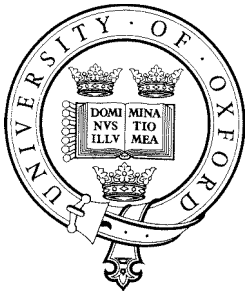
Bayesian Statistical Models of Shape and Appearance for Subcortical Brain Segmentation

Brian Patenaude



Trinity Term, 2007

Oxford Centre for Functional Magnetic Resonance Imaging of the Brain

Department of Clinical Neurology

University of Oxford

To my loving wife and son, Amanda and Sebastian, the two lights of my life.

Bayesian Statistical Models of Shape and Appearance for Subcortical Brain Segmentation

Brian Patenaude
Worcester College
University of Oxford

Abstract

Our motivation is to develop an automated technique for segmentation of sub-cortical human brain structures from MR images. To this purpose, models of shape-and-appearance are constructed and fit to new image data. The statistical models are trained from 317 manually labelled T1-weighted MR images. Shape is modelled using a surface-based point distribution model (PDM) such that the shape space is constrained to the linear combination of the mean shape and eigenvectors of the vertex coordinates. In addition, to model intensity at the structural boundary, intensities are sampled along the surface normal from the underlying image. We propose a novel Bayesian appearance model whereby the relationship between shape and intensity are modelled via the conditional distribution of intensity given shape. Our fully probabilistic approach eliminates the need for arbitrary weightings between shape and intensity as well as for tuning parameters that specify the relative contribution between the use of shape constraints and intensity information. Leave-one-out cross-validation is used to validate the model and fitting for 17 structures.

The PDM for shape requires surface parameterizations of the volumetric, manual labels such that vertices retain a one-to-one correspondence across the training subjects. Surface parameterizations with correspondence are generated through the use of deformable models under constraints that embed the correspondence criterion within the deformation process. A novel force that favours equal-area triangles throughout the mesh is introduced. The force adds stability to the mesh such that minimal smoothing or within-surface motion is required.

The use of the PDM for segmentation across a series of subjects results in a set surfaces that retain point correspondence. The correspondence facilitates landmark-based shape analysis. Amongst other metrics, vertex-wise multivariate statistics and discriminant analysis are used to investigate local and global size and shape differences between groups. The model is fit, and shape analysis is applied to two clinical datasets.

Thesis submitted for the degree Doctorate of Philosophy
at the University of Oxford
Trinity Term, 2007

Acknowledgements

We would like to thank David Kennedy and the Center for Morphometric Analysis (CMA) for providing the manually labelled T1-weighted images, without which this work would not have been possible. I would also like to thank Chirstian Haselgrove for the pratical aspects of procuring the data. Also, thanks to all the individual researchers that contributed their data through the CMA.

We are very grateful to Nick Fox and Richard Boyes, Dementia Research Centre, Institute of Neurology, UCL, UK for the MIRIAD data.

Also, to Dr. Clare Mackay, Prof. Tim Crow and MRC program "Brain Development and Structure in Psychosis" for providing the schizophrenia data that was used in chapter 4.

I would also like to acknowledge the ESPRC, IBIM grant for funding. I would like to thank the members of the IBIM collaboration, all of whom have influenced this research in some facet.

I would like to give a special thanks to Mark and Steve for providing me the opportunity to pursue my D.Phil as well as for their guidance and input into this research. Finally, I would like to thank Amanda and Sebastian for being troopers through the final stages of thesis, and for helping me keep my sanity.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Medical imaging allows researchers and clinicians to non-invasively investigate the structural variations of anatomy in-vivo. Structural variation in anatomy between a patient group and the normal population may be used to further understand the mechanisms underlying a pathology, to aid in the diagnosis of a disease, or to assess the response to treatment. Typically, the clinician or researcher is interested in a specific region of an image; most frequently this will correspond to an anatomical structure (for example, the hippocampus). The definition of these boundaries would historically require an expert operator to manually segment the structure from an image, which is a time consuming process and subject to inter- and intra-rater variability. In general, this provides the motivation for the field of medical image segmentation; automated methods eliminate the need for expertly trained operators, the extensive man-hours required, and the inter- and intra-rater variability.

Automated methods for image segmentation are continually improving in accuracy

and robustness as well as increasing in sophistication. Recent improvements in accuracy and robustness typically involve the incorporation of learnt shape information. Our method for automