbf{a}_i} \ , \tag{6.56}$$

and

$$\mathbf{S}_i = \begin{bmatrix} \mathbf{T}_i & \mathbf{t}_i \\ 0 & 1 \end{bmatrix} \ . \tag{6.57}$$

$$\begin{aligned} \frac{\partial^2 \mathcal{D}^{(i)}}{\partial \mathbf{a}_i^2} = & \sum_{j=1}^{N_i} \left(\sum_{k=1}^{K} \frac{w_{ik}\, \pi_k(\boldsymbol{\xi}_{ij})}{\sum_{c=1}^{K} w_{ic}\, \pi_c(\boldsymbol{\xi}_{ij})} \boldsymbol{g}^{\pi}_{jk} \right) \left(\sum_{k=1}^{K} \frac{w_{ik}\, (\pi_k(\boldsymbol{\xi}_{ij}))}{\sum_{c=1}^{K} w_{ic}\, (\pi_k(\boldsymbol{\xi}_{ij}))} \boldsymbol{g}^{\pi}_{jk} \right)^T \\ & - \sum_{j=1}^{N_i} \sum_{k=1}^{K} \frac{w_{ik}\, \pi_k(\boldsymbol{\xi}_{ij})}{\sum_{c=1}^{K} w_{ic}\, \pi_c(\boldsymbol{\xi}_{ij})} \boldsymbol{g}^{\pi}_{jk} \left(\boldsymbol{g}^{\pi}_{jk} \right)^T \ . \end{aligned} \tag{6.58}$$

Gradients and Hessians of the penalty term are instead given by

$$\frac{\partial \mathcal{R}^{(i)}_{af}}{\partial \mathbf{a}_i} = -\boldsymbol{\Sigma}_{\boldsymbol{a}}^{-1} \mathbf{a}_i \ , \tag{6.59}$$

$$\frac{\partial^2 \mathcal{R}^{(i)}_{af}}{\partial \mathbf{a}_i^2} = -\boldsymbol{\Sigma}_{\boldsymbol{a}}^{-1} \ . \tag{6.60}$$

6.4. Experimental results

6.4.1. Template construction

Data

The modelling scheme and the resulting algorithm illustrated in this chapter were used to construct average-shaped brain and cervical spinal cord templates, from a multivariate (i.e. multichannel) data set of structural MR images of the head and the neck.

The input data was obtained from three different databases, two of which are freely accessible for download, thus ensuring that the results presented here could readily be compared to those produced by competing algorithms for medical image registration or segmentation.

First data set

The first data set consists of thirty five T1-weighted MR scans from the OASIS (Open Access Series of Imaging Studies) database (Marcus et al., 2007). The data is freely available from the web site `http://www.oasis-brains.org`, where details on the population demographics and acquisition protocols are also reported. Additionally, the selected thirty five subjects are the same ones that were used within the 2012 MICCAI Multi-Atlas Labelling Challenge (Landman and Warfield, 2012).

Second data set

The second data set consists of scans of twenty healthy adults, acquired at University Hospital Balgrist with a 3T scanner (Siemens Magnetom Verio). Magnetisation-prepared rapid acquisition gradient echo (MPRAGE) sequences, at 1 mm isotropic resolution, were used to obtained T1-weighted data, while PD-weighted images of the same subjects were acquired with a multiecho 3D fast low-angle shot (FLASH) sequence, within a whole-brain multi-parameter mapping protocol (Helms et al., 2008; Weiskopf et al., 2013).

Third data set

The third and last data set comprises twenty five T1-, T2- and PD-weighted scans of healthy adults from the freely available IXI brain database, which were acquired at Guy's Hospital, in London, on a 1.5T system (Philips Medical Systems Gyroscan Intera). Additional information regarding the demographics of the population, as well as the acquisition protocols, can be found at `http://brain-development.org/ixi-dataset`.

The complete data set therefore consists of eighty multispectral scans of healthy adults, obtained with fairly diverse acquisition protocols and using scanning systems produced by different vendors.

Unfortunately, not all the three modalities of interest (T1-, T2- and PD-weighted) are available for all of the subjects. To circumvent the difficulties arising from the presence of missing imaging modalities, without neglecting any of the available data (indeed deletion of entries with missing data is still, in spite of its crudity, a common statistical practice), the Gaussian mixture modelling approach discussed in Chapter 5 was generalised by introducing an additional variational posterior distribution over the missing data points.

In practice, the resulting variational EM scheme iterates over first estimating an approximated posterior distribution on the unknown image intensities, secondly updating the sufficient statistics of the complete (observed and missing) data and finally computing variational posteriors on the Gaussian mixture parameters. Additional computational details relative to this strategy are provided in Appendix D.

In practice, even if the presence of missing data slows down the convergence of the inference algorithm, with this approach it was possible to fit the generative groupwise model described in this chapter to the entire data set, in spite of having different imaging modalities available from the different acquisition sites. This is indeed a very common scenario in real life medical imaging problems, therefore it should be actively addressed by processing or modelling solutions that claim to be applicable to large population data (van Tulder and de Bruijne, 2015).

Manual brain labels are available for all images in data set one. Such labels have been generated and made public by Neuromorphometrics, Inc. (http://Neuromorphometrics.com) under academic subscription and they provide a fine parcellation of cortical and non-cortical structures, for a total of 139 labels across the brain. A list of all the labelled structures and their average volume across the population is reported in Appendix E.

Part of this label data was used for training of the model while the remaining was left out for testing and validation. In particular, brain labels of twenty out of the thirty five OASIS subjects were used to create gray and white matter ground truth segmentations, which were then provided as training input for semisupervised model fitting.

Similarly, spinal cord manual labels were created for forty subjects (twenty from data set two and twenty from data set three). Such labels were randomly split in half for training and half for subsequent test analyses. Due to the limited resolution of the data it was not possible to manually delineate gray and white matter within the spinal cord. For this reason, each voxel classified as spinal cord in the training data was allowed to be assigned either to the gray or to the white matter tissue classes, based on the fit of its intensity value to the underlying Gaussian mixture model.

Analogously, in spite of having defined only one gray matter training label, two distinct gray matter classes were introduced in the mixture model (top two rows in Figure 6.2), to best capture the corresponding distribution of image intensities, which is poorly represented by a single Gaussian component, as opposed to the distribution of white matter intensities. Also in this case, membership probabilities of the labelled training data were computed by combining the available labels with the intensity model.

Tissue templates and intensity priors

The tissue probability maps obtained by applying the modelling framework presented in this chapter to the data set described above are depicted in Figure 6.2. The total number of tissue classes used for this experiment is equal to twelve but three classes, representing air in the background, are not shown. In principle it would have been possible to automatically estimate the optimal number of Gaussian components, as in chapter 5. However, due to the size of the data set used here, setting an initial number of components higher than the unknown optimal one was found to be too onerous from a computational point of view. In particular, Figure 6.2 shows how one of the two gray matter classes (first row) best fits the subcortical nuclei and also includes voxels affected by partial volume effects at the interface between gray and white matter, while the second one (second row) is more representative of cortical structures, with the presence of partial volume effects generated by the juxtaposition of gray matter and CSF. The third row in Figure 6.2 shows the white matter class, which also includes most of the brainstem and the spinal cord.

The remaining tissue classes were estimated in a purely unsupervised way. Therefore a non-ambiguous anatomical interpretation is not straightforward.

Tissue class four (fourth row) mainly contains CSF, even if other tissues are present, especially in the neck area. This should be attributed to the lack of CSF training labels as well as to a poor multivariate coverage of the cervical region in the available data. In fact, data from the OASIS set is truncated around the first cervical vertebra.

The T1-weighted scans of the IXI data set cover up the C2/C3 vertebral level, but the corresponding T2- and PD-weighted scans do not extend beyond the brainstem. Indeed, only the data from the second database (Balgrist hospital) provides more than one modality covering up to around the fourth cervical vertebra. In this case though, additional difficulties arose from poor inter-modality alignment of the data, a problem which turned out to be particularly severe in the cervical region and that could not be compensated for by rigid realignment (i.e. coregistration), due to the non-linearity of the changes in head positioning. Such problems, which occur rather commonly when working with medical image data, can significantly affect the performance of model fitting, therefore an interesting direction for future work could be the introduction of intra-subject deformation fields within the modelling framework presented in this thesis.

Bone tissue is also not easily identifiable from the data available for this experiment, but it could have potentially been much better extracted by incorporating some CT scans into the training data.

Fat and soft tissues are mainly represented in the last two classes (bottom two rows in Figure 6.2).

Figure 6.3 illustrates a three-dimensional rendering of the average-shaped brain and spinal cord (6.3a), obtained by extracting a boundary surface from the sum of the gray and white matter tissue classes, as well as a three-dimensional model of the white matter class (6.3b). A T1-weighted template, obtained by linear averaging of the spatially normalised and bias corrected data, is reported instead in Figure 6.4.

The empirical Bayes learning procedure, introduced in the previous chapter (see Section 5.6.1), to estimate suitable prior distributions on the parameters of the Gaussian mixture model, was applied here to the same data that was used to construct the templates. Some of the results are summarised in figure 6.5, which reports the estimated empirical prior distribution on the mean intensity of gray and wite matter in T1- and PD-weighted data, with overlaid contour plots showing some of the individual posteriors (randomly selected across the entire population).

Such results indicate that the proposed empirical Bayes learning scheme can serve to capture, not only the variability of mean tissue intensity across subjects, for each of the modalities of interest, but also the amount of covariance between such modalities. Information of this sort can potentially be used in a number of different frameworks, for

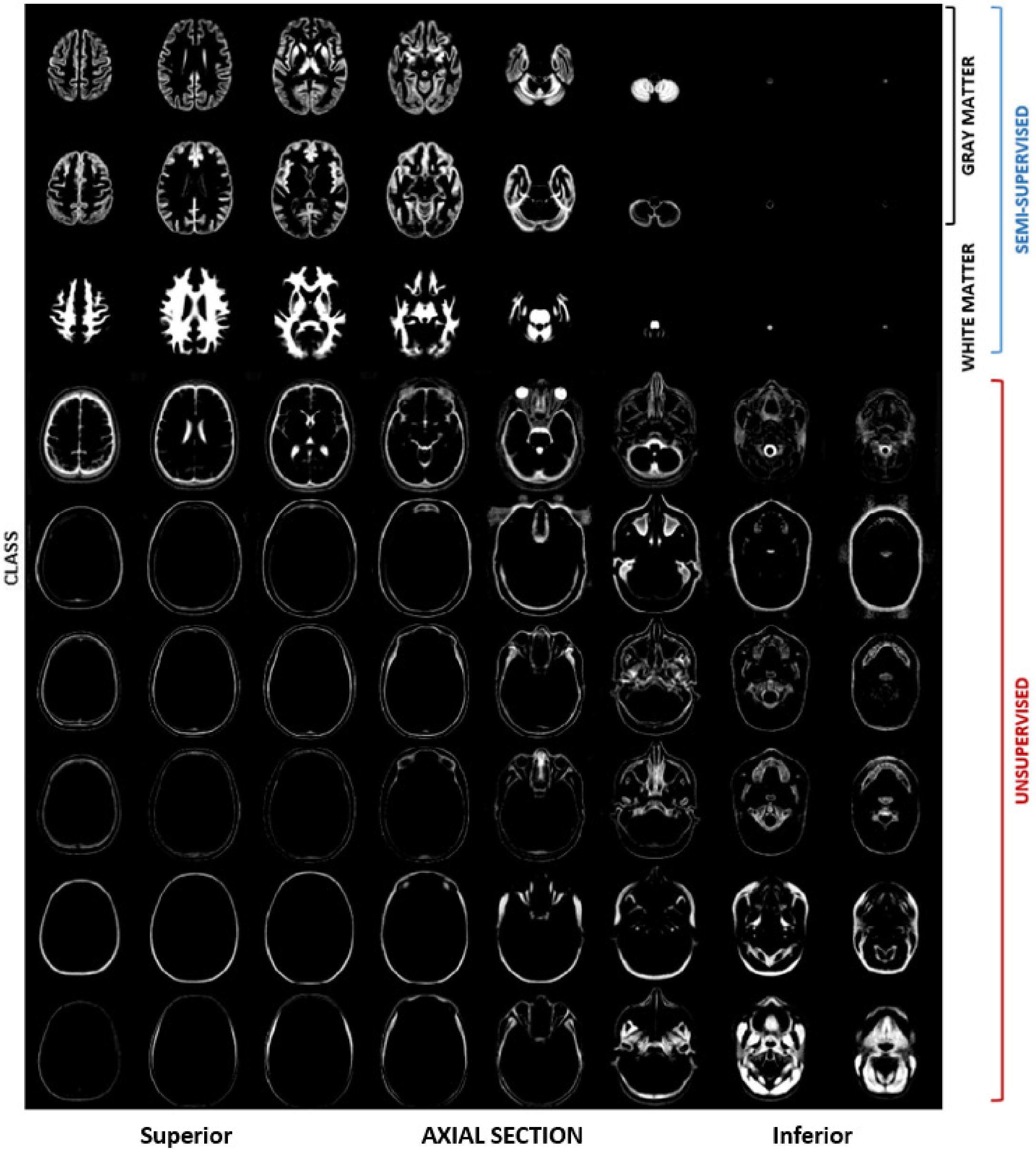


Figure 6.2: Head and neck tissue probability maps obtained by applying the presented group-wise generative model to a multispectral data set comprising scans of eighty healthy adults, from three different databases.

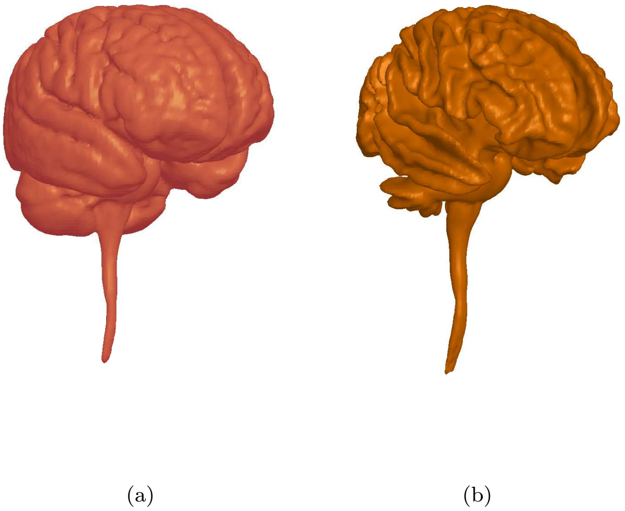


Figure 6.3: Average shaped brain and spine three-dimensional rendering, obtained by surface extraction of the sum of gray and white matter tissue classes (a), together with a 3-D rendering of the white matter prior model (b).

solving problems such as tissue segmentation, pathology detection or image synthesis.

Validity of groupwise registration

The performance of groupwise registration achieved by the presented algorithm was assessed by computing pairwise overlap measures for all possible couples of spatially normalised test images (i.e. the images provided with ground truth labels, which had not been used for training of the model). The Dice score coefficient was chosen as a metric of similarity.

Results are summarised in figures 6.6, 6.7 and 6.8, where the accuracy of the algorithm presented here is compared to that achieved by the method described in Avants et al. (2010), whose implementation is publicly available, as part of the Advanced normalisation Tools (ANTs 1.9) package, through the web site `http://stnava.github.io/ANTs/`. Indeed, the symmetric diffeomorphic registration framework implemented in ANTs has established itself as the state-of-the-art of medical image non-linear spatial normalisation (Klein et al., 2009).

A number of options can be customised within the template construction frame-

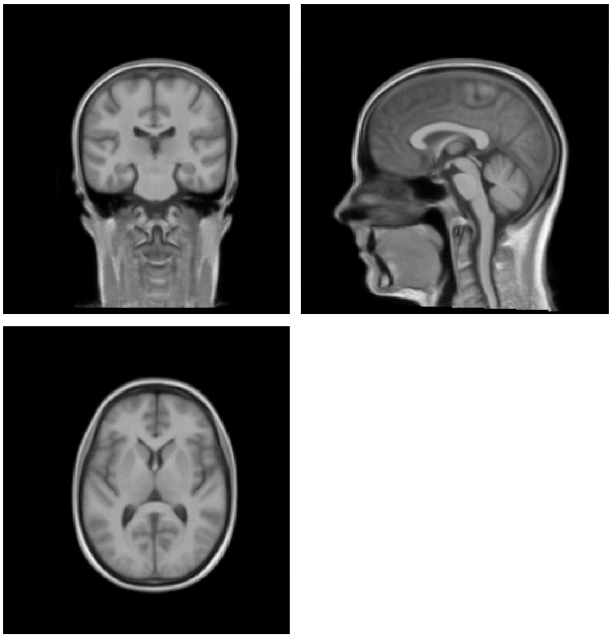


Figure 6.4: T1-weighted average-shaped template obtained by linear averaging of spatially normalised individual scans.

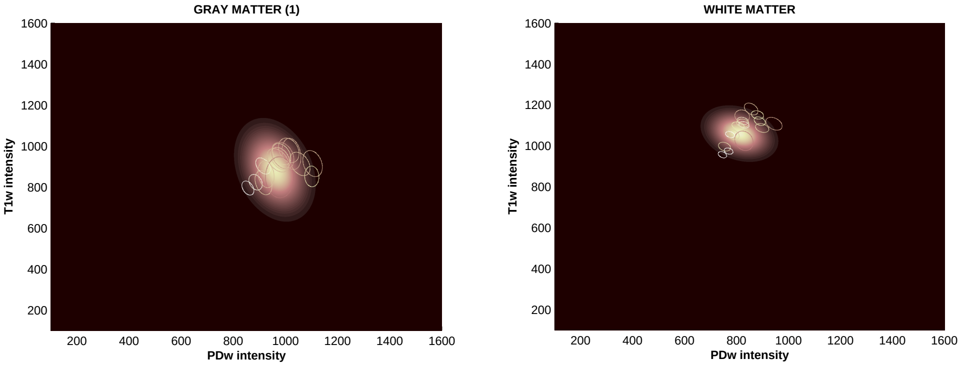


Figure 6.5: Prior distributions over the mean intensity of gray and white matter in T1- and PD-weighted data.

Table 6.2: Options selected to perform groupwise registration with ANTs, using the `antsMultivariateTemplateConstruction` *script provided with the ANTS package.*

Option	**Value**
Similarity Metric	**Cross-correlation (CC)**
Transformation model	**Greedy SyN (GR)**
Initial rigid body	**yes**
N4 Bias Correction	**yes**
Number of resolution levels	**4**
Number of iterations	$\mathbf{100 \times 70 \times 50 \times 10}$
Gradient step	**0.2**
Number of template updates	**4**

work distributed with ANTs. The experiments, whose results are reported here, were performed using only T1-weighted scans since the package does not handle the presence of missing data, with the settings recommended for brain MR data in the software documentation, which are also reported in table 6.2. Additionally, the data was not skull-stripped prior to model fitting.

Results of this validation analyses indicate that the method presented here, in spite of not being as accurate as ANTs for aligning some subcortical brain structures (e.g. thalamus, putamen, pallidum and brainstem), can provide significantly better overlap when registering cortical regions, as assessed by means of paired t-tests with a significance threshold of 0.05 and without correcting for multiple comparisons. No statistically significant differences were found between the two methods, with respect to registration of the spinal cord. This results should also be compared to the ones reported in Chapter 4 (see Figure 4.13 and Figure 4.14), which were obtained on a subset of the data used here (data set one only), in a ML setting, as opposed to the VB approach exploited in this chapter, and with a small deformation model rather than a diffeomorphic one. Not only does the model adopted in this chapter outperform the previous one in terms of mean accuracy but it exhibits much higher robustness, as indicated by the dramatic decrease in the number of negative outliers. Most of the cases of registration failure observed with the model presented in Chapter 3 corresponded to scans that were largely misaligned with the initial group average. For such images, due to a poor initialisation of model parameters, the optimisation algorithm got prematurely stuck in a local maximum. As indicated in Chapter 5 however, the introduction of intensity priors ensures much higher robustness to misregistration and this, in combination with a larger parametrisation of the deformations, explains the improvement in performance obtained here.

Accuracy of tissue classification

The accuracy of tissue classification achieved by the method presented here was first evaluated on test data, which was used to create the templates but without providing manual labels for training of the model. Dice scores were computed to compare the automated segmentations produced via semisupervised groupwise model fitting, with the ground truth, obtained by merging all the gray and white matter brain structures (labels) into two tissue classes respectively, and by considering the spinal cord as a third separate class.

All probabilistic brain segmentations were thresholded at a value equal to 0.5, in order to obtain binary label maps, directly comparable to the ground truth. To derive binary spine segmentations instead, the sum of gray and white matter posterior belonging probabilities was first computed in a subvolume containing the neck only, and then thresholded at 0.5.

Results are summarised in Figure 6.9, which shows the distributions of Dice scores obtained for brain gray matter, brain white matter and spinal cord.

Such results were then compared to those produced by the brain segmentation algorithm implemented in SPM12, using the standard tissue probability maps distributed with the software. Results of these analyses, which are summarised in Figure 6.10, indicate that the population specific atlases constructed with the method presented here enable higher tissue classification accuracy, at least for test data drawn from the same population that the model was trained on but whose labels were not exploited for training. A potential source of bias in the results of this experiment is the fact that the test data was actually employed for constructing the atlases, even though the corresponding labels were not seen by the algorithm. However a more cautious k-fold cross-validation,

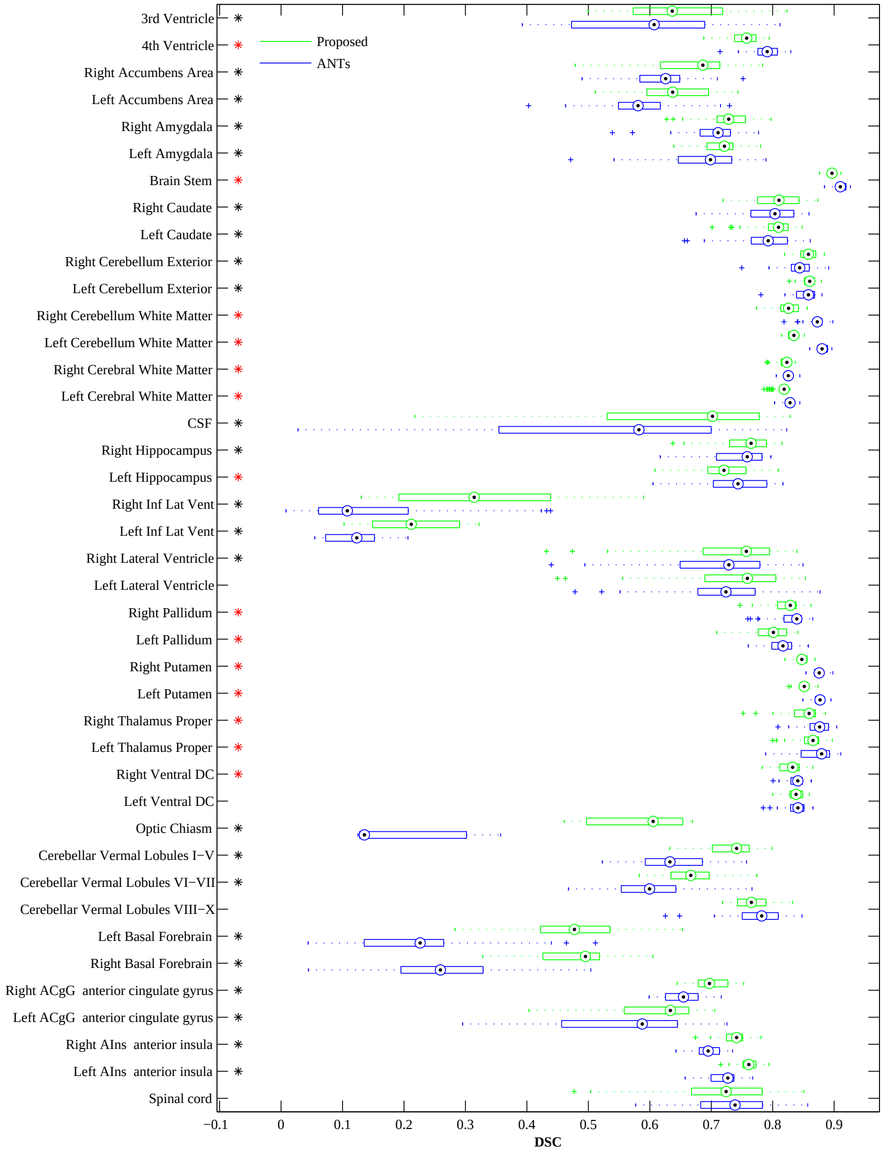


Figure 6.6: Accuracy of groupwise registration achieved by the presented method, compared to the performance of ANTs, for different neural regions. Stars indicate statistically significant differences between the two methods, assessed by means of paired t-tests with a significance level of 0.05.

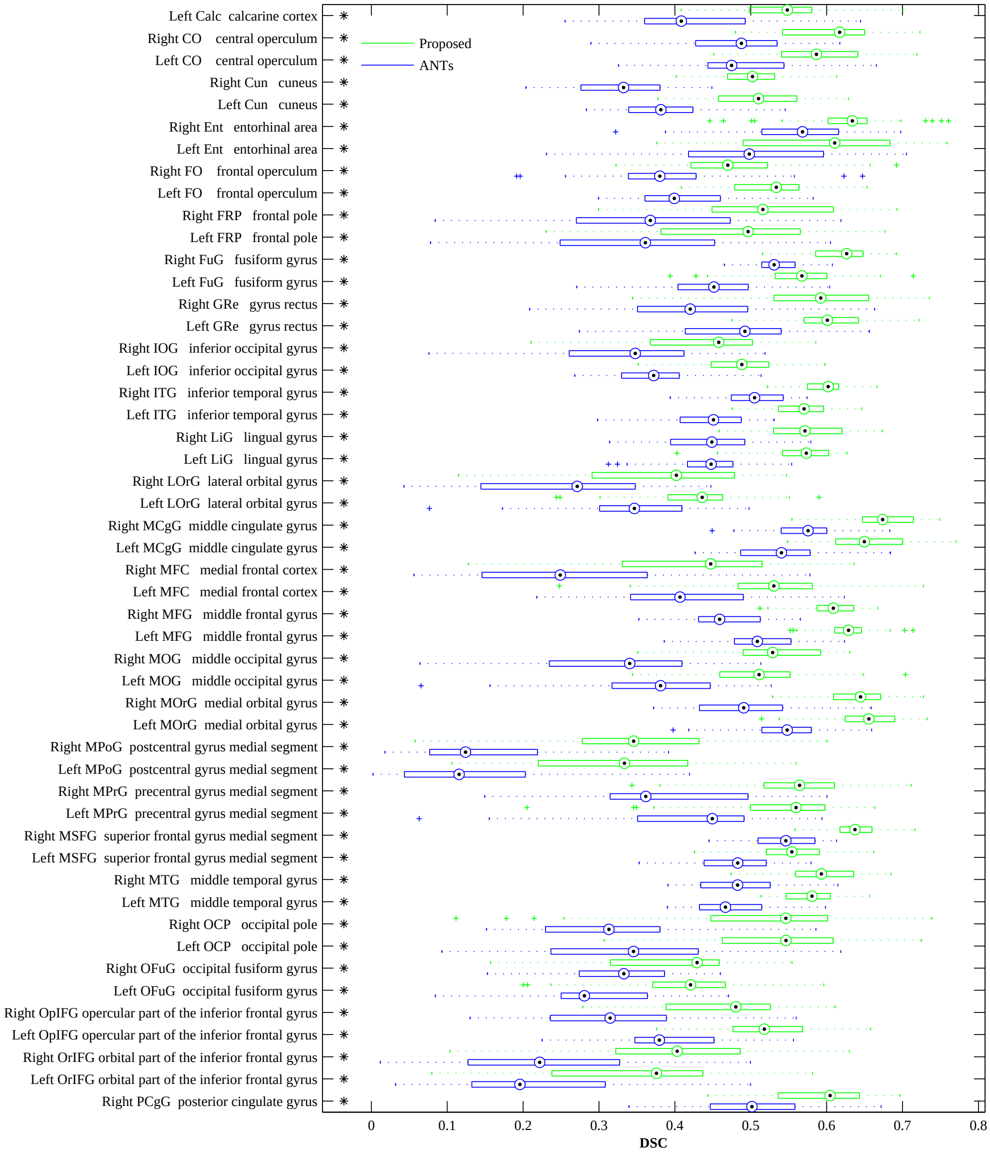


Figure 6.7: Accuracy of groupwise registration achieved by the presented method, compared to the performance of ANTs, for different neural regions. Stars indicate statistically significant differences between the two methods, assessed by means of paired t-tests with a significance level of 0.05.

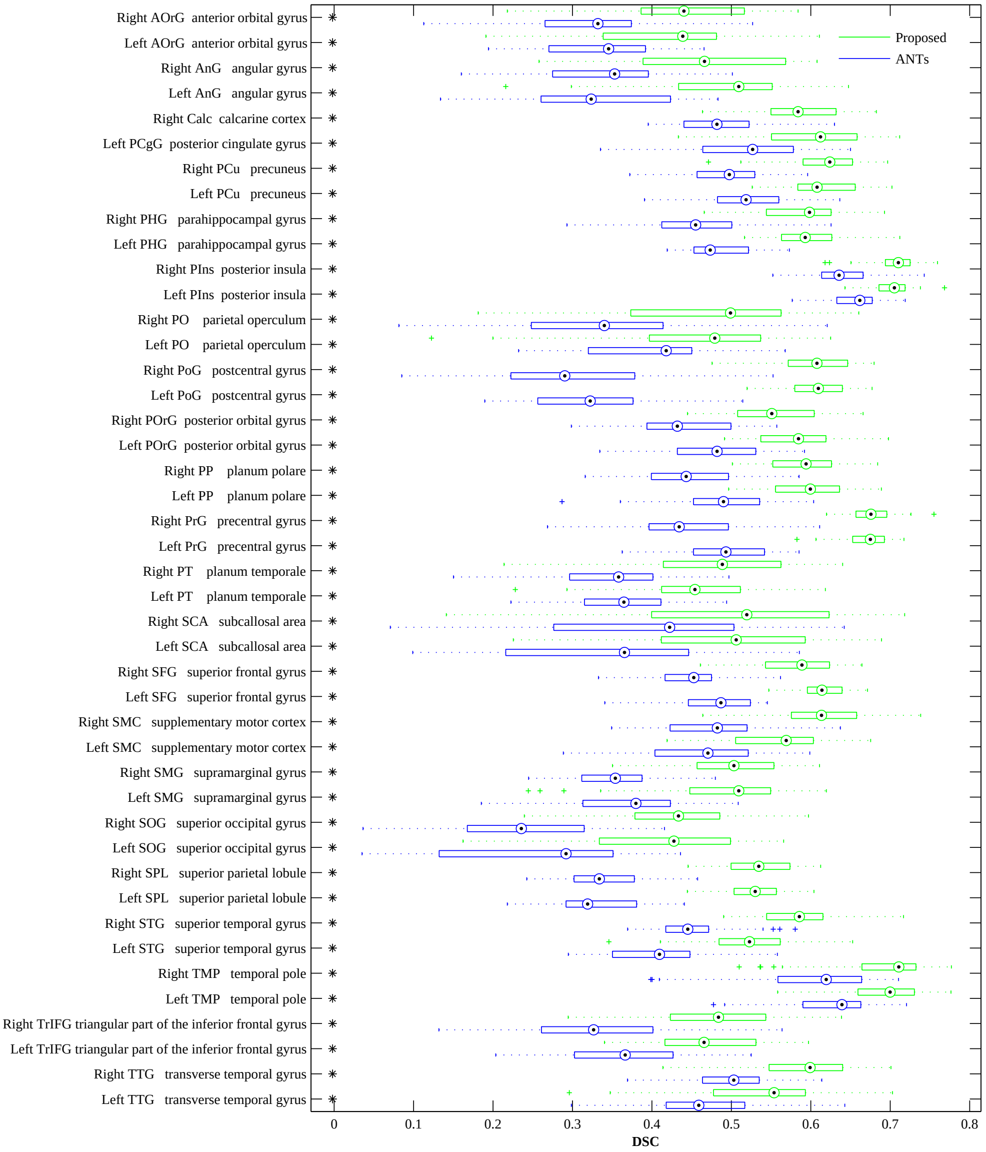


Figure 6.8: Accuracy of groupwise registration achieved by the presented method, compared to the performance of ANTs, for different neural regions. Stars indicate statistically significant differences between the two methods, assessed by means of paired t-tests with a significance level of 0.05.

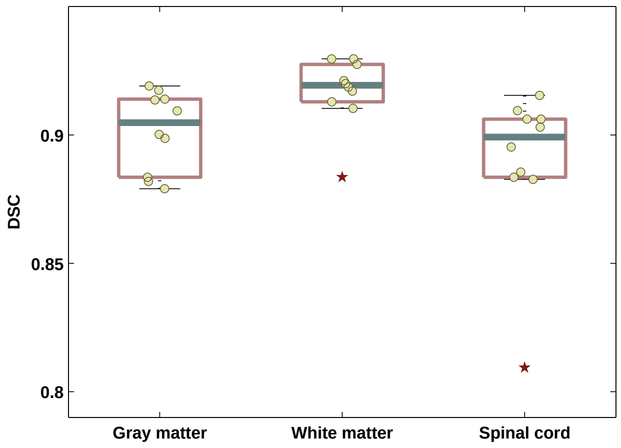


Figure 6.9: Brain and spinal cord segmentation accuracy of the presented method.

which would have required constructing numerous templates, was not feasible in this case due to the expensive computational cost of groupwise model fitting, which for the data set used here was around 160 hours on a Quad-Core PC at 3.19 GHz with 30 GB of RAM.

6.4.2. Modelling unseen data

Further validation experiments were performed to quantify the accuracy of the framework described in this chapter to model unseen data, that is to say data that was not included in the atlas generation process.

Such experiments were performed on synthetic T1-weighted brain MR scans from the Brainweb database (http://brainweb.bic.mni.mcgill.ca/).

Accuracy of bias correction

A healthy adult brain MR model was processed by means of the algorithm discussed here, using the head and neck templates previously constructed as tissue priors. Different noise and bias field levels were added to the uncorrupted synthetic data, to test the behavior of the proposed modelling scheme in different noise (1%, 3%, 7%) and bias conditions (20% and 40%).

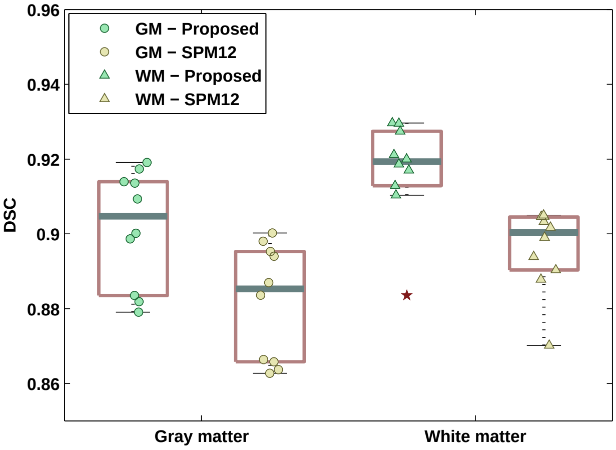


Figure 6.10: Brain segmentation accuracy of the presented modelling framework compared to SPM12. An equivalent comparison was not possible for the spinal cord, as the current SPM templates do not handle neck data.

The noise in these simulated images has Rayleigh statistics in the background and Rician statistics in the signal regions and its level is computed as a percent standard deviation ratio, relative to the MR signal, for a reference tissue (Cocosco et al., 1997).

Regarding the bias field instead, 20% bias is modelled as a smooth field in the range [0.9, 1.1] while 40% bias is obtained by rescaling of the 20% field, so as to range between 0.8 and 1.2 .

Table 6.3 reports the Pearson product-moment correlation coefficients between the ground truth and the estimated bias fields, for the different bias ranges and noise levels.

Table 6.3: Pearson's correlation coefficients between the ground truth bias fields and those estimated by the presented algorithm, for simulated T1-weighted data.

		Noise		
		1%	**3%**	**7%**
Bias	**20%**	0.86	0.86	0.70
	40%	0.72	0.72	0.51

Results indicate that the similarity between the estimated and true bias decreases for more intense non-uniformity fields and higher noise levels. Indeed this is not surprising, as the penalty term, which enforces smoothness of the bias field, has a greater impact in determining the shape of the estimated bias when the non-uniformities have a larger dynamic range, such as at higher field strengths (Vaughan et al., 2001). Nevertheless, results reported in the following section will show how this increased mismatch between the estimated and true bias does not seem to affect the accuracy of tissue segmentation. On the other hand, the accuracy of bias correction is directly related to the amount of noise corrupting the data, mainly due to how this affects the precision associated with estimatation of the Gaussian mixture parameters. Results reported here are in line with those presented in Chapter 5, using the probabilistic atlas publicly available in SPM12. This confirms the accuracy of the proposed approach and indicate that the templates learned with the model discussed in this chapter could be effectively integrated with existing modelling tools for neuroimaging data.

Accuracy of tissue classification

For the same data, the accuracy of tissue classification was also evaluated by comparing the similarity between the estimated gray and white matter segmentations and the underlying anatomical model.

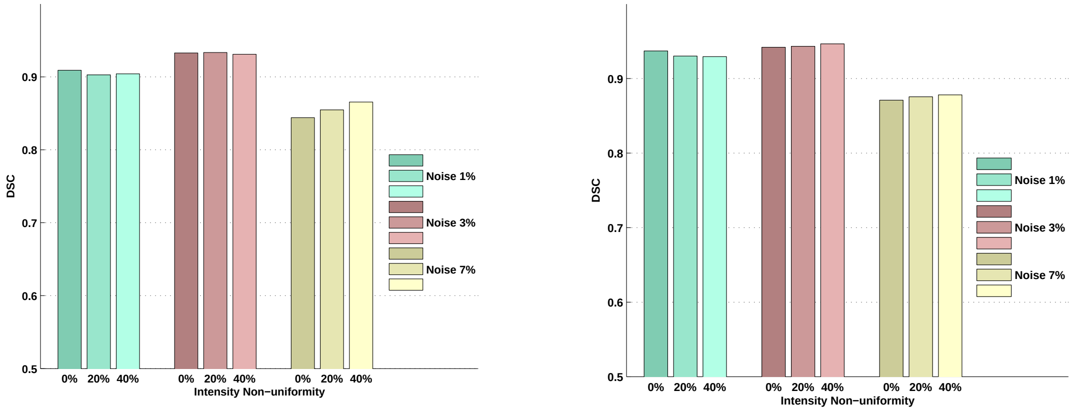
Results are reported in Figure 6.11, which shows the Dice score coefficients obtained under different bias and noise conditions.

The Brainweb database has been extensively used in the neuroimaging community to validate MR image processing algorithms. Therefore the results reported here should be directly comparable to the performance of many brain segmentation techniques present in the literature.

A potential framework for image synthesis

One of the potential applications of the generative model presented in this chapter is related to the creation of synthetic images.

This is a research problem that has been approached with a variety of different techniques, among which are compressed sensing (Roy et al., 2011), regression trees (Jog et al., 2013), convolutional neural networks (Li et al., 2014), patch-matching algorithms



(a)

(b)

Figure 6.11: Dice scores between the estimated and ground truth segmentations for brain white matter (a) and brain gray matter (b), under different noise and bias conditions, for synthetic T1-weighted data.

(Iglesias et al., 2013a), atlas-based fusion (Burgos et al., 2014), generative modelling of MR intensity distributions (Cordier et al., 2016) and Bayesian estimation of the physical quantities (i.e. relaxation times and proton density) that govern nuclear magnetisation phenomena (Maitra and Besag, 1998; Maitra and Riddles, 2010).

Such synthetic images can be useful for a number of different purposes. For example to alleviate the problems caused by the inconsistency of contrasts generated with different scanning systems and acquisition protocols in the context of large multi centre studies (Jovicich et al., 2009; Tofts, 1998), or to deal with missing data, without deleting incomplete observations, by means of data imputation (Campos et al., 2015; van Tulder and de Bruijne, 2015), or, just to provide another example, to augment data sets for the training of image processing algorithms (Ronneberger et al., 2015).

Within the model adopted here, a missing data value $\mathbf{x}_{mis}$, corresponding for example to the intensity of a non-acquired image contrast, can be estimated from the observed data (i.e. acquired contrasts) $\mathbf{x}_{obs}$ by

$$\hat{\mathbf{x}}_{mis} = \underset{\mathbf{x}_{mis}}{\arg\max} \log p(\mathbf{x}_{mis}|\mathbf{x}_{obs}, \hat{\Theta}_\pi, \hat{\Theta}_\beta, \hat{\Theta}_u) \; . \tag{6.61}$$

In a variational Bayes setting, an approximated posterior distribution over the missing data values can be computed as

$$\begin{aligned}
&\log p(\mathbf{x}_{mis}|\mathbf{x}_{obs}, \hat{\Theta}_{\pi}, \hat{\Theta}_{\beta}, \hat{\Theta}_{u}) \simeq \\
&\mathbb{E}_{\mathbf{z},\Theta_{\mu},\Theta_{\Sigma}}\left[\log p(\mathbf{x}_{obs}, \mathbf{x}_{mis}, \mathbf{z}, \Theta_{\mu}, \Theta_{\Sigma}, \hat{\Theta}_{\pi}, \hat{\Theta}_{\beta}, \hat{\Theta}_{u})\right] + \text{const} = \\
&\mathbb{E}_{\mathbf{z},\Theta_{\mu},\Theta_{\Sigma}}\left[\log p(\mathbf{x}_{obs}, \mathbf{x}_{mis}|\mathbf{z}, \Theta_{\mu}, \Theta_{\Sigma}, \hat{\Theta}_{\beta})\right] + \text{const} = \\
&\sum_{k=1}^{K} \gamma_k \, \mathbb{E}_{\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k}\left[\log \mathcal{N}(\mathbf{x}_{obs}, \mathbf{x}_{mis}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \hat{\Theta}_{\beta})\right] + \text{const} \; ,
\end{aligned} \tag{6.62}$$

where $\{\hat{\Theta}_{\pi}, \hat{\Theta}_{\beta}, \hat{\Theta}_{u}\}$ denotes a set of model parameter estimates relative respectively to the Gaussian mixing proportions, the bias field and the deformations, while γ_k is the posterior belonging probability of tissue class k.

The expectations that appear in the last line of (6.62) are computed with respect to posterior distributions on $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$. Such posteriors should capture, for every tissue type k, the patterns of (co)variability of image intensities across modalities. Therefore, if the missing contrast is unobserved on the entire volume of interest, informative intensity priors must be adopted. Additional mathematical details on how to perform variational inference on Gaussian mixtures, in the presence of missing data, are reported in Appendix D.

An example of a synthetic T2-weighted scan generated from a T1-weighted image included the IXI data set, making use of the tissue probability maps and the empirical intensity priors estimated from the training data sets described previously, is reported in Figure 6.12, together with a true T2-weighted image of the same subject. Such an example confirms that the proposed generative modelling framework could potentially have an application for image synthesis. One limitation however is that the magnitude of the covariance between different modalities in the intensity prior model is smaller than the corresponding variances, thus causing the simulated data to lie very close to the mean intensity value of the predominant tissue class at each voxel.

6.5. Summary

This chapter presented a general groupwise Bayesian modelling framework, which, in spite of having a number of potential applications, is primarily intended to enable simultaneous morphometric analyses of the brain and the cervical cord, from cross-sectional

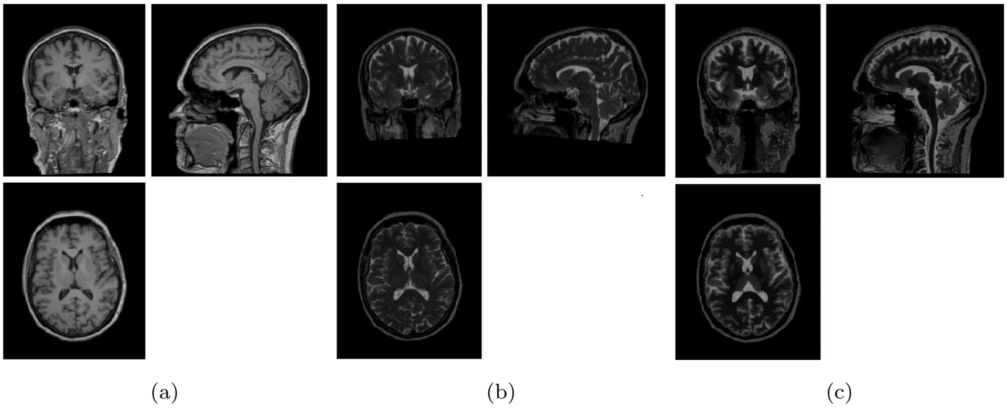


Figure 6.12: *Synthetic T2-weighted scan (c), generated from a T1-weighted volume (a) using informative intensity priors, and corresponding ground truth (b).*

MRI data sets. From a theoretical perspective, such a framework relies on variational probability density estimation techniques to model the observed data (i.e. MR signal intensities), by exploiting the VBEM framework introduced in Chapter 5. Additionally, a hierarchical modelling perspective is proposed, where observations from a population of subjects are integrated to construct both tissue probability maps and empirical intensity priors, which can then serve to inform models of new data.

Shape modelling, in this case, is performed via groupwise diffeomorphic registration, thus ensuring bijective (i.e. one-to-one) differentiable mappings between anatomical configurations (Miller, 2004). Such an approach enables a rigorous mathematical encoding of anatomical shapes via deformable template matching (Christensen et al., 1996), therefore providing a quantitative framework for the analysis of shape variation and covariation.

Data for training the method was collected from three different databases, two of which are publicly accessible to the research community. Results of validation experiments performed both on training and unseen test data indicated that the presented model is suitable to perform integrated brain and cervical cord morphometrics. Thus, the proposed algorithm could represent a concrete solution to extract anatomical volumetric and morphometric information from large neuroimaging data sets, in a fully automated manner. At the same time it could provide outputs that might be readily interpreted, for instance via statistical hypothesis testing, with the ultimate goal of comparing different populations, treatment effects etc. (Ashburner and Friston, 2000).

Finally, it was shown how the described framework also has potential for application in the field of image synthesis. Even if such a topic was not extensively addressed in this thesis, a proof of concept was provided, indicating the feasibility of such an approach.

7

Conclusion

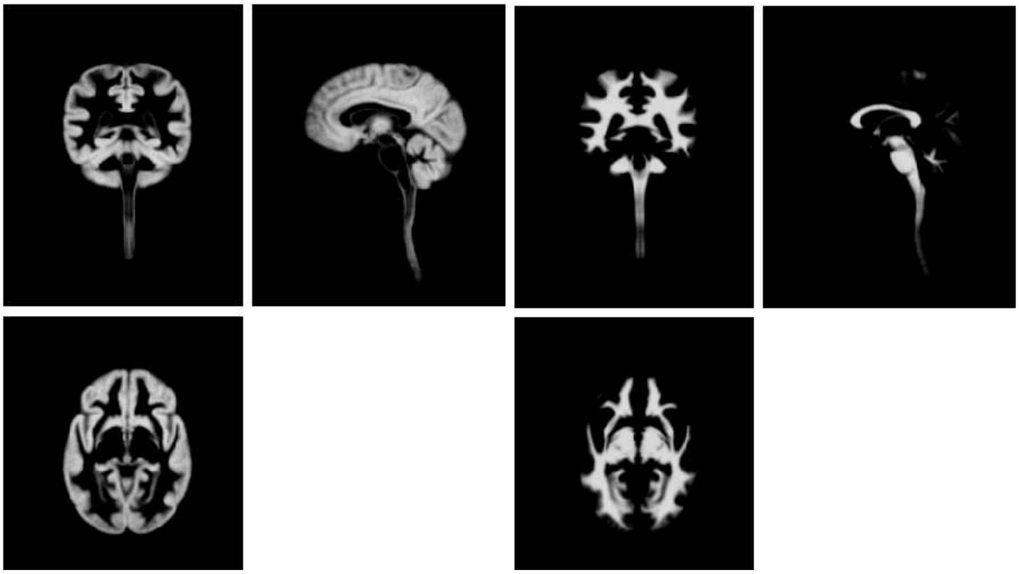
7.1. Contribution of this thesis

The work presented in this thesis has explored the potential of generative modelling approaches to capture anatomical and morphological features from large structural MR data sets. In particular, the proposed framework, which builds on the work of Ashburner and Friston (2005) to expand its hierarchical structure, allows combining image registration, tissue classification, bias correction, atlas construction and intensity prior learning in a single groupwise algorithm. This is achieved by formulating one comprehensive mathematical representation of the data, such that these individual processing tasks can be conceived as interdependent elements within the same generative process.

With such a perspective, not only diverse image processing problems can be addressed in the same framework, thus leading to general and flexible computational solutions, but each compartment of the mathematical model informs the others, resulting in higher accuracy, compared to independent model fitting of the single components.

Additionally, different model fitting strategies, namely maximum likelihood estimation, maximum a posteriori estimation and variational Bayes, have been implemented, compared and, when convenient, integrated to find a trade-off between accuracy and computational feasibility.

A number of experimental findings have been reported, mainly for the purpose of evaluating the behaviour of the proposed approach to process brain and spinal cord MR data, both in a fully unsupervised and semisupervised learning setting. Results



(a) (b)

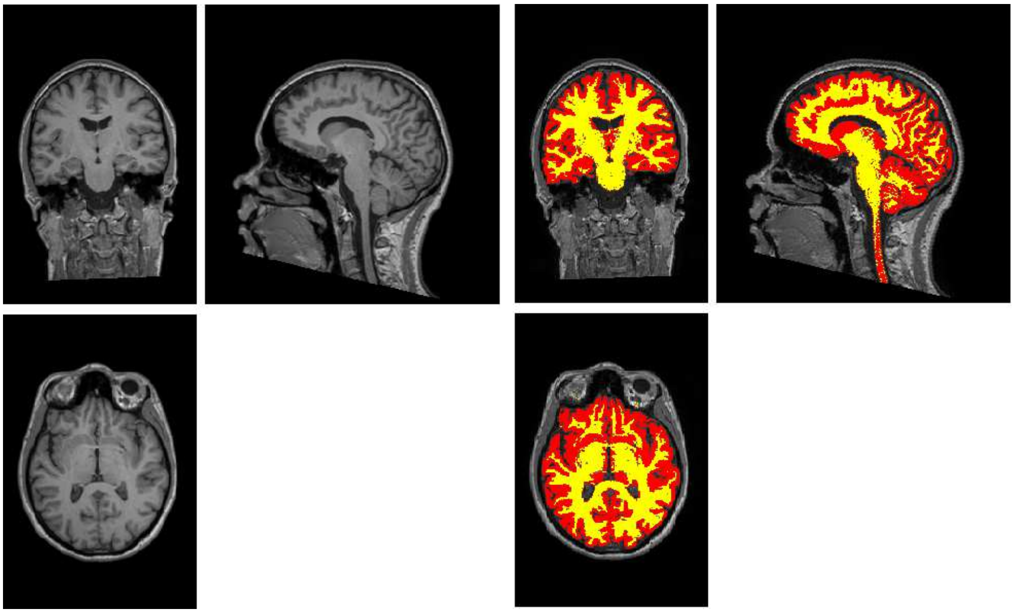
Figure 7.1: Brain and spine tissue probability maps of gray (a) and white (b) matter, constructed as described in Chapter 6.

of these analyses indicated the viability of integrated brain and spine morphometrics, thus opening up a new perspective in computational neuroimaging, where the central nervous system can actually be considered as an integrated structure and interactions between its compartments, for instance brain and spine, can be inferred.

7.2. Limitations

There are a number of limitations associated with the work presented in this thesis. A first crucial point is related to the amount of shape variability that the proposed model can capture. Indeed, in spite of having adopted a large deformation approach in Chapter 6, an intrinsic difficulty remains in modelling large shape variations that deviate significantly from the average shape model, as built from the training data. In such cases, finding a reasonable trade off between maximising image similarity and preserving topology becomes especially challenging, thus increasing the chance of incurring implausible warps or suboptimal local solutions (Crum et al., 2003).

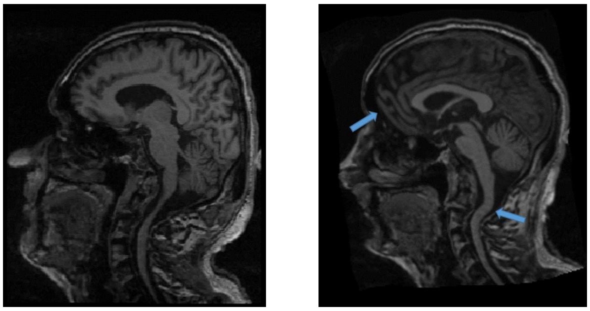
A practical example of misregistration is provided in Figure 7.3. In this case, a sub-



(a)

(b)

Figure 7.2: Example of brain and spine segmentation obtained by applying the modelling framework described in Chapter 6. The image in panel (b) is partitioned into gray and white matter (a).



(a) (b)

Figure 7.3: Example of registration failure of an individual scan (a) to the templates described in Chapter 6 (see Figure 6.2). Arrows indicate regions of severe misalignment in the warped image (b).

ject with pronounced cervical extension (a) fails to be registered to the templates shown in Figure 6.2, with the warped image (b) exhibiting an implausible spinal curvature and shrinkage of the frontal lobe.

Additionally, not having modelled the presence of partial volume effects can induce systematic misclassification of voxels that lie at the interface between different tissues. A typical example encountered in neuroimaging regards those voxels containing a mixture of white matter and cerebrospinal fluid, which in T1-weighted images tend to have an intensity overlapping with that of gray matter structures, thus easily leading to misclassification. Such a problem could be tackled, within the same mixture modelling scheme adopted here, by representing each voxel as an unknown mixture of different tissues, that is introducing continuous latent variables $\{\mathbf{z}_j\}_{j=1,\ldots,N}$, with $z_{jk} \in [0,1]$ rather than $z_{jk} \in \{0,1\}$ (Heller et al., 2008). In such a way, the pure tissue case would become just a special instance of a more general formulation (Van Leemput et al., 2003). Alternatively, a more simplistic approach would involve introducing additional classes to encode partial volume effects (Noe and Gee, 2001)

Another limitation is related to the difficulty of normalising intensity profiles across images acquired with different systems or protocols, which can lead to systematic biases (Weisenfeld and Warfteld, 2004). Results reported in the previous chapters indicated that the proposed method is capable of handling a certain amount of variability in the intensity scaling, by employing informative intensity priors and allowing a linear rescaling of the MRI signals. Nevertheless this remains a practical but perfectible solution, which could potentially benefit from exploring alternative intensity transforms, so as to increase robustness to acquisition dependent differences (Weisenfeld and Warfteld, 2004).

Finally, the model described here is generally unable to capture lesions or abnormalities, unless they are effectively represented in the training data, with consistent patterns across the entire population, which happens rarely in neuropathology.

7.3. Future directions

In relation to the limitations outlined above, a number of potential directions for future work arise. For instance, the difficulties induced by the inconsistency of MR signal intensities could be tackled by directly embedding physical models of the MR signal generation process into the proposed hierarchical Bayesian framework (Glad and Sebastiani, 1995), so as to derive an explicit model of how image intensities depend on the MR scanning parameters. Not only would this allow creating synthetic MR images corresponding to any combination of acquisition settings but, by including information on the covariation between MR and CT signal intensities, it could also define a strategy to obtain simulated CT images, photon attenuation maps or electron density maps, which are necessary for example for attenuation correction in PET/MR reconstruction (Burgos et al., 2013) or for accurate treatment planning in radiation therapy (Gudur et al., 2014).

Another topic, which could potentially be explored as part of future work, concerns the possibility of taking into account, within the proposed semi-supervised generative modelling framework, the uncertainty inherent in the process of manual rating. For such a purpose, posterior class probabilities could be computed by making use of the categorical output of manual labelling together with an estimate of the rater sensitivity and with a generative intensity model. In fact, the work of Warfield et al. (2004) has shown that reliable sensitivity estimates can be inferred in an automated manner by exploiting a probabilistic modelling scheme.

The medical imaging community has also shown considerable interest in deformable shape models that are constrained so as to ensure that all generated shape instances are statistically plausible for a given anatomical object (Cootes et al., 1995; Rueckert et al., 2003). For instance, active shape models, introduced by Cootes et al. (1995), describe the shape of an object through the relative position of a set of landmarks or labelled points, whose location is modelled by a mean plus a linear combination of a small number of modes of variation (point distribution model). Such modes of variation are computed via principal component analysis (PCA) on the deviations from the mean of the coordinates of corresponding data points in different training samples, after having accounted for positioning, orientation and size differences.

Such an approach is valid for landmark-based representations but not for more complex shape models. To circumvent this limitation, principal geodesic analysis (PGA) was introduced by Fletcher et al. (2004) as a generalisation of principal component analysis, which is only applicable in Euclidean vector spaces (Wold et al., 1987), for the purpose of describing geometric variability on curved manifolds. Analogously to PCA in the Euclidean space, PGA seeks lower dimensional subspaces that best capture the variability of data samples. In PCA these subspaces are linear subspaces. The corresponding generalisation in the manifold setting is provided by the notion of geodesic subspaces.

A PGA approach could be usefully incorporated in the shape modelling scheme presented here so as to increase its robustness, while inferring principal modes of brain and spine shape variability. Indeed, it has already been shown that such an approach can be rigorously formulated in a Bayesian setting (Zhang and Fletcher, 2013), which would be well suited for integration into the presented framework.

Another open research question is related to how models of healthy anatomy, such as the one presented in this work, can be generalised to handle the presence of pathological features. In this case additional challenges arise, since lesion morphology tends to exhibit even larger variability compared to healthy tissues, thus requiring very large training data sets in order to build informative priors of lesion shape, intensity, texture etc.

In a fully unsupervised setting, the simplest approach would involve treating lesions as outliers (Freifeld et al., 2007), which would essentially require a physiological model of intensity and shape variability for the tissues of interest, such as the one presented in this work. The variational framework discussed in this thesis might as well be useful for such a purpose, as it would allow to quantitatively evaluate the uncertainty relative to the estimates of the Gaussian mixture parameters, thus providing implicit criteria to detect the presence of abnormal intensity patterns. On the other hand, for pathologies that significantly alter anatomical shape but without affecting the Gaussian distribution parameters, the use of informative intensity priors, as illustrated in Chapter 5 and Chapter 6, should ensure better model fit compared to maximum likelihood techniques.

However, an approach of this sort is still likely to perform sub-optimally in the absence of contextual or shape information, especially if the available data is not multispectral. For this purpose, the generative framework presented in this thesis might only be useful when lesions tend to appear at similar spatial locations across different subjects. In such a case, a factorisation model, extending the binary logistic regression approach of Tipping (1999) to the multinomial case, could be used to provide a more flexible version of the prior information encoded in the tissue probability maps. Otherwise, different forms of priors should be incorporated, for instance in the form of Markov random fields to ensure spatial coherence (Schwarz et al., 2009) or by exploiting statistical shape models (Shepherd et al., 2012) to encode information on lesion shape variability. For instance, convolutional restricted Boltzmann machines have successfully been used to learn pathological shape priors from manually annotated data (Agn et al., 2016).

However, the fact that fully unsupervised generative learning tends to exhibit higher asymptotic prediction error compared to discriminative classification techniques (Jordan, 2002), might still hinder the application of generative models for capturing lesions, since they are intrinsically harder to explain compared to healthy features. Therefore an intuitive solution would be combining unsupervised learning methods, which can be easily trained on large data sets, with supervised classification approaches, such as deep neural networks, support vector machines or random forests (Havaei et al., 2017; Lao et al., 2008; Vaidya et al., 2015; Zacharaki et al., 2009), which yield high predictive performance but in some real life applications suffer from the limited availability of labelled examples. Approaches of this sort to the problems of lesion segmentation and computer assisted diagnosis have been explored, with encouraging preliminary results, by Alex et al. (2017); Batmanghelich et al. (2012); Guo et al. (2015); Jerman et al. (2015); Menze et al. (2016); Reddick et al. (1998); van Tulder and de Bruijne (2016). Therefore further research progress in this direction should possibly be pursued.

Appendix A

Gaussian-Wishart priors

The Gaussian-Wishart distribution is the conjugate prior of a multivariate D-dimensional normal distribution with unknown mean $\boldsymbol{\mu}$ and precision matrix $\boldsymbol{\Lambda}$. Its probability density function is

$$\begin{aligned} p(\boldsymbol{\mu}, \boldsymbol{\Lambda} | \boldsymbol{m}, \beta, \boldsymbol{W}, \nu) &= p(\boldsymbol{\mu} | \boldsymbol{\Lambda}, \boldsymbol{m}, \beta) p(\boldsymbol{\Lambda} | \boldsymbol{W}, \nu) \\ &= \mathcal{N}(\boldsymbol{\mu} | \boldsymbol{m}, (\beta \boldsymbol{\Lambda})^{-1}) \mathcal{W}(\boldsymbol{\Lambda} | \boldsymbol{W}, \nu) \ , \end{aligned} \tag{A.1}$$

with

$$\mathcal{N}(\boldsymbol{\mu} | \boldsymbol{m}, (\beta \boldsymbol{\Lambda})^{-1}) = \frac{|\beta \boldsymbol{\Lambda}|^{1/2}}{(2\pi)^{D/2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{m})^T \boldsymbol{\Lambda} \, (\boldsymbol{\mu} - \boldsymbol{m}) \right\} \ , \tag{A.2}$$

and

$$\mathcal{W}(\boldsymbol{\Lambda} | \boldsymbol{W}, \nu) = B_W(\boldsymbol{W}, \nu) |\boldsymbol{\Lambda}|^{\frac{\nu - D - 1}{2}} \exp\left\{ -\frac{1}{2} \operatorname{Tr}\left(\boldsymbol{W}^{-1} \boldsymbol{\Lambda} \right) \right\} \ . \tag{A.3}$$

The normalising constant B_W is given by

$$B_W(\boldsymbol{W}, \nu) = |\boldsymbol{W}|^{-\nu/2} \left(2^{\nu D/2} \pi^{D(D-1)/4} \prod_{i=1}^{D} \Gamma\left(\frac{\nu + 1 - i}{2} \right) \right)^{-1} \ , \tag{A.4}$$

where $\Gamma(\cdot)$ is the gamma function

$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} \mathrm{d}u \ . \tag{A.5}$$

The expectation of the determinant of the precision matrix, which appears in equation 5.25 (VE-step), is equal to (Bishop, 2006)

$$\mathbb{E}[\log |\boldsymbol{\Lambda}|] = \sum_{i=1}^{D} \psi \left(\frac{\nu + 1 - i}{2} \right) + D \log 2 + \log |\boldsymbol{W}| \ , \tag{A.6}$$

where $\psi(\cdot)$ indicates the digamma function, which is the logarithmic derivative of the gamma function

$$\psi(x) = \frac{d}{dx} \log \Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)} \; . \tag{A.7}$$

The following expectation has also to be computed during the VE-step

$$\mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\Lambda}}\left[(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Lambda}\,(\mathbf{x}-\boldsymbol{\mu})\right] = D\beta^{-1} + \nu(\mathbf{x}-\boldsymbol{m})^T \boldsymbol{W}(\mathbf{x}-\boldsymbol{m}) \; . \tag{A.8}$$

Appendix B

Variational Gaussian mixtures: derivation of the Gaussian-Wishart posterior update rules

Under the mean field theory assumption, a variational posterior on a subset of parameters $\boldsymbol{\Upsilon}_{\hat{s}}$ can be computed by

$$q_{\hat{s}}(\boldsymbol{\Upsilon}_{\hat{s}}) \propto \exp(\mathbb{E}_{s \neq \hat{s}}[\log p(\mathbf{X}, \boldsymbol{\Upsilon})]) \ . \tag{B.1}$$

where the expectations are evaluated with respect to variational posteriors on the remaining sets $\{\boldsymbol{\Upsilon}_s\}_{s \neq \hat{s}}$.

For a Gaussian mixture probability distribution with conjugate Gaussian-Wishart priors, this gives

$$q(\Theta_\mu, \Theta_\Sigma) \propto \exp\left\{ \sum_{j=1}^{N} \sum_{k=1}^{K} \gamma_{jk} \log \mathcal{N}(\mathbf{B}_j \mathbf{x}_j | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \sum_{k=1}^{K} \log\left(\mathcal{N}(\boldsymbol{\mu}_k | \boldsymbol{\Sigma}_k) \mathcal{W}(\boldsymbol{\Sigma}_k)\right) \right\} , \tag{B.2}$$

where $\{\mathbf{B}_j\}_{j=1,\ldots,N}$ is a multiplicative field applied to the observations $\{\mathbf{x}_j\}_{j=1,\ldots,N}$. From equation B.2 it is possible to obtain

$$\begin{aligned} \log q(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k^{-1}) = & \log \mathcal{N}(\boldsymbol{\mu}_k | \boldsymbol{m}_{0k}, \beta_{0k}^{-1} \boldsymbol{\Sigma}_k) \\ & + \log \mathcal{W}(\boldsymbol{\Sigma}_k^{-1} | \boldsymbol{W}_{0k}, \nu_{0k}) \\ & + \sum_{j=1}^{N} \gamma_{jk} \log \mathcal{N}(\mathbf{B}_j \mathbf{x}_j | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \text{const}, \end{aligned} \tag{B.3}$$

which can be expanded, to give

$$\begin{aligned}\log q(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k^{-1}) = & - \frac{\beta_{0k}}{2}(\boldsymbol{\mu}_k - \boldsymbol{m}_{0k})^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{\mu}_k - \boldsymbol{m}_{0k}) \\ & - \frac{1}{2}\sum_{j=1}^{N} \gamma_{jk}(\mathbf{B}_j\mathbf{x}_j - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{B}_j\mathbf{x}_j - \boldsymbol{\mu}_k) \\ & + \frac{1}{2}\log|\boldsymbol{\Sigma}_k^{-1}| + \frac{\nu_{0k} - D - 1}{2}\log|\boldsymbol{\Sigma}_k^{-1}| \\ & + \frac{1}{2}\sum_{j=1}^{N}\gamma_{jk}\log|\boldsymbol{\Sigma}_k^{-1}| - \frac{1}{2}\,\mathrm{Tr}\left((\boldsymbol{\Sigma}_k \boldsymbol{W}_{0k})^{-1}\right) + \text{const}\ .\end{aligned} \tag{B.4}$$

Let us first consider the terms containing $\boldsymbol{\mu}_k$

$$\begin{aligned}\log q(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k^{-1}) = & -\frac{1}{2}\sum_{j=1}^{N}\gamma_{jk}\left(\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}_k^{-1}(\mathbf{B}_j\mathbf{x}_j - \boldsymbol{\mu}_k) - (\mathbf{B}_j\mathbf{x}_j)^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k\right) \\ & - \frac{\beta_{0k}}{2}\left(\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{\mu}_k - \boldsymbol{m}_{0k}) - \boldsymbol{m}_{0k}^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k\right) \\ & + \text{const}\ .\end{aligned} \tag{B.5}$$

Rearranging and grouping of the different terms gives

$$\begin{aligned}\log q(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k^{-1}) = & + \boldsymbol{\mu}_k^T\boldsymbol{\Sigma}_k^{-1}\left(\beta_{0k}\boldsymbol{m}_{0k} + \sum_{j=1}^{N}\gamma_{jk}\mathbf{B}_j\mathbf{x}_j\right) \\ & - \frac{1}{2}\left(\beta_{0k} + \sum_{j=1}^{N}\gamma_{jk}\right)\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k + \text{const}\ .\end{aligned} \tag{B.6}$$

Finally, by completing the square, the following result is obtained

$$q(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k^{-1}) = \mathcal{N}(\boldsymbol{\mu}_k|\boldsymbol{m}_k, \beta_k^{-1}\boldsymbol{\Sigma}_k)\ , \tag{B.7}$$

with

$$\beta_k = \beta_{0k} + \sum_{j=1}^{N}\gamma_{jk}\ , \tag{B.8}$$

and

$$\boldsymbol{m}_k = \frac{\beta_{0k}\boldsymbol{m}_{0k} + \sum_{j=1}^{N}\gamma_{jk}\mathbf{B}_j\mathbf{x}_j}{\beta_{0k} + \sum_{j=1}^{N}\gamma_{jk}}\ . \tag{B.9}$$

The posterior $q(\boldsymbol{\Sigma}_k^{-1})$ can instead be computed by

$$\log q(\boldsymbol{\Sigma}_k^{-1}) = \log q(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k^{-1}) - \log q(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k^{-1})\ , \tag{B.10}$$

to give

$$\begin{aligned}\log q(\boldsymbol{\Sigma}_k^{-1}) =& -\frac{\beta_{0k}}{2}(\boldsymbol{\mu}_k - \boldsymbol{m}_{0k})^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{\mu}_k - \boldsymbol{m}_{0k}) \\ & -\frac{1}{2}\sum_{j=1}^{N}\gamma_{jk}(\mathbf{B}_j\mathbf{x}_j - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\mathbf{B}_j\mathbf{x}_j - \boldsymbol{\mu}_k) \\ & -\frac{1}{2}\,\mathrm{Tr}\left((\boldsymbol{\Sigma}_k\boldsymbol{W}_{0k})^{-1}\right) + \frac{\nu_{0k} - D - 1}{2}\log|\boldsymbol{\Sigma}_k^{-1}| \\ & +\frac{\beta_k}{2}(\boldsymbol{\mu}_k - \boldsymbol{m}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{\mu}_k - \boldsymbol{m}_k) \\ & +\frac{1}{2}\sum_{j=1}^{N}\gamma_{jk}\log|\boldsymbol{\Sigma}_k^{-1}| + \text{const}\ . \end{aligned} \tag{B.11}$$

Making use of the property $\mathbf{u}^T\boldsymbol{A}\mathbf{u} = \mathrm{Tr}(\boldsymbol{A}\mathbf{u}\mathbf{u}^T)$ allows rewriting of $q(\boldsymbol{\Sigma}_k^{-1})$ as

$$\begin{aligned} q(\boldsymbol{\Sigma}_k^{-1}) =& \frac{1}{2}\sum_{j=1}^{N}(\gamma_{jk} + \nu_{0k} - D - 1)\log|\boldsymbol{\Sigma}_k^{-1}| \\ & -\frac{1}{2}\,\mathrm{Tr}\Bigg\{\Big(\boldsymbol{W}_{0k}^{-1} + \beta_{0k}(\boldsymbol{\mu}_k - \boldsymbol{m}_{0k})(\boldsymbol{\mu}_k - \boldsymbol{m}_{0k})^T \\ & +\sum_{j=1}^{N}\gamma_{jk}(\mathbf{B}_j\mathbf{x}_j - \boldsymbol{\mu}_k)(\mathbf{B}_j\mathbf{x}_j - \boldsymbol{\mu}_k)^T \\ & -\beta_k(\boldsymbol{\mu}_k - \boldsymbol{m}_k)(\boldsymbol{\mu}_k - \boldsymbol{m}_k)^T\Big)\boldsymbol{\Sigma}_k^{-1}\Bigg\} + \text{const}\ . \end{aligned} \tag{B.12}$$

Finally, by substituting B.8 and B.9 into B.12, the following is obtained

$$q(\boldsymbol{\Sigma}_k^{-1}) = \mathcal{W}(\boldsymbol{\Sigma}_k^{-1}|\boldsymbol{W}_k, \nu_k)\ , \tag{B.13}$$

where

$$\nu_k = \nu_{0k} + \sum_{j=1}^{N}\gamma_{jk}\ , \tag{B.14}$$

and

$$\begin{aligned}\boldsymbol{W}_k^{-1} = \boldsymbol{W}_{0k}^{-1} + \sum_{j=1}^{N}\gamma_{jk}(\mathbf{B}_j\mathbf{x}_j)(\mathbf{B}_j\mathbf{x}_j)^T - \frac{\left(\sum_{j=1}^{N}\gamma_{jk}\mathbf{B}_j\mathbf{x}_j\right)\left(\sum_{j=1}^{N}\gamma_{jk}\mathbf{B}_j\mathbf{x}_j\right)^T}{\beta_{0k} + \sum_{j=1}^{N}\gamma_{jk}} \\ + \frac{\beta_{0k}\left(\sum_{j=1}^{N}\gamma_{jk}\right)\boldsymbol{m}_{0k}\boldsymbol{m}_{0k}^T}{\beta_{0k} + \sum_{j=1}^{N}\gamma_{jk}} - \frac{\beta_{0k}\left(\sum_{j=1}^{N}\gamma_{jk}\mathbf{B}_j\mathbf{x}_j\right)\boldsymbol{m}_{0k}^T}{\beta_{0k} + \sum_{j=1}^{N}\gamma_{jk}} - \frac{\beta_{0k}\boldsymbol{m}_{0k}\left(\sum_{j=1}^{N}\gamma_{jk}\mathbf{B}_j\mathbf{x}_j\right)^T}{\beta_{0k} + \sum_{j=1}^{N}\gamma_{jk}}\ . \end{aligned} \tag{B.15}$$

Appendix C

Variational Gaussian mixtures: derivatives of the lower bound with respect to the prior hyperparameters

Given a population of M independent observations (e.g. scans of different subjects), a lower bound on the marginal likelihood, for the Gaussian mixture model described in Chapter 5, can be expressed as a function of the set of Gaussian-Wishart hyperparameters $\Phi_0 = \{\beta_{0k}, \boldsymbol{m}_{0k}, \boldsymbol{W}_{0k}, \nu_{0k}\}_{k=1,\dots,K}$

$$
\begin{aligned}
\mathcal{L}(\Phi_0) &= \mathbb{E}_{\Theta_\mu,\Theta_\Sigma}[\log p(\Theta_\mu,\Theta_\Sigma)] \\
&= \sum_{i=1}^{m} \int\int q_i(\Theta_\mu,\Theta_\Sigma) \log p(\Theta_\mu,\Theta_\Sigma)\ \mathrm{d}\Theta_\mu \mathrm{d}\Theta_\Sigma + \text{const} \\
&= \frac{1}{2}\sum_{i=1}^{M}\sum_{k=1}^{K} \Bigg\{ \mathbb{E}\big[\log|\boldsymbol{\Sigma}_{ik}^{-1}|\big](\nu_{0k}-D) \\
&- \nu_{ik}\,\mathrm{Tr}(\boldsymbol{W}_{0k}^{-1}\boldsymbol{W}_{ik} + \beta_{0k}(\boldsymbol{m}_{ik}-\boldsymbol{m}_{0k})(\boldsymbol{m}_{ik}-\boldsymbol{m}_{0k})^T\boldsymbol{W}_{ik}) \Bigg\} \qquad \text{(C.1)} \\
&+ \frac{M}{2}\sum_{k=1}^{K} D\log\frac{\beta_{0k}}{2\pi} - D\sum_{i=1}^{M}\sum_{k=1}^{K}\frac{\beta_{0k}}{\beta_{ik}} \\
&+ 2M\sum_{k=1}^{K}\log B_W(\boldsymbol{W}_{0k},\nu_{0k}) + \text{const}\ ,
\end{aligned}
$$

where B_W indicates the normalising constant of a Wishart distribution and $\{\beta_{ik}, \boldsymbol{m}_{ik}, \boldsymbol{W}_{ik}, \nu_{ik}\}_{k=1,\dots,K}$ is a set of posterior Gaussian-Wishart hyperparameters relative to observation (e.g. subject) i.

The lower bound in (C.1) can be expressed as a function of the hyperparameters

$\{\beta_{0k}\}_{k=1,\dots,K}$ as follows

$$\begin{aligned}\mathcal{L}(\beta_{0k}) = \frac{MD}{2}\log\left(\frac{\beta_{0k}}{2\pi}\right) - \frac{1}{2}\sum_{i=1}^{M}\Bigg\{ & D\frac{\beta_{0k}}{\beta_{ik}} \\ & - \beta_{0k}\nu_{ik}(\boldsymbol{m}_{ik}-\boldsymbol{m}_{0k})^T\boldsymbol{W}_{ik}(\boldsymbol{m}_{ik}-\boldsymbol{m}_{0k})\Bigg\} \\ & + \text{const} \ , \end{aligned} \tag{C.2}$$

and the corresponding gradient and Hessian are given by

$$\begin{aligned} g_\beta &= \frac{MD}{\beta_{0k}} - \frac{1}{2}\sum_{i=1}^{M}\left\{\frac{D}{\beta_{ik}} - \nu_{ik}(\boldsymbol{m}_{ik}-\boldsymbol{m}_{0k})^T\boldsymbol{W}_{ik}(\boldsymbol{m}_{ik}-\boldsymbol{m}_{0k})\right\}, \\ \boldsymbol{H}_\beta &= -\frac{MD}{2\beta_{0k}^2}. \end{aligned} \tag{C.3}$$

Similarly for $\{\boldsymbol{m}_{0k}\}_{k=1,\dots,K}$ we find that $\mathcal{L}(\boldsymbol{m}_{0k})$ can be expressed as

$$\mathcal{L}(\boldsymbol{m}_{ok}) = \frac{1}{2}\sum_{i=1}^{M}\beta_{0k}\nu_{ik}(\boldsymbol{m}_{ik}-\boldsymbol{m}_{0k})^T\boldsymbol{W}_{ik}(\boldsymbol{m}_{ik}-\boldsymbol{m}_{0k}) + \text{const}. \tag{C.4}$$

The first and second derivatives are instead

$$\begin{aligned} g_{\boldsymbol{m}} &= -\sum_{i=1}^{M}\beta_{0k}\nu_{ik}(\boldsymbol{m}_{ik}-\boldsymbol{m}_{0k})^T\boldsymbol{W}_{ik} \ , \\ \boldsymbol{H}_{\boldsymbol{m}} &= \sum_{i=1}^{M}\beta_{0k}\nu_{ik}\boldsymbol{W}_{ik} \ . \end{aligned} \tag{C.5}$$

The following indicates the dependency of $\mathcal{L}$ on the degrees of freedom of the Wishart priors

$$\begin{aligned}\mathcal{L}(\nu_{0k}) = & \sum_{i=1}^{M}\frac{\nu_{0k}}{2}\mathbb{E}\big[\log|\boldsymbol{\Sigma}_{ik}^{-1}|\big] + M\log|\boldsymbol{W}_{0k}|^{-\frac{\nu_{0k}}{2}} \\ & + M\log\left(2^{\frac{D\nu_{0k}}{2}}\pi^{\frac{D(D-1)}{4}}\prod_{d=1}^{D}\Gamma\left(\frac{\nu_{0k}+1-d}{2}\right)\right)^{-1} \\ & + \text{const} \ . \end{aligned} \tag{C.6}$$

In this case the gradient and Hessian can be computed by

$$\begin{aligned} g_\nu = & \frac{1}{2}\sum_{i=1}^{M}\mathbb{E}\big[\log|\boldsymbol{\Sigma}_{ik}^{-1}|\big] \\ & - \frac{M}{2}\left\{\log|\boldsymbol{W}_{0k}| + D\log 2 + \sum_{d=1}^{D}\psi\left(\frac{\nu_{0k}+1-d}{2}\right)\right\} \ , \\ \boldsymbol{H}_\nu = & M\psi_1\left(\frac{\nu_{0k}+1-d}{2}\right) \ , \end{aligned} \tag{C.7}$$

where $\psi(\cdot)$ and $\psi_1(\cdot)$ are the digamma and trigamma functions respectively, that is the first and second logarithmic derivatives of the gamma function.

Finally for the Wishart scale matrices we find that

$$\begin{aligned}\mathcal{L}(\boldsymbol{W}_{0k}) &= M\nu_{0k}\log|\boldsymbol{C}_{0k}| \\ &-\frac{1}{2}\sum_{i=1}^{M}\nu_{ik}\,\mathrm{Tr}(\boldsymbol{C}_{0k}^{T}\boldsymbol{W}_{ik}\boldsymbol{C}_{0k}) + \mathrm{const}\ ,\end{aligned} \tag{C.8}$$

where $\boldsymbol{C}_{0k}$ is the Cholesky factor of $\boldsymbol{W}_{0k}^{-1}$

$$\boldsymbol{W}_{0k}^{-1} = \boldsymbol{C}_{0k}\boldsymbol{C}_{0k}^{T}\ . \tag{C.9}$$

The first and second derivatives are given by

$$\begin{aligned} g_{\boldsymbol{W}} &= M\nu_{0k}\,\mathrm{diag}(1/C_{11},\ldots,1/C_{DD}) - \sum_{i=1}^{M}\nu_{ik}\boldsymbol{W}_{ik}\boldsymbol{C}_{0k}\ , \\ \boldsymbol{H}_{\boldsymbol{W}} &= -M\nu_{0k}\,\mathrm{diag}(1/C_{11}^{2},\ldots,1/C_{DD}^{2}) - \sum_{i=1}^{M}\nu_{ik}\boldsymbol{W}_{ik}\ . \end{aligned} \tag{C.10}$$

Appendix D

Variational Gaussian mixtures: inference of missing data

The variational Bayes EM algorithm for fitting Gaussian mixture models, described in Chapter 5, can be generalised to handle the case where some components of the D-dimensional observation $\mathbf{x}_j$ are missing.

Having denoted

$$\mathbf{x}_j = \begin{bmatrix} \mathbf{o}_j \\ \mathbf{h}_j \end{bmatrix} , \tag{D.1}$$

with $\mathbf{o}_j$ being the observed data and $\mathbf{h}_j$ the missing data, the Gaussian likelihood $p(\mathbf{x}_j|z_{jk}=1,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$ can be expressed as

$$p(\mathbf{x}_j|z_{jk}=1,\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k) = \mathcal{N}\left(\begin{bmatrix} \mathbf{o}_j \\ \mathbf{h}_j \end{bmatrix} \middle| \begin{bmatrix} \boldsymbol{\mu}_k^o \\ \boldsymbol{\mu}_k^h \end{bmatrix} , \begin{bmatrix} \boldsymbol{\Lambda}_k^{o,o} & \boldsymbol{\Lambda}_k^{o,h} \\ \boldsymbol{\Lambda}_k^{o,h} & \boldsymbol{\Lambda}_k^{h,h} \end{bmatrix} \right) , \tag{D.2}$$

by making use of block matrix notation to partition the mean vector $\boldsymbol{\mu}_k$ and the precision matrix $\boldsymbol{\Lambda}_k$.

In this case $\mathbf{h}_j$ is treated as an unobserved random variable. Thus, in a variational Bayes setting, an additional posterior factor can be introduced for each missing data point $\mathbf{h}_j$ to give

$$q(\mathbf{H},\mathbf{Z},\Theta_\mu,\Theta_\Sigma) = q(\mathbf{H})q(\mathbf{Z})q(\Theta_\mu,\Theta_\Sigma) = q(\mathbf{Z})q(\Theta_\mu,\Theta_\Sigma)\prod_{j=1}^{N} q(\mathbf{h}_j) . \tag{D.3}$$

Making use of the general result in (5.10), an approximated posterior on the missing data point $\mathbf{h}_j$ can be computed by

$$\begin{aligned}\log q(\mathbf{h}_j) &= \mathbb{E}_{\mathbf{Z},\Theta_\mu,\Theta_\Sigma}\left[\log p(\mathbf{x}_j, \mathbf{z}_j, \Theta_\mu, \Theta_\Sigma | \Theta_\pi)\right] + \text{const} \\ &= \mathbb{E}_{\mathbf{Z},\Theta_\mu,\Theta_\Sigma}\left[\log p(\mathbf{z}_j|\Theta_\pi) + \log p(\mathbf{x}_j|\mathbf{z}_j, \Theta_\mu, \Theta_\Sigma) + \log p(\Theta_\mu, \Theta_\Sigma)\right] + \text{const} \;,\end{aligned} \tag{D.4}$$

where Θ_π denotes the mixing proportion parameter set, treated here via maximum likelihood, and $p(\Theta_\mu, \Theta_\Sigma)$ is a conjugate Gaussian-Wishart prior on the means and covariances of the model.

Ignoring the terms independent from $\mathbf{h}_j$, equation (D.4) can be rewritten as

$$\begin{aligned}\log q(\mathbf{h}_j) =& \sum_{k=1}^{K} \gamma_{jk}\, \mathbb{E}_{\Theta_\mu,\Theta_\Sigma}\left[\log \mathcal{N}(\mathbf{x}_j|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right] + \text{const} \\ =& \frac{1}{2}\sum_{k=1}^{K} \gamma_{jk} \mathbf{h}_j^T \mathbb{E}_{\Theta_\mu,\Theta_\Sigma}\left[\boldsymbol{\Lambda}_k^{h,h}\right] \mathbf{h}_j \\ & + \sum_{k=1}^{K} \gamma_{jk} \mathbf{h}_j^T \mathbb{E}_{\Theta_\mu,\Theta_\Sigma}\left[\boldsymbol{\Lambda}_k^{o,h}\right] \left(\mathbf{o}_j - \mathbb{E}_{\Theta_\mu,\Theta_\Sigma}\left[\boldsymbol{\mu}_k^o\right]\right) \\ & - \sum_{k=1}^{K} \gamma_{jk} \mathbf{h}_j^T \mathbb{E}_{\Theta_\mu,\Theta_\Sigma}\left[\boldsymbol{\Lambda}_k^{h,h}\right] \mathbb{E}_{\Theta_\mu,\Theta_\Sigma}\left[\boldsymbol{\mu}_k^h\right] + \text{const} \;.\end{aligned} \tag{D.5}$$

The previous equation indicates that the unobserved value $\mathbf{h}_j$ is drawn from a Gaussian mixture distribution with mixing proportions equal to the posterior (after having observed $\mathbf{o}_j$) membership probabilities $\{\gamma_{jk}\}_{k=1,\dots,K}$, while the Gaussian means $\{\mathbf{n}_{jk}\}_{k=1,\dots,K}$ and covariances $\{\mathbf{P}_{jk}\}_{k=1,\dots,K}$ are given by

$$\mathbf{n}_{jk} = \mathbb{E}_{\Theta_\mu,\Theta_\Sigma}\left[\boldsymbol{\mu}_k^h\right] + \left(\mathbb{E}_{\Theta_\mu,\Theta_\Sigma}\left[\boldsymbol{\Lambda}_k^{h,h}\right]\right)^{-1} \mathbb{E}_{\Theta_\mu,\Theta_\Sigma}\left[\boldsymbol{\Lambda}_k^{o,h}\right] \left(\mathbb{E}_{\Theta_\mu,\Theta_\Sigma}\left[\boldsymbol{\mu}_k^o\right] - \mathbf{o}_j\right) \;, \tag{D.6}$$

$$\mathbf{P}_k = \mathbb{E}_{\Theta_\mu,\Theta_\Sigma}\left[\boldsymbol{\Lambda}_k^{h,h}\right] \;. \tag{D.7}$$

Given the posteriors $q(\mathbf{Z})$ and $q(\mathbf{H})$, the following sufficient statistics of $\mathbf{X}$ can be computed

$$\boldsymbol{s}_{1k} = \begin{bmatrix} \sum_{j=1}^{N} \gamma_{jk} \mathbf{o}_j \\ \sum_{j=1}^{N} \gamma_{jk} \mathbf{n}_{jk} \end{bmatrix} \;, \tag{D.8}$$

$$\boldsymbol{S}_{2k} = \begin{bmatrix} \sum_{j=1}^{N} \gamma_{jk} \mathbf{o}_j \mathbf{o}_j^T & \sum_{j=1}^{N} \gamma_{jk} \mathbf{o}_j \mathbf{n}_{jk}^T \\ \sum_{j=1}^{N} \gamma_{jk} \mathbf{n}_{jk} \mathbf{o}_j^T & \sum_{j=1}^{N} \gamma_{jk} \left(\mathbf{n}_k \mathbf{n}_{jk}^T + (\mathbf{P}_k)^{-1}\right) \end{bmatrix} \;. \tag{D.9}$$

Once such sufficient statistics have been computed, they can be used to update the Gaussian-Wishart posteriors $q(\Theta_\mu, \Theta_\Sigma)$ in the exact same way as in equation (5.32).

Such posteriors are in turn used to compute the expectations that appear in equations (D.6) and (D.7).

Appendix E

Brain parcellation protocol adopted by Neuromorphometrics, Inc.

The following table reports a list of anatomical labels used for brain parcellation by Neuromorphometrics (http://www.neuromorphometrics.com/). Ground truth labels generated according to such a protocol are used in Chapter 4, Chapter 5 and Chapter 6 for both training and validation purposes. For each label the table reports the corresponding tissue class (WM for with matter, sGM for subcortical gray matter, cGM for cortical gray matter) and the average volume attained on a subset of the OASIS data set (http://www.oasis-brains.org).

Table E.1: Brain parcellation protocol adopted by Neuromorphometrics, Inc. Labels, tissue classes (WM for with matter, sGM for subcortical gray matter, cGM for cortical gray matter) and average volumes across thirty five subjects form the OASIS database.

Region	**Tissue class**	**Average volume** (mm^3)
Right Accumbens Area	sGM	233
Left Accumbens Area	sGM	257
Right Amygdala	sGM	603
Left Amygdala	sGM	636
Brain Stem	WM	16220
Right Caudate	sGM	2584
Left Caudate	sGM	2560
Right Cerebellum Exterior	cGM	43370
Left Cerebellum Exterior	dGM	43689

Right Cerebellum White Matter	WM	10053
Left Cerebellum White Matter	WM	10390
Right Cerebral Exterior	cGM	105
Left Cerebral Exterior	cGM	109
Right Cerebral White Matter	WM	174708
Left Cerebral White Matter	WM	170141
Right Hippocampus	sGM	2524
Left Hippocampus	sGM	2494
Right Pallidum	sGM	1133
Left Pallidum	sGM	1054
Right Putamen	sGM	3457
Left Putamen	sGM	3652
Right Thalamus Proper	sGM	6779
Left Thalamus Proper	sGM	7182
Right Ventral DC	sGM	3621
Left Ventral DC	sGM	3824
Cerebellar Vermal Lobules I-V	cGM	2746
Cerebellar Vermal Lobules VI-VII	cGM	1196
Cerebellar Vermal Lobules VIII-X	cGM	1931
Left Basal Forebrain	sGM	182
Right Basal Forebrain	sGM	181
Right ACgG anterior cingulate gyrus	cGM	2674
Left ACgG anterior cingulate gyrus	cGM	3667
Right AIns anterior insula	cGM	2785
Left AIns anterior insula	cGM	3009
Right AOrG anterior orbital gyrus	cGM	1121
Left AOrG anterior orbital gyrus	cGM	1237
Right AnG angular gyrus	cGM	6576
Left AnG angular gyrus	cGM	6420
Right Calc calcarine cortex	cGM	1782
Left Calc calcarine cortex	cGM	1936
Right CO central operculum	cGM	2464

Left CO central operculum	cGM	2225
Right Cun cuneus	cGM	2787
Left Cun cuneus	cGM	2567
Right Ent entorhinal area	cGM	1037
Left Ent entorhinal area	cGM	1024
Right FO frontal operculum	cGM	1057
Left FO frontal operculum	cGM	1033
Right FRP frontal pole	cGM	2605
Left FRP frontal pole	cGM	1951
Right FuG fusiform gyrus	cGM	4508
Left FuG fusiform gyrus	cGM	4535
Right GRe gyrus rectus	cGM	1325
Left GRe gyrus rectus	cGM	1432
Right IOG inferior occipital gyrus	cGM	4201
Left IOG inferior occipital gyrus	cGM	3966
Right ITG inferior temporal gyrus	cGM	8132
Left ITG inferior temporal gyrus	cGM	7824
Right LiG lingual gyrus	cGM	4114
Left LiG lingual gyrus	cGM	3997
Right LOrG lateral orbital gyrus	cGM	1417
Left LOrG lateral orbital gyrus	cGM	1580
Right MCgG middle cingulate gyrus	cGM	2829
Left MCgG middle cingulate gyrus	cGM	3052
Right MFC medial frontal cortex	cGM	1187
Left MFC medial frontal cortex	cGM	1285
Right MFG middle frontal gyrus	cGM	12396
Left MFG middle frontal gyrus	cGM	13346
Right MOG middle occipital gyrus	cGM	3757
Left MOG middle occipital gyrus	cGM	4375
Right MOrG medial orbital gyrus	cGM	2399
Left MOrG medial orbital gyrus	cGM	3007
Right MPoG postcentral gyrus	cGM	430

Left MPoG postcentral gyrus	cGM	483
Right MPrG precentral gyrus	cGM	1435
Left MPrG precentral gyrus	cGM	1519
Right MSFG superior frontal gyrus	cGM	4833
Left MSFG superior frontal gyrus	cGM	4481
Right MTG middle temporal gyrus	cGM	9994
Left MTG middle temporal gyrus	cGM	9502
Right OCP occipital pole	cGM	2366
Left OCP occipital pole	cGM	1987
Right OFuG occipital fusiform gyrus	cGM	2658
Left OFuG occipital fusiform gyrus	cGM	2476
Right OpIFG opercular inferior frontal gyrus	cGM	1842
Left OpIFG opercular inferior frontal gyrus	cGM	1649
Right OrIFG orbital inferior frontal gyrus	cGM	794
Left OrIFG orbital inferior frontal gyrus	cGM	761
Right PCgG posterior cingulate gyrus	cGM	2215
Left PCgG posterior cingulate gyrus	cGM	2691
Right PCu precuneus	cGM	6010
Left PCu precuneus	cGM	6377
Right PHG parahippocampal gyrus	cGM	1392
Left PHG parahippocampal gyrus	cGM	1589
Right PIns posterior insula	cGM	1452
Left PIns posterior insula	cGM	1401
Right PO parietal operculum	cGM	1170
Left PO parietal operculum	cGM	1376
Right PoG postcentral gyrus	cGM	4890
Left PoG postcentral gyrus	cGM	5954
Right POrG posterior orbital gyrus	cGM	1498
Left POrG posterior orbital gyrus	cGM	1548
Right PP planum polare	cGM	1029
Left PP planum polare	cGM	1133
Right PrG precentral gyrus		8454

Left PrG precentral gyrus	cGM	7550
Right PT planum temporale	cGM	978
Left PT planum temporale cGM		1168
Right SCA subcallosal area	cGM	582
Left SCA subcallosal area	cGM	635
Right SFG superior frontal gyrus	cGM	9083
Left SFG superior frontal gyrus	cGM	9218
Right SMC supplementary motor cortex	cGM	3632
Left SMC supplementary motor cortex	cGM	3707
Right SMG supramarginal gyrus	cGM	5321
Left SMG supramarginal gyrus	cGM	5443
Right SOG superior occipital gyrus	cGM	2301
Left SOG superior occipital gyrus	cGM	2231
Right SPL superior parietal lobule	cGM	6351
Left SPL superior parietal lobule	cGM	6441
Right STG superior temporal gyrus	cGM	4844
Left STG superior temporal gyrus	cGM	4521
Right TMP temporal pole	cGM	5470
Left TMP temporal pole	cGM	5321
Right TrIFG triangular inferior frontal gyrus	cGM	1976
Left TrIFG triangular inferior frontal gyrus	cGM	2402
Right TTG transverse temporal gyrus	cGM	661
Left TTG transverse temporal gyrus	cGM	708

Publications

Journal Papers

Claudia Blaiotta, Patrick Freund, John Ashburner. A new probabilistic atlas of the brain and cervical cord for SPM. *(in preparation)*.

Claudia Blaiotta, Patrick Freund, Manuel Jorge Cardoso, John Ashburner. Generative diffeomorphic atlas construction from brain and spinal cord MRI data. *Neuroimage (under review)*, 2017.

Claudia Blaiotta, Manuel Jorge Cardoso, John Ashburner. Variational inference for medical image segmentation. *Computer Vision and Image Understanding*, 2016.

Patrick Grabher, **Claudia Blaiotta**, John Ashburner and Patrick Freund. Relationship between brainstem neurodegeneration and clinical impairment in traumatic spinal cord injury. *Neuroimage Clinical*, 2017.

Ferran Prados, John Ashburner, **Claudia Blaiotta**, Tom Brosch, Julio Carballido-Gamio, Manuel Jorge Cardoso, Benjamin Conrad, Esha Datta, Gergely David, Benjamin De Leener and others. Spinal cord grey matter segmentation challenge. *Neuroimage*, 2017.

Conference Proceedings

Claudia Blaiotta, Manuel Jorge Cardoso, John Ashburner. Variational inference for image segmentation. *MICCAI, Workshop on Bayesian and grAphical Models for Medical Imaging*, 2015.

Claudia Blaiotta, Patrick Freund, Armin Curt, Manuel Jorge Cardoso, John Ashburner. A probabilistic framework to learn average-shaped tissue templates and its application to spinal cord image segmentation. *Proceedings of the 24th Annual Meeting of ISMRM*, 2016.

Patrick Grabher, **Claudia Blaiotta**, Armin Curt, John Ashburner and Patrick Freund. Subcortical brainstem changes in the motor system in patients with chronic spinal cord injury revealed by quantitative MRI protocols. *Proceedings of the 24th Annual Meeting of ISMRM*, 2016.

Patrick Grabher, **Claudia Blaiotta**, Armin Curt, John Ashburner and Patrick Freund. Subcortical brainstem changes in patients with spinal cord injury using quantitative MRI protocols. *Proceedings of Organization for Human Brain Mapping Annual Meeting*, 2016.

Bibliography

Mikael Agn, Ian Law, Per Munck af Rosenschöld, and Koen Van Leemput. A generative model for segmentation of tumor and organs-at-risk for radiation therapy planning of glioblastoma patients. In *SPIE Medical Imaging*, pages 97841D–97841D. International Society for Optics and Photonics, 2016.

Mohamed N Ahmed, Sameh M Yamany, Nevin Mohamed, Aly A Farag, and Thomas Moriarty. A modified fuzzy c-means algorithm for bias field estimation and segmentation of MRI data. *IEEE Transactions on Medical Imaging*, 21(3):193–199, 2002.

Santiago Aja-Fernández and Antonio Tristán-Vega. A review on statistical noise models for magnetic resonance imaging. *LPI, ETSI Telecomunicacion, Universidad de Valladolid, Spain, Tech. Rep*, 2013.

Varghese Alex, Mohammed Safwan KP, Sai Saketh Chennamsetty, and Ganapathy Krishnamurthi. Generative adversarial networks for brain lesion detection. In *SPIE Medical Imaging*, pages 101330G–101330G. International Society for Optics and Photonics, 2017.

Paul Aljabar, R Heckemann, Alexander Hammers, Joseph V Hajnal, and Daniel Rueckert. Classifier selection strategies for label fusion using large atlas databases. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI 2007*, pages 523–531. Springer, 2007.

Paul Aljabar, Rolf A Heckemann, Alexander Hammers, Joseph V Hajnal, and Daniel Rueckert. Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy. *Neuroimage*, 46(3):726–738, 2009.

Stéphanie Allassonnière and Estelle Kuhn. Stochastic algorithm for Bayesian mixture effect template estimation. *ESAIM: Probability and Statistics*, 14:382–408, 2010.

Stéphanie Allassonnière, Alain Trouvé, and Laurent Younes. Geodesic shooting and diffeomorphic matching via textured meshes. In *International Workshop on Energy minimisation Methods in Computer Vision and Pattern Recognition*, pages 365–381. Springer, 2005.

Stéphanie Allassonnière, E Kuhn, Alain Trouve, and Yali Amit. Generative model and consistent estimation algorithms for non-rigid deformable models. In *IEEE international conference on Acoustics, Speech and Signal Processing, 2006*, volume 5, pages V–V. IEEE, 2006.

Stéphanie Allassonnière, Yali Amit, and Alain Trouvé. Towards a coherent statistical framework for dense deformable template estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(1):3–29, 2007.

Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to MCMC for machine learning. *Machine learning*, 50(1-2):5–43, 2003.

J Ashburner and K Friston. Multimodal image coregistration and partitioninga unified framework. *Neuroimage*, 6(3):209–217, 1997.

John Ashburner. A fast diffeomorphic image registration algorithm. *Neuroimage*, 38 (1):95–113, 2007.

John Ashburner and Karl J Friston. Voxel-based morphometry: The methods. *Neuroimage*, 11(6):805–821, 2000.

John Ashburner and Karl J Friston. Unified segmentation. *Neuroimage*, 26(3):839–851, 2005.

John Ashburner and Karl J Friston. Computing average shaped tissue probability templates. *Neuroimage*, 45(2):333–341, 2009.

John Ashburner and Karl J. Friston. Diffeomorphic registration using geodesic shooting and Gauss–Newton optimisation. *Neuroimage*, 55(3):954–967, 2011.

John Ashburner and Gerard R. Ridgway. Symmetric diffeomorphic modeling of longitudinal structural MRI. *Frontiers in Neuroscience*, 6:197, 2013.

John Ashburner, Karl J Friston, et al. Nonlinear spatial normalization using basis functions. *Human Brain Mapping*, 7(4):254–266, 1999.

John Ashburner, John G Csernansk, Christos Davatzikos, Nick C Fox, Giovanni B Frisoni, and Paul M Thompson. Computer-assisted imaging to assess brain structure in healthy and diseased brains. *The Lancet Neurology*, 2(2):79–88, 2003.

Hagai Attias. Inferring parameters and structure of latent variable models by variational Bayes. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, pages 21–30. Morgan Kaufmann Publishers Inc., 1999.

Berengere Aubert-Broche, Mark Griffin, G Bruce Pike, Alan C Evans, and D Louis Collins. Twenty new digital brain phantoms for creation of validation image data bases. *IEEE transactions on medical imaging*, 25(11):1410–1416, 2006.

Brian Avants and James C Gee. Geodesic estimation for large deformation anatomical shape averaging and interpolation. *Neuroimage*, 23:S139–S150, 2004.

Brian B Avants, P Thomas Schoenemann, and James C Gee. Lagrangian frame diffeomorphic image registration: Morphometric comparison of human and chimpanzee cortex. *Medical Image Analysis*, 10(3):397–412, 2006.

Brian B Avants, Charles L Epstein, Murray Grossman, and James C Gee. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, 12(1):26–41, 2008.

Brian B Avants, Paul Yushkevich, John Pluta, David Minkoff, Marc Korczykowski, John Detre, and James C Gee. The optimal template effect in hippocampus studies of diseased populations. *Neuroimage*, 49(3):2457–2466, 2010.

Brian B Avants, Nicholas J Tustison, Gang Song, Philip A Cook, Arno Klein, and James C Gee. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *Neuroimage*, 54(3):2033–2044, 2011a.

Brian B Avants, Nicholas J Tustison, Jue Wu, Philip A Cook, and James C Gee. An open source multivariate framework for n-tissue segmentation with evaluation on public data. *Neuroinformatics*, 9(4):381–400, 2011b.

Ruzena Bajcsy, Robert Lieberson, and Martin Reivich. A computerised system for the elastic matching of deformed radiographic images to idealized atlas images. *Journal of Computer Assisted Tomography*, 7(4):618–625, 1983.

Rohit Bakshi, Venkata SR Dandamudi, Mohit Neema, Chitradeep De, and Robert A Bermel. Measurement of brain and spinal cord atrophy by magnetic resonance imaging as a tool to monitor multiple sclerosis. *Journal of Neuroimaging*, 15(s4):30S–45S, 2005.

Serdar K Balci, Polina Golland, Martha Shenton, and William M Wells. Free-form B-spline deformation model for groupwise registration. In *Proc. International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI 2007, Statistical Registration Workshop: Pair-wise and Group-wise Alignment and Atlas Formation*, volume 10, pages 23–30, 2007.

Nematollah K Batmanghelich, Ben Taskar, and Christos Davatzikos. Generative-discriminative basis learning for medical imaging. *IEEE Transactions on Medical Imaging*, 31(1):51–69, 2012.

M Faisal Beg, Michael I Miller, Alain Trouvé, and Laurent Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International Journal of Computer Vision*, 61(2):139–157, 2005.

Mirza Faisal Beg and Ali Khan. Computing an average anatomical atlas using LDDMM and geodesic shooting. In *3rd IEEE International Symposium on Biomedical Imaging: Nano to Macro, 2006.*, pages 1116–1119. IEEE, 2006.

Serge Belongie, Chad Carson, Hayit Greenspan, and Jitendra Malik. Color-and texture-based image segmentation using EM and its application to content-based image retrieval. In *6th International Conference on Computer Vision, 1998*, pages 675–682. IEEE, 1998.

Dimitri P Bertsekas. *Nonlinear Programming.* Athena scientific Belmont, 1999.

James C Bezdek, LO Hall, and L_P Clarke. Review of MR image segmentation techniques using pattern recognition. *Medical Physics*, 20(4):1033–1048, 1992.

K K Bhatia, J V Hajnal, B K Puri, A D Edwards, and D Rueckert. Consistent groupwise non-rigid registration for atlas construction. In *IEEE International Symposium on Biomedical Imaging: Nano to Macro, 2004*, volume 1, pages 908–911. IEEE, April 2004.

Kanwal K Bhatia, Paul Aljabar, James P Boardman, Latha Srinivasan, Maria Murgasova, Serena J Counsell, Mary A Rutherford, Joseph V Hajnal, A David Edwards, and Daniel Rueckert. Groupwise combined segmentation and registration for atlas construction. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI 2007*, pages 532–540. Springer, 2007.

Christophe Biernacki and Stéphane Chrétien. Degeneracy in the maximum likelihood estimation of univariate gaussian mixtures with EM. *Statistics & probability letters*, 61(4):373–382, 2003.

Christophe Biernacki, Gilles Celeux, and Gérard Govaert. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics & Data Analysis*, 41(3):561–575, 2003.

Christopher M Bishop. *Neural Networks for Pattern Recognition.* Oxford university press, 1995.

Christopher M Bishop. *Pattern Recognition and Machine Learning*, volume 1. Springer New York, 2006.

Christopher M Bishop, Julia Lasserre, et al. Generative or discriminative? Getting the best of both worlds. *Bayesian Statistics*, 8:3–24, 2007.

Ronald Boellaard. Standards for PET image acquisition and quantitative data analysis. *Journal of Nuclear Medicine*, 50(Suppl 1):11S–20S, 2009.

Fred L Bookstein. Shape and the information in medical images: A decade of the morphometric synthesis. In *Mathematical Methods in BioMedical Image Analysis, 1996., Proceedings of the Workshop on*, pages 2–12. IEEE, 1996.

Anna Bosch, Andrew Zisserman, and Xavier Muñoz. Scene classification using a hybrid generative/discriminative approach. *IEEE transactions on Pattern Analysis and machine intelligence*, 30(4):712–727, 2008.

Joost Bot, Frederik Barkhof, Chris Polman, Lycklama à Nijeholt, Corline De Groot, Elisabeth Bergers, Herman Ader, and Jonas Castelijns. Spinal cord abnormalities in recently diagnosed MS patients added value of spinal MRI examination. *Neurology*, 62(2):226–233, 2004.

Guillaume Bouchard and Bill Triggs. The tradeoff between generative and discriminative classifiers. In *16th IASC International Symposium on Computational Statistics (COMPSTAT'04)*, pages 721–728, 2004.

Anton E Bowden, Richard D Rabbitt, and Jeffrey A Weiss. Anatomical registration and segmentation by warping template finite element models. In *BiOS'98 International Biomedical Optics Symposium*, pages 469–476. International Society for Optics and Photonics, 1998.

Hamparsum Bozdogan. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370, 1987.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Stephanie Bricq, Ch Collet, and Jean-Paul Armspach. Unifying framework for multimodal brain MRI segmentation based on Hidden Markov Chains. *Medical Image Analysis*, 12(6):639–652, 2008.

Benjamin H Brinkmann, Armando Manduca, and Richard A Robb. Optimised homomorphic unsharp masking for MR grayscale inhomogeneity correction. *IEEE Transactions on Medical Imaging*, 17(2):161–171, 1998.

Tom Brosch, Lisa YW Tang, Youngjin Yoo, David KB Li, Anthony Traboulsee, and Roger Tam. Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. *IEEE transactions on medical imaging*, 35(5):1229–1239, 2016.

Martin Burger, Jan Modersitzki, and Lars Ruthotto. A hyperelastic regularization energy for image registration. *SIAM Journal on Scientific Computing*, 35(1):B132–B148, 2013.

Ninon Burgos, Manuel Jorge Cardoso, Marc Modat, Stefano Pedemonte, John Dickson, Anna Barnes, John S Duncan, David Atkinson, Simon R Arridge, Brian F Hutton, et al. Attenuation correction synthesis for hybrid PET-MR scanners. In *Proc. International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI 2013*, pages 147–154. Springer, 2013.

Ninon Burgos, M Jorge Cardoso, Kris Thielemans, Marc Modat, Stefano Pedemonte, John Dickson, Anna Barnes, Rebekah Ahmed, Colin J Mahoney, Jonathan M Schott, et al. Attenuation correction synthesis for hybrid PET-MR scanners: Application to brain studies. *IEEE Transactions on Medical Imaging*, 33(12):2332–2341, 2014.

Mariano Cabezas, Arnau Oliver, Xavier Lladó, Jordi Freixenet, and Meritxell Bach Cuadra. A review of atlas-based segmentation for magnetic resonance brain images. *Computer methods and programs in biomedicine*, 104(3):e158–e177, 2011.

Pascal Cachier and David Rey. Symmetrization of the non-rigid registration problem using inversion-invariant energies: Application to multiple sclerosis. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI 2000*, pages 472–481. Springer, 2000.

Sergio Campos, Luis Pizarro, Carlos Valle, Katherine R Gray, Daniel Rueckert, and Héctor Allende. Evaluating imputation techniques for missing data in ADNI: A patient classification study. In *Iberoamerican Congress on Pattern Recognition*, pages 3–10. Springer International Publishing, 2015.

M Jorge Cardoso, Matthew J Clarkson, Gerard R Ridgway, Marc Modat, Nick C Fox, Sebastien Ourselin, Alzheimer's Disease Neuroimaging Initiative, et al. LoAd: A locally adaptive cortical segmentation algorithm. *Neuroimage*, 56(3):1386–1397, 2011.

M Jorge Cardoso, Carole H Sudre, Marc Modat, and Sebastien Ourselin. Template-based multimodal joint generative model of brain data. In *Information Processing in Medical Imaging*, pages 17–29. Springer, 2015.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1721–1730. ACM, 2015.

Gilles Celeux and Gilda Soromenho. An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, 13(2):195–212, 1996.

Olivier Chapelle, Bernhard Schölkopf, Alexander Zien, et al. *Semi-supervised learning*. MIT press Cambridge, 2006.

Min Chen, Aaron Carass, Jiwon Oh, Govind Nair, Dzung L. Pham, Daniel S. Reich, and Jerry L Prince. Automatic magnetic resonance spinal cord segmentation with topology constraints for variable fields of view. *Neuroimage*, 83:1051–1062, 2013.

Gary Christensen. Consistent linear-elastic transformations for image matching. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 224–237. Springer, 1999.

Gary E Christensen and Hans J Johnson. Consistent image registration. *IEEE Transactions on Medical Imaging*, 20(7):568–582, 2001.

Gary E Christensen, Richard D Rabbitt, and Michael I Miller. Deformable templates using large deformation kinematics. *IEEE Transactions on Image Processing*, 5(10): 1435–1447, 1996.

Keh-Shih Chuang, Hong-Long Tzeng, Sharon Chen, Jay Wu, and Tzong-Jer Chen. Fuzzy c-means clustering with spatial information for image segmentation. *Computerized Medical Imaging and Graphics*, 30(1):9–15, 2006.

Haili Chui, Lawrence Win, Robert Schultz, James Duncan, and Anand Rangarajan. A unified feature registration method for brain mapping. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 300–314. Springer, 2001.

Se Young Chun and Jeffrey A Fessler. Regularized methods for topology-preserving smooth nonrigid image registration using B-spline basis. In *5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2008*, pages 1099–1102. IEEE, 2008.

Delia Ciardo, M Peroni, M Riboldi, Daniela Alterio, G Baroni, and Roberto Orecchia. The role of regularization in deformable image registration for head and neck adaptive radiotherapy. *Technology in cancer research & treatment*, 12(4):323–331, 2013.

Dan Ciresan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. In *Advances in Neural Information Processing Systems*, pages 2843–2851, 2012.

Chris A Cocosco, Vasken Kollokian, Remi K-S Kwan, G Bruce Pike, and Alan C Evans. Brainweb: Online interface to a 3D MRI simulated brain database. In *Neuroimage*. Citeseer, 1997.

J Cohen-Adad, A Mareyam, B Keil, JR Polimeni, and LL Wald. 32-channel RF coil optimised for brain and cervical spinal cord at 3T. *Magnetic Resonance in Medicine*, 66(4):1198–1208, 2011.

D. Louis Collins and Alan C. Evans. Animal: Validation and applications of nonlinear registration-based segmentation. *International Journal of Pattern recognition and Artificial Intelligence*, 11(08):1271–1294, 1997.

Louis Collins, Alex Zijdenbos, Vasken Kollokian, John Sled, Noor Kabani, Colin J Holmes, and Alan C Evans. Design and construction of a realistic digital brain phantom. *IEEE Transactions on Medical Imaging*, 17(3):463–468, 1998.

Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

Timothy F Cootes, Stephen Marsland, Carole J Twining, Kate Smith, and Christopher J Taylor. Groupwise diffeomorphic non-rigid registration for automatic model building. In *European conference on computer vision*, pages 316–327. Springer, 2004.

Nicolas Cordier, Hervé Delingette, Matthieu Lê, and Nicholas Ayache. Extended modality propagation: Image synthesis of pathological cases. *IEEE Transactions on Medical Imaging*, page 11, 2016.

Adrian Corduneanu and Christopher M Bishop. Variational Bayesian model selection for mixture distributions. In *Artificial intelligence and Statistics*, volume 2001, pages 27–34. Morgan Kaufmann Waltham, MA, 2001.

Olivier Coulon, Simon Hickman, Geoff Parker, Gareth Barker, David Miller, and Simon Arridge. Quantification of spinal cord atrophy from magnetic resonance images via a B-spline active surface model. *Magnetic resonance in medicine*, 47(6):1176–1185, 2002.

Robert W Cox. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical research*, 29(3):162–173, 1996.

George R Cross and Anil K Jain. Markov random field texture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1:25–39, 1983.

William R. Crum, Oscar Camara, and Derek L.G. Hill. Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE Transactions on Medical Imaging*, 25(11):1451–1461, 2006.

William R Crum, Thomas Hartkens, and DLG Hill. Non-rigid image registration: Theory and practice. *The British Journal of Radiology*, 2014.

WR Crum, LD Griffin, DLG Hill, and DJ Hawkes. Zen and the art of medical image registration: correspondence, homology, and quality. *NeuroImage*, 20(3):1425–1437, 2003.

Meritxell Bach Cuadra, Claudio Pollo, Anton Bardera, Olivier Cuisenaire, J-G Villemure, and J-P Thiran. Atlas-based segmentation of pathological MR brain images using a model of lesion growth. *IEEE Transactions on Medical Imaging*, 23(10): 1301–1314, 2004.

Adelino R Ferreira da Silva. Bayesian mixture models of variable dimension for image segmentation. *Computer Methods and Programs in Biomedicine*, 94(1):1–14, 2009.

AR Davies and RS Anderssen. Optimisation in the regularisation ill-posed problems. *The Journal of the Australian Mathematical Society. Series B. Applied Mathematics*, 28(01):114–133, 1986.

Benoit M Dawant, Alex P Zijdenbos, and Richard A Margolin. Correction of intensity variations in MR images for computer-aided tissue classification. *IEEE Transactions on Medical Imaging*, 12(4):770–781, 1993.

Mathieu De Craene, Aloys Du Bois Daische, Benot Macq, and Simon K. Warfield. Multi-subject registration for unbiased statistical atlas construction. In Pierre Hellier

Christian Barillot, David R. Haynor, editor, *Proc. International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI 2004*, volume 3216 of *LNCS*, pages 655–662, Berlin, 2004. Springer.

Benjamin De Leener, Samuel Kadoury, and Julien Cohen-Adad. Robust, accurate and fast automatic segmentation of the spinal cord. *Neuroimage*, 98:528–536, 2014.

Benjamin De Leener, Simon Lévy, Sara M Dupont, Vladimir S Fonov, Nikola Stikov, D Louis Collins, Virginie Callot, and Julien Cohen-Adad. SCT: Spinal Cord Toolbox, an open-source software for processing spinal cord MRI data. *Neuroimage*, 2016.

Benjamin De Leener, Simon Lévy, Sara M Dupont, Vladimir S Fonov, Nikola Stikov, D Louis Collins, Virginie Callot, and Julien Cohen-Adad. SCT: Spinal Cord Toolbox, an open-source software for processing spinal cord MRI data. *NeuroImage*, 145:24–43, 2017.

Yves Delignon, Abdelwaheb Marzouki, and Wojciech Pieczynski. Estimation of generalized mixtures and its application in image segmentation. *IEEE Transactions on Image Processing*, 6(10):1364–1375, 1997.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, pages 1–38, 1977.

Erika RE Denton, Luke I Sonoda, Daniel Rueckert, Sheila C Rankin, Carmel Hayes, Martin O Leach, Derek LG Hill, and David J Hawkes. Comparison and evaluation of rigid, affine, and nonrigid registration of breast MR images. *Journal of Computer Assisted Tomography*, 23(5):800–805, 1999.

Edgar A DeYoe, Peter Bandettini, Jay Neitz, David Miller, and Paula Winans. Functional magnetic resonance imaging (FMRI) of the human brain. *Journal of Neuroscience Methods*, 54(2):171–187, 1994.

Jean Dieudonné. *Foundations of Modern Analysis.* Read Books Ltd, 2013.

David Draper. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 45–97, 1995.

Guillaume Dugas-Phocion, Miguel Angel González Ballester, Grégoire Malandain, Christine Lebrun, and Nicholas Ayache. Improved EM-based tissue segmentation and partial volume effect quantification in multi-sequence brain MRI. In *Proc. International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI 2004*, pages 26–33. Springer, 2004.

Emiliano DAgostino, Frederik Maes, Dirk Vandermeulen, and Paul Suetens. A unified framework for atlas based brain image segmentation and registration. In *Biomedical Image Registration*, pages 136–143. Springer, 2006.

Simon B Eickhoff, Klaas E Stephan, Hartmut Mohlberg, Christian Grefkes, Gereon R Fink, Katrin Amunts, and Karl Zilles. A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *Neuroimage*, 25(4):1325–1335, 2005.

Markus Enzweiler and Dariu M Gavrila. A mixed generative-discriminative framework for pedestrian classification. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

James E Falk. Lagrange multipliers and nonlinear programming. *Journal of Mathematical Analysis and Applications*, 19(1):141–159, 1967.

Ayres Fan, William M Wells, John W Fisher, Müjdat Cetin, Steven Haker, Robert Mulkern, Clare Tempany, and Alan S Willsky. A unified variational approach to denoising and bias correction in MR. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 148–159. Springer, 2003.

Ayres C Fan, John W Fisher III, William M Wells III, James J Levitt, and Alan S Willsky. MCMC curve sampling for image segmentation. In *Proc. International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI 2007*, pages 477–485. Springer, 2007.

Lingzhong Fan, Hai Li, Junjie Zhuo, Yu Zhang, Jiaojian Wang, Liangfu Chen, Zhengyi Yang, Congying Chu, Sangma Xie, Angela R Laird, et al. The human brainnetome atlas: A new brain atlas based on connectional architecture. *Cerebral Cortex*, page 157, 2016.

Roman Filipovych, Christos Davatzikos, Alzheimer's Disease Neuroimaging Initiative, et al. Semi-supervised pattern classification of medical images: Application to mild cognitive impairment (MCI). *Neuroimage*, 55(3):1109–1119, 2011.

Massimo Filippi, Adriana Campi, Bruno Colombo, Clodoaldo Pereira, Vittorio Martinelli, Corrado Baratti, and Giancarlo Comi. A spinal cord MRI study of benign and secondary progressive multiple sclerosis. *Journal of Neurology*, 243(7):502–505, 1996.

Bernd Fischer and Jan Modersitzki. Ill-posed medicinean introduction to image registration. *Inverse Problems*, 24(3):034008, 2008.

Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.

Bruce Fischl, David H Salat, Evelina Busa, Marilyn Albert, Megan Dieterich, Christian Haselgrove, Andre Van Der Kouwe, Ron Killiany, David Kennedy, Shuna Klaveness, et al. Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33(3):341–355, 2002.

Bruce Fischl, André van der Kouwe, Christophe Destrieux, Eric Halgren, Florent Ségonne, David H Salat, Evelina Busa, Larry J Seidman, Jill Goldstein, David Kennedy, et al. Automatically parcellating the human cerebral cortex. *Cerebral Cortex*, 14(1):11–22, 2004.

P Thomas Fletcher, Conglin Lu, Stephen M Pizer, and Sarang Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Transactions on Medical Imaging*, 23(8):995–1005, 2004.

Vladimir S Fonov, Arnaud Le Troter, Manuel Taso, Benjamin De Leener, G Lévêque, Matilde Benhamou, Michael Sdika, Habib Benali, Pierre-Francois Pradat, Louis Collins, et al. Framework for integrated Mri average of the spinal cord white and gray matter: The MNI–Poly–AMU template. *Neuroimage*, 102:817–827, 2014.

Oren Freifeld, Hayit Greenspan, and Jacob Goldberger. Lesion detection in noisy MR brain images using constrained GMM and active contours. In *4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2007*, pages 596–599. IEEE, 2007.

Patrick Freund, Armin Curt, Karl Friston, and Alan Thompson. Tracking changes following spinal cord injury insights from neuroimaging. *The Neuroscientist*, 19(2): 116–128, 2013a.

Patrick Freund, Nikolaus Weiskopf, John Ashburner, Katharina Wolf, Reto Sutter, Daniel R Altmann, Karl Friston, Alan Thompson, and Armin Curt. MRI investigation of the sensorimotor cortex and the corticospinal tract after acute spinal cord injury: A prospective longitudinal study. *The Lancet Neurology*, 12(9):873–881, 2013b.

Patrick Freund, Karl Friston, Alan J Thompson, Klaas E Stephan, John Ashburner, Dominik R Bach, Zoltan Nagy, Gunther Helms, Bogdan Draganski, Siawoosh Mohammadi, et al. Embodied neurology: An integrative framework for neurological disorders. *Brain*, page aww076, 2016.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

Nir Friedman and Stuart Russell. Image segmentation in video sequences: A probabilistic approach. In *Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence*, pages 175–181. Morgan Kaufmann Publishers Inc., 1997.

Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010.

Karl Friston, J.ohn Ashburner, Chris Frith, Jean-Baprtiste Poline, J. D. Heather, and Richard Frackowiak. Spatial registration and normalisation of images. *Human Brain Mapping*, 3(3):165–189, 1995.

Watson Fulks and JO Sather. Asymptotics. II. Laplaces method for multiple integrals. *Pacific Journal of Mathematics*, 11(1):185–192, 1961.

Glenn Fung and Jonathan Stoeckel. SVM feature selection for classification of SPECT images of Alzheimer's disease using spatial information. *Knowledge and Information Systems*, 11(2):243–258, 2007.

Marco Ganzetti, Nicole Wenderoth, and Dante Mantini. Quantitative evaluation of intensity inhomogeneity correction methods for structural MR brain images. *Neuroinformatics*, 14(1):5–21, 2016.

S Geisser, J Hodges, S Press, and A ZeUner. The validity of posterior expansions based on Laplaces method. *Bayesian and Likelihood Methods in Statistics and Econometrics*, 7:473, 1990.

Guido Gerig, John Martin, Ron Kikinis, Olaf Kubler, Martha Shenton, and Ferenc A Jolesz. Unsupervised tissue type segmentation of 3D dual-echo MR head data. *Image and vision computing*, 10(6):349–360, 1992.

David T Gering, Arya Nabavi, Ron Kikinis, W Eric L Grimson, Noby Hata, Peter Everett, Ferenc Jolesz, and William M Wells. An integrated visualization system for surgical planning and guidance using image fusion and interventional imaging. In *Proc. International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI 1999*, pages 809–819. Springer, 1999.

Zoubin Ghahramani. Unsupervised learning. In *Advanced Lectures on Machine learning*, pages 72–112. Springer, 2004.

Zoubin Ghahramani. Bayesian non-parametrics and the probabilistic approach to modelling. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 371(1984):20110553, 2013.

Satrajit S Ghosh, Sita Kakunoori, Jean Augustinack, Alfonso Nieto-Castanon, Ioulia Kovelman, Nadine Gaab, Joanna A Christodoulou, Christina Triantafyllou, John DE Gabrieli, and Bruce Fischl. Evaluating the validity of volume-based and surface-based brain image registration for developmental cognitive neuroscience studies in children 4 to 11 years of age. *Neuroimage*, 53(1):85–93, 2010.

Antonio Giorgio and Nicola De Stefano. Clinical use of brain volumetry. *Journal of Magnetic Resonance Imaging*, 37(1):1–14, 2013.

Ingrid K Glad and Giovanni Sebastiani. A Bayesian approach to synthetic magnetic resonance imaging. *Biometrika*, 82(2):237–250, 1995.

William L Goffe, Gary D Ferrier, and John Rogers. Global optimisation of statistical functions with simulated annealing. *Journal of econometrics*, 60(1-2):65–99, 1994.

Sally Goldman and Yan Zhou. Enhancing supervised learning with unlabeled data. In *International Conference on Machine Learning, 2000*, pages 327–334. Citeseer, 2000.

Ali Gooya, Kilian M Pohl, Michel Bilello, Luigi Cirillo, George Biros, Elias R Melhem, and Christos Davatzikos. Glistr: Glioma image segmentation and registration. *IEEE Transactions on Medical Imaging*, 31(10):1941–1954, 2012.

Patrick Grabher, Martina F Callaghan, John Ashburner, Nikolaus Weiskopf, Alan J Thompson, Armin Curt, and Patrick Freund. Tracking sensory system atrophy and outcome prediction in spinal cord injury. *Annals of neurology*, 78(5):751–761, 2015.

Henry Gray. *Gray's Anatomy: With original illustrations by Henry Carter*. Arcturus Publishing, 2009.

Hayit Greenspan, Amit Ruf, and Jacob Goldberger. Constrained Gaussian mixture model framework for automatic segmentation of MR brain images. *IEEE Transactions on Medical Imaging*, 25(9):1233–1245, 2006.

Ed Gronenschild, Petra Habets, Heidi Jacobs, Ron Mengelers, Nico Rozendaal, Jim Van Os, and Machteld Marcelis. The effects of FreeSurfer version, workstation type, and Macintosh operating system version on anatomical volume and cortical thickness measurements. *PloS one*, 7(6):e38234, 2012.

Robert I Grossman, Fredrick Barkhof, and Massimo Filippi. Assessment of spinal cord damage in MS using MRI. *Journal of the Neurological Sciences*, 172:S36–S39, 2000.

Hákon Gudbjartsson and Samuel Patz. The Rician distribution of noisy MRI data. *Magnetic resonance in medicine*, 34(6):910–914, 1995.

Madhu Sudhan Reddy Gudur, Wendy Hara, Quynh-Thu Le, Lei Wang, Lei Xing, and Ruijiang Li. A unifying probabilistic Bayesian approach to derive electron density from MRI for radiation therapy treatment planning. *Physics in medicine and Biology*, 59(21):6595, 2014.

Régis Guillemaud. Uniformity correction with homomorphic filtering on region of interest. In *1998 International Conference on Image Processing, 1998*, volume 2, pages 872–875. IEEE, 1998.

Régis Guillemaud and Michael Brady. Estimating the bias field of MR images. *IEEE Transactions on Medical Imaging*, 16(3):238–251, 1997.

Alexandre Guimond, Jean Meunier, and Jean-Philippe Thirion. Average brain models: A convergence study. *Computer Vision and Image Understanding*, 77(2):192–210, 2000.

Dazhou Guo, Julius Fridriksson, Paul Fillmore, Christopher Rorden, Hongkai Yu, Kang Zheng, and Song Wang. Automated lesion detection on MRI scans using combined unsupervised and supervised methods. *BMC Medical Imaging*, 15(1):50, 2015.

Lalit Gupta and Thotsapon Sortrakul. A Gaussian-mixture-based image segmentation algorithm. *Pattern Recognition*, 31(3):315–325, 1998.

John Haselgrove and Manfred Prammer. An algorithm for compensation of surface-coil images for sensitivity of the surface coil. *Magnetic Resonance Imaging*, 4(6):469–472, 1986.

Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical Image Analysis*, 35:18–31, 2017.

Lei He, Zhigang Peng, Bryan Everding, Xun Wang, Chia Y Han, Kenneth L Weiss, and William G Wee. A comparative study of deformable contour methods on medical image segmentation. *Image and Vision Computing*, 26(2):141–163, 2008.

Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4): 18–28, 1998.

Rolf A Heckemann, Joseph V Hajnal, Paul Aljabar, Daniel Rueckert, and Alexander Hammers. Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. *NeuroImage*, 33(1):115–126, 2006.

Mattias P Heinrich, Mark Jenkinson, Manav Bhushan, Tahreema Matin, Fergus V Gleeson, Michael Brady, and Julia A Schnabel. MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration. *Medical Image Analysis*, 16 (7):1423–1435, 2012.

Karsten Held, Elena Rota Kops, Bernd J Krause, William M Wells III, Ron Kikinis, and Hans-Wilhelm Muller-Gartner. Markov random field segmentation of brain MR images. *IEEE Transactions on Medical Imaging*, 16(6):878–886, 1997.

Katherine A Heller, Sinead Williamson, and Zoubin Ghahramani. Statistical models for partial membership. In *Proceedings of the 25th international conference on Machine learning*, pages 392–399. ACM, 2008.

Gunther Helms, Henning Dathe, and Peter Dechent. Quantitative FLASH MRI at 3t using a rational approximation of the Ernst equation. *Magnetic Resonance in Medicine*, 59(3):667–672, 2008.

Monica Hernandez and Salvador Olmos. Gauss-Newton optimisation in diffeomorphic registration. In *5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2008*, pages 1083–1086. IEEE, 2008.

Derek LG Hill, Philipp G Batchelor, Mark Holden, and David J Hawkes. Medical image registration. *Physics in medicine and biology*, 46(3):R1, 2001.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Steven CH Hoi, Rong Jin, Jianke Zhu, and Michael R Lyu. Semisupervised SVM batch mode active learning with applications to image retrieval. *ACM Transactions on Information Systems (TOIS)*, 27(3):16, 2009.

Mark A Horsfield, Stefania Sala, Mohit Neema, Martina Absinta, Anshika Bakshi, Maria Pia Sormani, Maria A Rocca, Rohit Bakshi, and Massimo Filippi. Rapid semi-automatic segmentation of the spinal cord from magnetic resonance images: Application in multiple sclerosis. *Neuroimage*, 50(2):446–455, 2010.

David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

Hsieh Hou and H Andrews. Cubic splines for image interpolation and digital filtering. *IEEE Transactions on acoustics, speech, and signal processing*, 26(6):508–517, 1978.

Zujun Hou. A review on MR image intensity inhomogeneity correction. *International Journal of Biomedical Imaging*, 2006, 2006.

Eveline Huber, Armin Curt, and Patrick Freund. Tracking trauma-induced structural and functional changes above the level of spinal cord injury. *Current opinion in neurology*, 28(4):365–372, 2015.

Juan Iglesias, Ender Konukoglu, Albert Montillo, Zhuowen Tu, and Antonio Criminisi. Combining generative and discriminative models for semantic segmentation of CT scans via active learning. In *Information Processing in Medical Imaging*, pages 25–36. Springer, 2011.

Juan Eugenio Iglesias and Mert R Sabuncu. Multi-atlas segmentation of biomedical images: A survey. *Medical Image Analysis*, 24(1):205–219, 2015.

Juan Eugenio Iglesias, Mert Rory Sabuncu, and Koen Van Leemput. A generative model for multi-atlas segmentation across modalities. In *9th IEEE International Symposium on Biomedical Imaging, 2012*, pages 888–891. IEEE, 2012a.

Juan Eugenio Iglesias, Mert Rory Sabuncu, and Koen Van Leemput. Incorporating parameter uncertainty in Bayesian segmentation models: Application to hippocampal subfield volumetry. In *Proc. International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI 2012*, pages 50–57. Springer, 2012b.

Juan Eugenio Iglesias, Ender Konukoglu, Darko Zikic, Ben Glocker, Koen Van Leemput, and Bruce Fischl. Is synthesizing MRI contrast useful for inter-modality analysis? In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 631–638. Springer, 2013a.

Juan Eugenio Iglesias, Mert Rory Sabuncu, and Koen Van Leemput. A probabilistic, non-parametric framework for inter-modality label fusion. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI 2013*, pages 576–583. Springer, 2013b.

Juan Eugenio Iglesias, Mert Rory Sabuncu, Koen Van Leemput, Alzheimers Disease Neuroimaging Initiative, et al. Improved inference in Bayesian segmentation using Monte Carlo sampling: Application to hippocampal subfield volumetry. *Medical image analysis*, 17(7):766–778, 2013c.

Ivana Išgum, Manon JNL Benders, Brian Avants, M Jorge Cardoso, Serena J Counsell, Elda Fischi Gomez, Laura Gui, Petra S Hűppi, Karina J Kersbergen, Antonios Makropoulos, et al. Evaluation of automatic neonatal brain segmentation algorithms: The NeoBrains12 challenge. *Medical Image Analysis*, 20(1):135–151, 2015.

Tommi S Jaakkola, David Haussler, et al. Exploiting generative models in discriminative classifiers. *Advances in Neural Information Processing Systems*, pages 487–493, 1999.

Mark Jenkinson and Stephen Smith. A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, 5(2):143–156, 2001.

Mark Jenkinson, Peter Bannister, Michael Brady, and Stephen Smith. Improved optimisation for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, 17(2):825–841, 2002.

Tim Jerman, Alfiia Galimzianova, Franjo Pernuš, Boštjan Likar, and Žiga Špiclin. Combining unsupervised and supervised methods for lesion segmentation. In *International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 45–56. Springer, 2015.

Amod Jog, Snehashis Roy, Aaron Carass, and Jerry L Prince. Magnetic resonance image synthesis through patch regression. In *IEEE 10th International Symposium on Biomedical Imaging, 2013*, pages 350–353. IEEE, 2013.

Hans J Johnson and Gary E Christensen. Consistent landmark and intensity-based image registration. *IEEE Transactions on Medical Imaging*, 21(5):450–461, 2002.

Michael I Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. *Advances in Neural Information Processing Systems*, 14: 841, 2002.

Sarang Joshi, Brad Davis, Matthieu Jomier, and Guido Gerig. Unbiased diffeomorphic atlas construction for computational anatomy. *Neuroimage*, 23:S151–S160, 2004.

Sarang C Joshi and Michael I Miller. Landmark matching via large deformation diffeomorphisms. *IEEE Transactions on Image Processing*, 9(8):1357–1370, 2000.

Jorge Jovicich, Silvester Czanner, Xiao Han, David Salat, Andre van der Kouwe, Brian Quinn, Jenni Pacheco, Marilyn Albert, Ronald Killiany, Deborah Blacker, et al. MRI-derived measurements of human subcortical, ventricular and intracranial brain volumes: Reliability effects of scan sessions, acquisition sequences, data analyses, scanner upgrade, scanner vendors and field strengths. *Neuroimage*, 46(1):177–192, 2009.

Tina Kapur, W Eric, L Grimson, Ron Kikinis, and William M Wells. Enhanced spatial priors for segmentation of magnetic resonance imagery. In *Proc. International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI 1998*, pages 457–468. Springer, 1998.

Zoltan Kato. Segmentation of color images via reversible jump MCMC sampling. *Image and Vision Computing*, 26(3):361–371, 2008.

Ali R Khan, Lei Wang, and Mirza Faisal Beg. Freesurfer-initiated fully-automated subcortical brain segmentation in MRI using large deformation diffeomorphic metric mapping. *Neuroimage*, 41(3):735–746, 2008.

D Kidd, John Thorpe, Alan Thompson, Brian Kendall, IF Moseley, David MacManus, W Ian McDonald, and David H Miller. Spinal cord MRI using multi-array coils and fast spin echo ii. findings in multiple sclerosis. *Neurology*, 43(12):2632–2632, 1993.

Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pages 3581–3589, 2014.

Christopher Kipps, AJ Duggins, Neil Mahant, Regina LE Gomes, John Ashburner, and Elizabeth A McCusker. Progression of structural neuropathology in preclinical Huntingtons disease: A tensor based morphometry study. *Journal of Neurology, Neurosurgery & Psychiatry*, 76(5):650–655, 2005.

Frederick Klauschen, Aaron Goldman, Vincent Barra, Andreas Meyer-Lindenberg, and Arvid Lundervold. Evaluation of automated brain MR image segmentation and volumetry methods. *Human Brain Mapping*, 30(4):1310–1327, 2009.

Arno Klein and Jason Tourville. 101 labeled brain images and a consistent human cortical labeling protocol. *Frontiers in neuroscience*, 6:171, 2012.

Arno Klein, Brett Mensh, Satrajit Ghosh, Jason Tourville, and Joy Hirsch. Mindboggle: Automated brain labeling with multiple atlases. *BMC Medical Imaging*, 5(1):7, 2005.

Arno Klein, Jesper Andersson, Babak A Ardekani, John Ashburner, Brian Avants, Ming-Chang Chiang, Gary E Christensen, D Louis Collins, James Gee, Pierre Hellier, et al. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *Neuroimage*, 46(3):786–802, 2009.

Stefan Klein, Marius Staring, and Josien PW Pluim. Evaluation of optimisation methods for nonrigid medical image registration using mutual information and B-splines. *IEEE transactions on image processing*, 16(12):2879–2890, 2007.

Stefan Klein, Marius Staring, Keelin Murphy, Max A Viergever, and Josien PW Pluim. Elastix: A toolbox for intensity-based medical image registration. *IEEE Transactions on Medical Imaging*, 29(1):196–205, 2010.

Lisa M Koch, Martin Rajchl, Tong Tong, Jonathan Passerat-Palmbach, Paul Aljabar, and Daniel Rueckert. Multi-atlas segmentation as a graph labelling problem: Application to partially annotated atlas data. In *International Conference on Information Processing in Medical Imaging*, pages 221–232. Springer, 2015.

Natasha Kovacevic, Nancy Lobaugh, Michael Bronskill, Beth Levine, Alvan Feinstein, and Sandra Black. A robust method for extraction and automatic segmentation of brain images. *Neuroimage*, 17(3):1087–1100, 2002.

Jouni Kuha. AIC and BIC comparisons of assumptions and performance. *Sociological Methods & Research*, 33(2):188–229, 2004.

Maria Kuklisova-Murgasova, Paul Aljabar, Latha Srinivasan, Serena J Counsell, Valentina Doria, Ahmed Serag, Ioannis S Gousias, James P Boardman, Mary A Rutherford, A David Edwards, et al. A dynamic 4D probabilistic atlas of the developing brain. *NeuroImage*, 54(4):2750–2763, 2011.

Remi KS Kwan, Alan C Evans, and G Bruce Pike. MRI simulation-based evaluation of image-processing and classification methods. *IEEE Transactions on Medical Imaging*, 18(11):1085–1097, 1999.

Angela R Laird, Simon B Eickhoff, P Mickle Fox, Angela M Uecker, Kimberly L Ray, Juan J Saenz, D Reese McKay, Danilo Bzdok, Robert W Laird, Jennifer L Robinson, et al. The BrainMap strategy for standardization, sharing, and meta-analysis of neuroimaging data. *BMC research notes*, 4(1):349, 2011.

Bennet Allan Landman and Simon Warfield. MICCAI 2012: Grand challenge and Workshop on Multi-atlas labeling. In *Proc. International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI 2012*, 2012.

Thomas Robin Langerak, Uulke A van der Heide, Alexis NTJ Kotte, Max A Viergever, Marco Van Vulpen, and Josien PW Pluim. Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE). *IEEE Transactions on Medical Imaging*, 29(12):2000–2008, 2010.

Zhiqiang Lao, Dinggang Shen, Dengfeng Liu, Abbas F Jawad, Elias R Melhem, Lenore J Launer, R Nick Bryan, and Christos Davatzikos. Computer-assisted segmentation of white matter lesions in 3D MR images using support vector machine. *Academic Radiology*, 15(3):300–313, 2008.

Christian Thode Larsen, J Eugenio Iglesias, and Koen Van Leemput. N3 bias field correction explained as a Bayesian modeling method. In *Proc. International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI 2015, Bambi Workshop on Bayesian Models for Medical Image Analysis*, pages 1–12. Springer, 2015.

Julia A Lasserre, Christopher M Bishop, and Thomas P Minka. Principled hybrids of generative and discriminative models. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006*, volume 1, pages 87–94. IEEE, 2006.

I Nigel C Lawes, Thomas R Barrick, Vengadasalam Murugam, Natalia Spierings, David R Evans, Marie Song, and Chris A Clark. Atlas-based segmentation of white matter tracts of the human brain using diffusion tensor tractography and comparison with classical dissection. *Neuroimage*, 39(1):62–79, 2008.

Neil D Lawrence and John C Platt. Learning to learn with the informative vector machine. In *Proceedings of the 21th International Conference on Machine Learning*, page 65. ACM, 2004.

Matthieu Lê, Jan Unkelbach, Nicholas Ayache, and Hervé Delingette. Gpssi: Gaussian process for sampling segmentations of images. In *Proc. International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI 2015*, pages 38–46. Springer, 2015.

Tong Hau Lee, Mohammad Faizal Ahmad Fauzi, and Ryoichi Komiya. Segmentation of CT brain images using K-means and EM clustering. In *5th International Conference on Computer Graphics, Imaging and Visualisation, 2008*, pages 339–344. IEEE, 2008.

Thomas Martin Lehmann, Claudia Gonner, and Klaus Spitzer. Survey: Interpolation methods in medical image processing. *IEEE Transactions on Medical Imaging*, 18 (11):1049–1075, 1999.

Hava Lester and Simon R Arridge. A survey of hierarchical non-linear medical image registration. *Pattern Recognition*, 32(1):129–149, 1999.

Susan Levy, Marie-Anne Benhamou, Charles Naaman, Pierre Rainville, Virginie Callot, and J Cohen-Adad. White matter atlas of the human spinal cord with estimation of partial volume effect. *Neuroimage*, 119:262–271, 2015.

Emma B Lewis and Nicholas C Fox. Correction of differential intensity inhomogeneity in longitudinal MR images. *Neuroimage*, 23(1):75–83, 2004.

Rongjian Li, Wenlu Zhang, Heung-Il Suk, Li Wang, Jiang Li, Dinggang Shen, and Shuiwang Ji. Deep learning based imaging data completion for improved brain disease diagnosis. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI 2014*, pages 305–312. Springer, 2014.

Shuo Li, Thomas Fevens, Adam Krzyżak, and Song Li. Automatic clinical image segmentation using pathological modeling, PCA and SVM. *Engineering Applications of Artificial Intelligence*, 19(4):403–410, 2006.

Zhengrong Liang, Ronald J Jaszczak, and R Edward Coleman. Parameter estimation of finite mixtures using the EM algorithm and information criteria with application to medical image processing. *IEEE Transactions on Nuclear Science*, 39(4):1126–1133, 1992.

Zhengrong Liang, James R MacFall, and Donald P Harrington. Parameter estimation and tissue segmentation from multispectral MR images. *IEEE Transactions on Medical Imaging*, 13(3):441–449, 1994.

Boštjan Likar, Max A Viergever, and Franjo Pernuš. Retrospective correction of MR intensity inhomogeneity by information minimisation. In *Proc. International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI 2000*, pages 375–384. Springer, 2000.

Boštjan Likar, Max A Viergever, and Franjo Pernus. Retrospective correction of MR intensity inhomogeneity by information minimisation. *IEEE Transactions on Medical Imaging*, 20(12):1398–1410, 2001.

Antonio Criminisi Nicholas Ayache Loic le Folgoc, Herve Delingette. Sparse Bayesian registration of medical images for self-tuning of parameters and spatially adaptive parametrisation of displacements. *Medical Image Analysis*, December 2016.

Peter Lorenzen, Brad C. Davis, and Sarang Joshi. Unbiased atlas formation via large deformations metric mapping. In James S. Duncan and Guido Gerig, editors, *Proc. International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI 2005*, pages 411–418, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

Maria Lorenzo-Valdés, Gerardo I Sanchez-Ortiz, Andrew G Elkington, Raad H Mohiaddin, and Daniel Rueckert. Segmentation of 4D cardiac MR images using a probabilistic atlas and the EM algorithm. *Medical Image Analysis*, 8(3):255–265, 2004.

Nick Losseff and David H Miller. Measures of brain and spinal cord atrophy in multiple sclerosis. *Journal of Neurology, Neurosurgery and Psychiatry*, 64:S102, 1998.

Nick Losseff, SL Webb, Jonathan O'riordan, R Page, Liang Wang, Gareth Barker, Paul S Tofts, W Ian McDonald, David H Miller, and Alan J Thompson. Spinal cord atrophy and disability in multiple sclerosis. *Brain*, 119(3):701–708, 1996.

Miet Loubele, Frederik Maes, Filip Schutyser, Guy Marchal, Reinhilde Jacobs, and Paul Suetens. Assessment of bone segmentation quality of cone-beam CT versus multislice spiral CT: A pilot study. *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology*, 102(2):225–234, 2006.

Geert Lycklama, Alan Thompson, Massimo Filippi, David Miller, Christ Polman, Franz Fazekas, and Frederik Barkhof. Spinal-cord MRI in multiple sclerosis. *The Lancet Neurology*, 2(9):555–562, 2003.

David JC MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.

Frederik Maes, Andre Collignon, Dirk Vandermeulen, Guy Marchal, and Paul Suetens. Multimodality image registration by maximization of mutual information. *IEEE transactions on medical imaging*, 16(2):187–198, 1997.

Frederik Maes, Dirk Vandermeulen, and Paul Suetens. Medical image registration using mutual information. *Proceedings of the IEEE*, 91(10):1699–1722, 2003.

Dwarikanath Mahapatra. Joint segmentation and groupwise registration of cardiac perfusion images using temporal information. *Journal of Digital Imaging*, 26(2):173–182, 2013.

Dwarikanath Mahapatra and Ying Sun. Integrating segmentation information for improved MRF-based elastic image registration. *IEEE Transactions on Image Processing*, 21(1):170–183, 2012.

Ranjan Maitra and Julian Besag. Bayesian reconstruction in synthetic magnetic resonance imaging. In *SPIE's International Symposium on Optical Science, Engineering, and Instrumentation*, pages 39–47. International Society for Optics and Photonics, 1998.

Ranjan Maitra and John J Riddles. Synthetic magnetic resonance imaging revisited. *IEEE Transactions on Medical Imaging*, 29(3):895–902, 2010.

Suvrajit Maji and Marcel P Bruchez. Inferring biological structures from super-resolution single molecule images using generative models. *PloS one*, 7(5):e36973, 2012.

Joseph A Maldjian, Paul J Laurienti, Robert A Kraft, and Jonathan H Burdette. An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *Neuroimage*, 19(3):1233–1239, 2003.

Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive Neuroscience*, 19(9):1498–1507, 2007.

Donald W Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial & Applied Mathematics*, 11(2):431–441, 1963.

Gloria P Mazzara, Robert P Velthuizen, James L Pearlman, Harvey M Greenberg, and Henry Wagner. Brain tumor target volume determination for radiation treatment planning through automated MRI segmentation. *International Journal of Radiation Oncology* Biology* Physics*, 59(1):300–312, 2004.

John Mazziotta, Arthur Toga, Alan Evans, Peter Fox, Jack Lancaster, Karl Zilles, Roger Woods, Tomas Paus, Gregory Simpson, Bruce Pike, et al. A probabilistic atlas and reference system for the human brain: International consortium for brain mapping (ICBM). *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 356(1412):1293–1322, 2001.

John C Mazziotta, Arthur W Toga, Alan Evans, Peter Fox, and Jack Lancaster. A probabilistic atlas of the human brain: Theory and rationale for its development the international consortium for brain mapping (ICBM). *Neuroimage*, 2(2PA):89–101, 1995.

Andrew McCallum, Chris Pal, Gregory Druck, and Xuerui Wang. Multi-conditional learning: Generative/discriminative training for clustering and classification. In *AAAI*, pages 433–439, 2006.

Tim McInerney and Demetri Terzopoulos. Deformable models in medical image analysis: A survey. *Medical Image Analysis*, 1(2):91–108, 1996.

Andrea Mechelli, Cathy J Price, Karl J Friston, and John Ashburner. Voxel-based morphometry of the human brain: methods and applications. *Current Medical Imaging Reviews*, 1(2):105–113, 2005.

Xiao-Li Meng and Donald B Rubin. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2):267–278, 1993.

Bjoern H Menze, Koen Van Leemput, Danial Lashkari, Marc-André Weber, Nicholas Ayache, and Polina Golland. A generative model for brain tumor segmentation in multi-modal images. In *Proc. International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI 2010*, pages 151–159. Springer, 2010.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2015.

Bjoern H Menze, Koen Van Leemput, Danial Lashkari, Tammy Riklin-Raviv, Ezequiel Geremia, Esther Alberts, Philipp Gruber, Susanne Wegener, Marc-André Weber, Gabor Székely, et al. A generative probabilistic model and discriminative extensions for brain lesion segmentationwith application to tumor and stroke. *IEEE Transactions on Medical Imaging*, 35(4):933–946, 2016.

Hiroaki Mihara, Norio Iriguchi, and Shogo Ueno. A method of RF inhomogeneity correction in MR imaging. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 7(2):115–120, 1998.

David H Miller, Frederik Barkhof, Joseph A Frank, Geoffrey JM Parker, and Alan J Thompson. Measurement of atrophy in multiple sclerosis: pathological basis, methodological aspects and clinical relevance. *Brain*, 125(8):1676–1695, 2002.

Michael I Miller. Computational anatomy: Shape, growth, and atrophy comparison via diffeomorphisms. *Neuroimage*, 23:S19–S33, 2004.

Michael I. Miller, Alain Trouvé, and Laurent Younes. Geodesic shooting for computational anatomy. *Journal of Mathematical Imaging and Vision*, 24(2):209–228, 2006.

Fausto Milletari, Seyed-Ahmad Ahmadi, Christine Kroll, Annika Plate, Verena Rozanski, Juliana Maiostre, Johannes Levin, Olaf Dietrich, Birgit Ertl-Wagner, Kai Bötzel, et al. Hough-CNN: Deep learning for segmentation of deep brain regions in MRI and ultrasound. *arXiv preprint arXiv:1601.07014*, 2016.

Marc Modat, Gerard R Ridgway, Zeike A Taylor, Manja Lehmann, Josephine Barnes, David J Hawkes, Nick C Fox, and Sébastien Ourselin. Fast free-form deformation using graphics processing units. *Computer methods and programs in biomedicine*, 98 (3):278–284, 2010.

Jan Modersitzki. *Numerical Methods for Image Registration*. Oxford University Press on Demand, 2004.

Siawoosh Mohammadi, Patrick Freund, Thorsten Feiweier, Armin Curt, and Nikolaus Weiskopf. The impact of post-processing on spinal cord diffusion tensor imaging. *Neuroimage*, 70:377–385, 2013.

Nathan Moon, Elizabeth Bullitt, Koen Van Leemput, and GuidGo Erig. Model-based brain and tumor segmentation. In *16th International Conference on Pattern Recognition, 2002*, volume 1, pages 528–531. IEEE, 2002.

Jorge J Moré. The Levenberg-Marquardt algorithm: Implementation and theory. In *Numerical analysis*, pages 105–116. Springer, 1978.

Anam Mustaqeem, Ali Javed, and Tehseen Fatima. An efficient brain tumor detection algorithm using watershed & thresholding based segmentation. *International Journal of Image, Graphics and Signal Processing*, 4(10):34, 2012.

Radford M Neal and Geoffrey E Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer, 1998.

John Neter, Michael H Kutner, Christopher J Nachtsheim, and William Wasserman. *Applied linear statistical models*, volume 4. Irwin Chicago, 1996.

Christophoros Nikou, Aristidis C Likas, and Nikolaos P Galatsanos. A Bayesian framework for image segmentation with spatially varying mixtures. *IEEE Transactions on Image Processing*, 19(9):2278–2289, 2010.

Vincent Noblet, Christian Heinrich, Fabrice Heitz, and J-P Armspach. 3-D deformable image registration: A topology preservation scheme based on hierarchical deformation models and interval analysis optimisation. *IEEE Transactions on Image Processing*, 14(5):553–566, 2005.

Jorge Nocedal. Updating quasi-Newton matrices with limited storage. *Mathematics of computation*, 35(151):773–782, 1980.

Aljaž Noe and James C Gee. Partial volume segmentation of cerebral MRI scans with mixture model clustering. In *Information Processing in Medical Imaging*, pages 423–430. Springer, 2001.

Alireza Norouzi, Mohd Shafry Mohd Rahim, Ayman Altameem, Tanzila Saba, Abdolvahab Ehsani Rad, Amjad Rehman, and Mueen Uddin. Medical image segmentation methods, algorithms, and applications. *IETE Technical Review*, 31(3):199–213, 2014.

László G Nyúl, Jayaram K Udupa, and Xuan Zhang. New variants of a method of MRI scale standardization. *IEEE Transactions on Medical Imaging*, 19(2):143–150, 2000.

Jeff Orchard, Chen Greif, Gene H Golub, Bruce Bjornson, and M Stella Atkins. Simultaneous registration and activation detection for fMRI. *IEEE Transactions on Medical Imaging*, 22(11):1427–1435, 2003.

Yoshito Otake, Mehran Armand, Robert S Armiger, Michael D Kutzer, Ehsan Basafa, Peter Kazanzides, and Russell H Taylor. Intraoperative image-based multiview 2D/3D registration for image-guided orthopaedic surgery: incorporation of fiducial-based c-arm tracking and gpu-acceleration. *IEEE Transactions on Medical Imaging*, 31(4): 948–962, 2012.

Nikhil R Pal and Sankar K Pal. A review on image segmentation techniques. *Pattern Recognition*, 26(9):1277–1294, 1993.

Giorgio Parisi and Francesco Zamponi. Mean-field theory of hard sphere glasses and jamming. *Reviews of Modern Physics*, 82(1):789, 2010.

J Anthony Parker, Robert V Kenyon, and Donald E Troxel. Comparison of interpolating methods for image resampling. *IEEE Transactions on Medical Imaging*, 2(1):31–39, 1983.

Brian Patenaude, Stephen M Smith, David N Kennedy, and Mark Jenkinson. A Bayesian model of shape and appearance for subcortical brain segmentation. *Neuroimage*, 56(3):907–922, 2011.

Zhigang Peng, William Wee, and Jing-Huei Lee. Automatic segmentation of MR brain images using spatial-varying Gaussian mixture and Markov random field approach. In *Conference on Computer Vision and Pattern Recognition Workshop, 2006*, pages 80–80. IEEE, 2006.

Graeme P Penney, Jürgen Weese, John A Little, Paul Desmedt, Derek LG Hill, et al. A comparison of similarity measures for use in 2-D-3-D medical image registration. *IEEE Transactions on Medical Imaging*, 17(4):586–595, 1998.

Joao MS Pereira, L Xiong, Julio Acosta-Cabronero, George Pengas, Guy B Williams, and Peter J Nestor. Registration accuracy for VBM studies varies according to region and degenerative disease grouping. *Neuroimage*, 49(3):2205–2215, 2010.

Cheryl A Petersilge, Jonathan S Lewin, Jeffrey L Duerk, Jung U Yoo, and Alexander J Ghaneyem. optimising imaging parameters for MR evaluation of the spine with titanium pedicle screws. *AJR. American Journal of Roentgenology*, 166(5):1213–1218, 1996.

Vladimir S Petrovic, Timothy F Cootes, AM Mills, Carole J Twining, and Christopher J Taylor. Automated analysis of deformable structure in groups of images. In *British Machine Vision Conference*, volume 2, pages 1060–1069, 2007.

Dzung L Pham, Chenyang Xu, and Jerry L Prince. Current methods in medical image segmentation. *Annual review of biomedical engineering*, 2(1):315–337, 2000.

John Pluta, Brian B. Avants, Simon Glynn, Suyash Awate, James C. Gee, and John A. Detre. Appearance and incomplete label matching for diffeomorphic template based hippocampus segmentation. *Hippocampus*, 19(6):565–571, 2009.

Kilian M Pohl, John Fisher, James J Levitt, Martha E Shenton, Ron Kikinis, W Eric L Grimson, and William M Wells. A unifying approach to registration, segmentation, and intensity correction. In *Proc. International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI 2005*, pages 310–318. Springer, 2005.

Kilian M Pohl, John Fisher, W Eric L Grimson, Ron Kikinis, and William M Wells. A

Bayesian model for joint segmentation and registration. *Neuroimage*, 31(1):228–239, 2006.

Regina Pohle and Klaus D Toennies. Segmentation of medical images using adaptive region growing. In *Medical Imaging 2001*, pages 1337–1346. International Society for Optics and Photonics, 2001.

Valeriu Popescu, M Battaglini, WS Hoogstrate, SCJ Verfaillie, IC Sluimer, Ronald A van Schijndel, BW Van Dijk, Keith S Cover, Dirk L Knol, Mark Jenkinson, et al. optimising parameter choice for FSL-Brain Extraction Tool (bet) on 3D T1 images in multiple sclerosis. *Neuroimage*, 61(4):1484–1494, 2012.

Michael JD Powell. A fast algorithm for nonlinearly constrained optimisation calculations. In *Numerical Analysis*, pages 144–157. Springer, 1978.

Stephanie Powell, Vincent A Magnotta, Hans Johnson, Vamsi K Jammalamadaka, Ronald Pierson, and Nancy C Andreasen. Registration and machine learning-based automated segmentation of subcortical and cerebellar brain structures. *Neuroimage*, 39(1):238–247, 2008.

Ferran Prados, John Ashburner, Claudia Blaiotta, Tom Brosch, Julio Carballido-Gamio, Manuel Jorge Cardoso, Benjamin N. Conrad, Esha Datta, Gergely Dvid, Benjamin De Leener, Sara M. Dupont, Patrick Freund, Claudia A.M. Gandini Wheeler-Kingshott, Francesco Grussu, Roland Henry, Bennett A. Landman, Emil Ljungberg, Bailey Lyttle, Sebastien Ourselin, Nico Papinutto, Salvatore Saporito, Regina Schlaeger, Seth A. Smith, Paul Summers, Roger Tam, Marios C. Yiannakas, Alyssa Zhu, and Julien Cohen-Adad. Spinal cord grey matter segmentation challenge. *NeuroImage*, 152:312 – 329, 2017.

Marcel Prastawa. Automatic brain tumor segmentation by subject specific modification of atlas priors. *Academic Radiology*, 10(12):1341–1348, 2003.

Marcel Prastawa, Elizabeth Bullitt, Sean Ho, and Guido Gerig. A brain tumor segmentation framework based on outlier detection. *Medical Image Analysis*, 8(3):275–283, 2004.

Carey E Priebe and David J Marchette. Alternating kernel and mixture density estimates. *Computational Statistics & Data Analysis*, 35(1):43–65, 2000.

Rajat Raina, Yirong Shen, Andrew Y Ng, and Andrew McCallum. Classification with hybrid generative/discriminative models. In *NIPS*, volume 16, 2003.

Rajat Raina, Andrew Y Ng, and Daphne Koller. Constructing informative priors using transfer learning. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 713–720. ACM, 2006.

Jagath C Rajapakse and Frithjof Kruggel. Segmentation of MR images with intensity inhomogeneities. *Image and Vision Computing*, 16(3):165–180, 1998.

Wilburn E Reddick, Raymond K Mulhern, T David Elkin, John O Glass, Thomas E Merchant, and James W Langston. A hybrid neural network analysis of subtle brain volume differences in children surviving brain tumors. *Magnetic Resonance Imaging*, 16(4):413–421, 1998.

Annemie Ribbens, Jeroen Hermans, Frederik Maes, Dirk Vandermeulen, and Paul Suetens. SPARC: Unified framework for automatic segmentation, probabilistic atlas construction, registration and clustering of brain MR images. In *IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2010*, pages 856–859. IEEE, 2010.

Tammy Riklin-Raviv, Koen Van Leemput, Bjoern H Menze, William M Wells, and Polina Golland. Segmentation of image ensembles via latent atlases. *Medical image analysis*, 14(5):654–665, 2010.

Petter Risholm, Steve Pieper, Eigil Samset, and William M Wells III. Summarizing and visualizing uncertainty in non-rigid registration. In *Proc. International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI 2010*, pages 554–561. Springer, 2010.

Maria A Rocca, Paolo Preziosa, and Massimo Filippi. *Spinal Cord Diseases*. Oxford University Press, USA, 2015.

Torsten Rohlfing and Calvin R. Maurer. Multi-classifier framework for atlas-based image segmentation. *Pattern Recognition Letters*, 26(13):2070–2079, 2005.

Torsten Rohlfing, Robert Brandt, Randolf Menzel, and Calvin R Maurer Jr. Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. *Neuroimage*, 21(4):1428–1442, 2004.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI 2015*, pages 234–241. Springer, 2015.

Snehashis Roy, Aaron Carass, and Jerry Prince. A compressed sensing approach for MR tissue contrast synthesis. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 371–383. Springer, 2011.

Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Ansgar Rudisch, Christian Kremser, Siegfried Peer, Anton Kathrein, Werner Judmaier, and Herwing Daniaux. Metallic artifacts in magnetic resonance imaging of patients with spinal fusion: A comparison of implant materials and imaging sequences. *Spine*, 23(6):692–699, 1998.

Daniel Rueckert, Alejandro F Frangi, and Julia A Schnabel. Automatic construction of 3-D statistical deformation models of the brain using nonrigid registration. *IEEE Transactions on Medical Imaging*, 22(8):1014–1025, 2003.

Daniel Rueckert, Paul Aljabar, Rolf A Heckemann, Joseph V Hajnal, and Alexander Hammers. Diffeomorphic registration using B-splines. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 702–709. Springer, 2006.

Daniel Rueckert, Ben Glocker, and Bernhard Kainz. Learning clinically useful information from images: Past, present and future, 2016.

Mert R Sabuncu, Serdar K Balci, and Polina Golland. Discovering modes of an image population through mixture modeling. In *Proc. International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI 2008*, pages 381–389. Springer, 2008.

Mert R Sabuncu, BT Thomas Yeo, Koen Van Leemput, Bruce Fischl, and Polina Golland. A generative model for image segmentation based on label fusion. *IEEE Transactions on Medical Imaging*, 29(10):1714–1729, 2010.

Rachael I Scahill, Chris Frost, Rhian Jenkins, Jennifer L Whitwell, Martin N Rossor, and Nick C Fox. A longitudinal study of brain volume changes in normal aging using serial registered magnetic resonance imaging. *Archives of Neurology*, 60(7):989–994, 2003.

Tanya Schmah, Geoffrey E Hinton, Steven L Small, Stephen Strother, and Richard S Zemel. Generative versus discriminative training of RBMs for classification of fMRI images. In *Advances in Neural Information Processing Systems*, pages 1409–1416, 2008.

Christopher Schwarz, Evan Fletcher, Charles DeCarli, and Owen Carmichael. Fully-automated white matter hyperintensity detection with anatomical prior knowledge and without FLAIR. In *International Conference on Information Processing in Medical Imaging*, pages 239–251. Springer, 2009.

Matthias Seeger. Covariance kernels from Bayesian generative models. *Advances in Neural Information Processing Systems*, 2:905–912, 2002.

Jasjeet S Sekhon and Walter R Mebane Jr. Genetic optimisation using derivatives. *Political Analysis*, pages 187–210, 1998.

S Kamaledin Setarehdan and Sameer Singh. *Advanced algorithmic approaches to medical image segmentation: state-of-the-art applications in cardiology, neurology, mammography and pathology.* Springer Science & Business Media, 2012.

Giorgos Sfikas, Christophoros Nikou, and Nikolaos Galatsanos. Robust image segmentation with mixtures of Student's t-distributions. In *IEEE International Conference on Image Processing, 2007*, volume 1, pages I–273. IEEE, 2007.

Neeraj Sharma and Lalit M Aggarwal. Automated medical image segmentation techniques. *Journal of medical physics/Association of Medical Physicists of India*, 35(1): 3, 2010.

Ravi K Sharma, Todd K Leen, and Misha Pavel. Bayesian sensor image fusion using local linear generative models. *Optical Engineering*, 40(7):1364–1376, 2001.

David W Shattuck, Stephanie R Sandor-Leahy, Kirt A Schaper, David A Rottenberg, and Richard M Leahy. Magnetic resonance image tissue classification using a partial volume model. *Neuroimage*, 13(5):856–876, 2001.

David W Shattuck, Mubeena Mirza, Vitria Adisetiyo, Cornelius Hojatkashani, Georges Salamon, Katherine L Narr, Russell A Poldrack, Robert M Bilder, and Arthur W Toga. Construction of a 3D probabilistic atlas of human cortical structures. *Neuroimage*, 39(3):1064–1080, 2008.

Dinggang Shen and Christos Davatzikos. Measuring temporal morphological changes robustly in brain MR images via 4-dimensional template warping. *Neuroimage*, 21 (4):1508–1517, 2004.

Tony Shepherd, Simon JD Prince, and Daniel C Alexander. Interactive lesion segmentation with shape priors from offline and online learning. *IEEE Transactions on Medical Imaging*, 31(9):1698–1712, 2012.

IJA Simpson, MJ Cardoso, M Modat, DM Cash, MW Woolrich, JLR Andersson, JA Schnabel, S Ourselin, Alzheimers Disease Neuroimaging Initiative, et al. Probabilistic non-linear registration with spatially adaptive regularisation. *Medical Image Analysis*, 26(1):203–216, 2015.

Ivor JA Simpson, Julia A Schnabel, Adrian R Groves, Jesper LR Andersson, and Mark W Woolrich. Probabilistic inference of regularisation in non-rigid registration. *Neuroimage*, 59(3):2438–2451, 2012.

John G Sled and G Bruce Pike. Understanding intensity non-uniformity in MRI. In William M. Wells, Alan Colchester, and Scott Delp, editors, *Proc. International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI 1998*, volume 1496 of *LNCS*, pages 614–622, Berlin, 1998. Springer.

John G Sled, Alex P Zijdenbos, and Alan C Evans. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE transactions on medical imaging*, 17(1):87–97, 1998.

Stephen M Smith, Mark Jenkinson, Mark W Woolrich, Christian F Beckmann, Timothy EJ Behrens, Heidi Johansen-Berg, Peter R Bannister, Marilena De Luca, Ivana Drobnjak, David E Flitney, et al. Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage*, 23:S208–S219, 2004.

Carlos Oscar Sánchez Sorzano, Philippe Thévenaz, and Michael Unser. Elastic registration of biological images using vector-spline regularization. *IEEE Transactions on Biomedical Engineering*, 52(4):652–663, 2005.

Aristeidis Sotiras, Christos Davatzikos, and Nikos Paragios. Deformable medical image registration: A survey. *IEEE transactions on medical imaging*, 32(7):1153–1190, 2013.

Radu Stefanescu, Xavier Pennec, and Nicholas Ayache. Grid powered nonlinear image registration with locally adaptive regularization. *Medical Image Analysis*, 8(3):325–342, 2004.

Patrick W Stroman, Claudia Wheeler-Kingshott, M Bacon, Joseph Schwab, Rachel Bosma, J Brooks, David Cadotte, Thomas Carlstedt, Olga Ciccarelli, Julien Cohen-Adad, et al. The current state-of-the-art of spinal cord imaging: Methods. *Neuroimage*, 84:1070–1081, 2014.

Irene Sturm, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Interpretable deep neural networks for single-trial EEG classification. *Journal of neuroscience methods*, 274:141–145, 2016.

Martin Styner, Christian Brechbuhler, G Szckely, and Guido Gerig. Parametric estimate of intensity inhomogeneities applied to mri. *IEEE Transactions on Medical Imaging*, 19(3):153–165, 2000.

Carole H Sudre, M Jorge Cardoso, Willem H Bouvy, Geert Jan Biessels, Josephine Barnes, and Sebastien Ourselin. Bayesian model selection for pathological neuroimaging data applied to white matter lesion segmentation. *IEEE Transactions on Medical Imaging*, 34(10):2079–2102, 2015.

Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT press, 1998.

Katherine H Taber, Richard C Herrick, Susan W Weathers, Ashok J Kumar, Donald F Schomer, and L Anne Hayman. Pitfalls and artifacts encountered in clinical MR imaging of the spine. *Radiographics*, 18(6):1499–1521, 1998.

Manuel Taso, Arnaud Le Troter, Michaël Sdika, Jean-Philippe Ranjeva, Maxime Guye, Monique Bernard, and Virginie Callot. Construction of an in vivo human spinal cord atlas based on high-resolution MR images at cervical and thoracic levels: preliminary results. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 27(3):257–267, 2014.

Paul M Thompson and Arthur W Toga. Detection, visualization and animation of abnormal anatomic structure with a deformable probabilistic brain atlas based on random vector field transformations. *Medical Image Analysis*, 1(4):271–294, 1997.

Paul M Thompson, Roger P Woods, Michael S Mega, and Arthur W Toga. Mathematical/computational challenges in creating deformable and probabilistic atlases of the human brain. *Human Brain Mapping*, 9(2):81–92, 2000.

GuangJian Tian, Yong Xia, Yanning Zhang, and Dagan Feng. Hybrid genetic and variational expectation-maximisation algorithm for Gaussian-mixture-model-based brain MR image segmentation. *IEEE Transactions on Information Technology in Biomedicine*, 15(3):373–380, 2011.

Luke Tierney and Joseph B Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393): 82–86, 1986.

M Tincher, Charles Meyer, R Gupta, and DM Williams. Polynomial modeling and reduction of RF body coil spatial inhomogeneity in MRI. *IEEE Transactions on Medical Imaging*, 12(2):361–365, 1993.

Michael E Tipping. Probabilistic visualisation of high-dimensional binary data. In *Advances in neural information processing systems*, pages 592–598, 1999.

Michael Titterington, Adrian Smith, and Udi Makov. *Statistical Analysis of Finite Mixture Distributions.* Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons, 1985.

PS Tofts. Standardisation and optimisation of magnetic resonance techniques for multicentre studies. *Journal of Neurology, Neurosurgery and Psychiatry*, 64:S37, 1998.

Arthur W Toga and Paul M Thompson. Image registration and the construction of multidimensional brain atlases. In Isaac N Bankman, editor, *Handbook of Medical Imaging: Medical image processing and Analysis*, Biomedical Engineering, pages 635 – 653. Academic Press, San Diego, 2000.

Alain Trouvé. Diffeomorphisms groups and pattern matching in image analysis. *International Journal of Computer Vision*, 28(3):213–221, 1998.

Jeffrey Tsao. Interpolation artifacts in multimodality image registration based on maximisation of mutual information. *IEEE Transactions on Medical Imaging*, 22(7): 854–864, 2003.

Zhuowen Tu, Katherine L Narr, Piotr Dollár, Ivo Dinov, Paul M Thompson, and Arthur W Toga. Brain anatomical structure segmentation by hybrid discriminative/generative models. *IEEE Transactions on Medical Imaging*, 27(4):495–508, 2008.

Nicholas J Tustison, Brian B Avants, Philip A Cook, Yuanjie Zheng, Alexander Egan, Paul A Yushkevich, and James C Gee. N4ITK: improved N3 bias correction. *IEEE transactions on medical imaging*, 29(6):1310–1320, 2010.

Dimitris G Tzikas, Aristidis C Likas, and Nickolaos P Galatsanos. The variational approximation for Bayesian inference. *Signal Processing Magazine, IEEE*, 25(6):131–146, 2008.

Naonori Ueda, Ryohei Nakano, Zoubin Ghahramani, and Geoffrey E Hinton. SMEM algorithm for mixture models. *Neural computation*, 12(9):2109–2128, 2000.

Suthirth Vaidya, Abhijith Chunduru, Ramanathan Muthuganapathy, and Ganapathy Krishnamurthi. Longitudinal multiple sclerosis lesion segmentation using 3D convolutional neural networks. *Proceedings of the 2015 Longitudinal Multiple Sclerosis Lesion Segmentation Challenge*, pages 1–2, 2015.

Koen Van Leemput. Encoding probabilistic brain atlases using Bayesian inference. *IEEE Transactions on Medical Imaging*, 28(6):822–837, 2009.

Koen Van Leemput, Frederik Maes, Fernando Bello, Dirk Vandermeulen, Alan Colchester, and Paul Suetens. Automated segmentation of MS lesions from multi-channel MR images. In *Proc. International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI 1999*, pages 11–21. Springer, 1999a.

Koen Van Leemput, Frederik Maes, Dirk Vandermeulen, and Paul Suetens. Automated model-based tissue classification of MR images of the brain. *IEEE Transactions on Medical Imaging*, 18(10):897–908, 1999b.

Koen Van Leemput, Frederik Maes, Dirk Vandermeulen, and Paul Suetens. Automated model-based tissue classification of MR images of the brain. *IEEE Transactions on Medical Imaging*, 18(10):897–908, 1999c.

Koen Van Leemput, Frederik Maes, Dirk Vandermeulen, and Paul Suetens. Automated model-based bias field correction of MR images of the brain. *IEEE Transactions on Medical Imaging*, 18(10):885–896, 1999d.

Koen Van Leemput, Frederik Maes, Dirk Vandermeulen, Alan Colchester, and Paul Suetens. Automated segmentation of multiple sclerosis lesions by model outlier detection. *IEEE Transactions on Medical Imaging*, 20(8):677–688, 2001.

Koen Van Leemput, Frederik Maes, Dirk Vandermeulen, and Paul Suetens. A unifying framework for partial volume segmentation of brain MR images. *IEEE Transactions on Medical Imaging*, 22(1):105–119, 2003.

Gijs van Tulder and Marleen de Bruijne. Why does synthesized data improve multi-sequence classification? In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI 2015*, pages 531–538. Springer, 2015.

Gijs van Tulder and Marleen de Bruijne. Combining generative and discriminative representation learning for lung CT analysis with convolutional restricted boltzmann machines. *IEEE Transactions on Medical Imaging*, 35(5):1262–1272, 2016.

Robert Van Uitert, Ingmar Bitter, and John A Butman. Semi-automatic spinal cord segmentation and quantification. In *International Congress Series*, volume 1281, pages 224–229. Elsevier, 2005.

John Thomas Vaughan, Michael Garwood, CM Collins, W Liu, L DelaBarre, G Adriany, P Andersen, H Merkle, R Goebel, MB Smith, et al. 7T vs. 4T: RF power, homogeneity, and signal-to-noise comparison in head images. *Magnetic resonance in medicine*, 46 (1):24–30, 2001.

Alfredo Vellido, José David Martín-Guerrero, and Paulo JG Lisboa. Making machine learning models interpretable. In *ESANN*, volume 12, pages 163–172, 2012.

Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache. Diffeomorphic demons: Efficient non-parametric image registration. *Neuroimage*, 45(1): S61–S72, 2009.

François-Xavier Vialard, Laurent Risser, Daniel Rueckert, and Darryl D Holm. Diffeomorphic atlas estimation using geodesic shooting on volumetric images. *Ann. BMVA*, 2012, 2012.

Hongzhi Wang, Jung W Suh, Sandhitsu R Das, John B Pluta, Caryne Craige, and Paul A Yushkevich. Multi-atlas segmentation with joint label fusion. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 35(3):611–623, 2013.

Lei Wang, David Y Lee, Ellen Bailey, Johanna M Hartlein, Mohktar H Gado, Michael I Miller, and Kevin J Black. Validity of large-deformation high dimensional brain mapping of the basal ganglia in adults with Tourette syndrome. *Psychiatry Research: Neuroimaging*, 154(2):181–190, 2007.

Liqun Wang, H Lai, Gareth J Barker, David H Miller, Paul S Tofts, et al. Correction for variations in MRI scanner sensitivity in brain studies with histogram matching. *Magnetic Resonance in Medicine*, 39(2):322–327, 1998.

Qian Wang, Dieter Seghers, Emiliano DAgostino, Frederik Maes, Dirk Vandermeulen, Paul Suetens, and Alexander Hammers. Construction and validation of mean shape atlas templates for atlas-based brain image segmentation. In GaryE. Christensen and Milan Sonka, editors, *Information Processing in Medical Imaging*, volume 3565 of *LNCS*, pages 689–700, Berlin, 2005. Springer.

Simon Warfield, Andre Robatino, Joachim Dengler, Ferenc Jolesz, and Ron Kikinis.

Nonlinear registration and template driven segmentation. *Brain Warping*, 4:67–84, 1999.

Simon K Warfield, Kelly H Zou, and William M Wells. Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–921, 2004.

Frank W Warner. *Foundations of Differentiable Manifolds and Lie Groups*, volume 94. Springer Science & Business Media, 2013.

Neil L Weisenfeld and SK Warfteld. Normalisation of joint image-intensity statistics in MRI using the Kullback-Leibler divergence. In *IEEE International Symposium on Biomedical Imaging: Nano to Macro, 2004*, pages 101–104. IEEE, 2004.

Nikolaus Weiskopf, John Suckling, Guy Williams, Marta Morgado Correia, Becky Inkster, Roger Tait, Cinly Ooi, Edward T Bullmore, and Antoine Lutti. Quantitative multi-parameter mapping of R, PD*, MT, and R2* at 3T: A multi-center validation. *Frontiers in Neuroscience*, 7:95, 2013.

William M Wells, Paul Viola, Hideki Atsumi, Shin Nakajima, and Ron Kikinis. Multimodal volume registration by maximization of mutual information. *Medical image analysis*, 1(1):35–51, 1996.

Williams M Wells III, W Eric L Grimson, Ron Kikinis, and Ferenc A Jolesz. Adaptive segmentation of MRI data. *IEEE Transactions on Medical Imaging*, 15(4):429–442, 1996.

Elisabeth Wenger, Johan Mårtensson, Hannes Noack, Nils Christian Bodammer, Simone Kühn, Sabine Schaefer, Hans-Jochen Heinze, Emrah Düzel, Lars Bäckman, Ulman Lindenberger, et al. Comparing manual and automatic segmentation of hippocampal volumes: Reliability and validity issues in younger and older brains. *Human Brain Mapping*, 35(8):4236–4248, 2014.

Claudia Wheeler-Kingshott, Patrick W Stroman, Joseph Schwab, M Bacon, Rachel Bosma, J Brooks, David Cadotte, Thomas Carlstedt, Olga Ciccarelli, Julien Cohen-Adad, et al. The current state-of-the-art of spinal cord imaging: Applications. *Neuroimage*, 84:1082–1093, 2014.

Peter M Williams. Bayesian regularization and pruning using a Laplace prior. *Neural Computation*, 7(1):117–143, 1995.

Bertram J Wilm, Jonas Svensson, Alice Henning, Klaas P Pruessmann, Peter Boesiger, and Spyros S Kollias. Reduced field-of-view MRI using outer volume suppression for spinal cord diffusion imaging. *Magnetic Resonance in Medicine*, 57(3):625–630, 2007.

Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3):37–52, 1987.

Gert Wollny and Frithjof Kruggel. Computational cost of nonrigid registration algorithms based on fluid dynamics. *IEEE Transactions on Medical Imaging*, 21(8): 946–952, 2002.

Roger P Woods. Characterizing volume and surface deformations in an atlas framework: theory, applications, and implementation. *NeuroImage*, 18(3):769–788, 2003.

Mark W Woolrich, Saad Jbabdi, Brian Patenaude, Michael Chappell, Salima Makni, Timothy Behrens, Christian Beckmann, Mark Jenkinson, and Stephen M Smith. Bayesian analysis of neuroimaging data in FSL. *Neuroimage*, 45(1):S173–S186, 2009.

Mark William Woolrich and Timothy E Behrens. Variational Bayes inference of spatial mixture models for segmentation. *IEEE Transactions on Medical Imaging*, 25(10): 1380–1391, 2006.

Guorong Wu, Qian Wang, Daoqiang Zhang, Feiping Nie, Heng Huang, and Dinggang Shen. A generative probability model of joint label fusion for multi-atlas based brain segmentation. *Medical Image Analysis*, 18(6):881–890, 2014a.

Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding. Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, 26(1):97–107, 2014b.

Yiming Wu, Xiangyu Yang, and Kap Luk Chan. Unsupervised color image segmentation based on gaussian mixture model. In *Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia.*, volume 1, pages 541–544. IEEE, 2003.

Paul P Wyatt and J Alison Noble. MAP MRF joint segmentation and registration of medical images. *Medical Image Analysis*, 7(4):539–552, 2003.

Chen Xiaohua, Michael Brady, and Daniel Rueckert. Simultaneous segmentation and registration for medical image. In Christian Barillot, DavidR. Haynor, and Pierre Hellier, editors, *Proc. International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI 2004*, volume 3216 of *LNCS*, pages 663–670, Berlin, 2004a. Springer.

Chen Xiaohua, Michael Brady, and Daniel Rueckert. Simultaneous segmentation and registration for medical image. In *Proc. International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI 2004*, pages 663–670. Springer, 2004b.

Hao Xu, Bertrand Thirion, and Stéphanie Allassonniere. Probabilistic atlas and geometric variability estimation to drive tissue segmentation. *Statistics in Medicine*, 33 (20):3576–3599, 2014.

B Yegnanarayana. *Artificial neural networks*. PHI Learning Pvt. Ltd., 2009.

BT Thomas Yeo, Mert R Sabuncu, Rahul Desikan, Bruce Fischl, and Polina Golland. Effects of registration regularization and atlas sharpness on segmentation accuracy. *Medical Image Analysis*, 12(5):603–615, 2008.

Anthony Yezzi, Lilla Zollei, and Tina Kapur. A variational framework for joint segmentation and registration. In *Mathematical Methods in BioMedical Image Analysis, MMBIA 2001, IEEE Workshop on*, pages 44–51. IEEE, 2001.

Zhao Yi, Antonio Criminisi, Jamie Shotton, and Andrew Blake. Discriminative, semantic segmentation of brain tissue in MR images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 558–565. Springer, 2009.

MC Yiannakas, H Kearney, RS Samson, Declan T Chard, Olga Ciccarelli, David H Miller, and Claudia AM Wheeler-Kingshott. Feasibility of grey matter and white matter segmentation of the upper cervical cord in vivo: A pilot study with application to magnetisation transfer measurements. *Neuroimage*, 63(3):1054–1059, 2012.

Laurent Younes. *Shapes and Diffeomorphisms*, volume 171. Springer Science & Business Media, 2010.

Laurent Younes, Felipe Arrate, and Michael I Miller. Evolutions equations in computational anatomy. *Neuroimage*, 45(1):S40–S50, 2009.

Evangelia I Zacharaki, Sumei Wang, Sanjeev Chawla, Dong Soo Yoo, Ronald Wolf, Elias R Melhem, and Christos Davatzikos. Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme. *Magnetic resonance in medicine*, 62(6):1609–1618, 2009.

Miaomiao Zhang and P Thomas Fletcher. Probabilistic principal geodesic analysis. In *Advances in Neural Information Processing Systems*, pages 1178–1186, 2013.

Miaomiao Zhang and P Thomas Fletcher. Bayesian principal geodesic analysis in diffeomorphic image registration. In *Proc. International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI 2014*, pages 121–128. Springer, 2014.

Miaomiao Zhang, Nikhil Singh, and P Thomas Fletcher. Bayesian estimation of regularization and atlas building in diffeomorphic image registration. In *Information Processing in Medical Imaging*, pages 37–48. Springer, 2013.

Miaomiao Zhang, Hang Shao, and P Thomas Fletcher. A mixture model for automatic diffeomorphic multi-atlas building. In *Proc. International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI 2015, Bambi Workshop on Bayesian Models for Medical Image Analysis*, 2015a.

Tongjie Zhang and Ching Y. Suen. A fast parallel algorithm for thinning digital patterns. *Communications of the ACM*, 27(3):236–239, 1984.

Wenlu Zhang, Rongjian Li, Houtao Deng, Li Wang, Weili Lin, Shuiwang Ji, and Dinggang Shen. Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. *Neuroimage*, 108:214–224, 2015b.

Yongyue Zhang, Michael Brady, and Stephen Smith. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximisation algorithm. *IEEE Transactions on Medical Imaging*, 20(1):45–57, 2001.

Xiaojin Zhu. Semi-supervised learning literature survey. *Computer Science, University of Wisconsin-Madison*, 2(3):4, 2006.

Barbara Zitova and Jan Flusser. Image registration methods: A survey. *Image and vision computing*, 21(11):977–1000, 2003.

Lilla Zöllei, John W Fisher III, and William M Wells III. A unified statistical and information theoretic framework for multi-modal image registration. In *IPMI*, pages 366–377. Springer, 2003.

Lilla Zöllei, Mark Jenkinson, Samson Timoner, and William Wells. A marginalized MAP approach and EM optimisation for pair-wise registration. In *Information Processing in Medical Imaging*, pages 662–674. Springer, 2007a.

Lilla Zöllei, Martha Shenton, William Wells, and Kilian Pohl. The impact of atlas formation methods on atlas-guided brain segmentation. In *Proc. International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI 2007, Statistical Registration Workshop: Pair-wise and Group-wise Alignment and Atlas Formation*, volume 10, pages 39–46, 2007b.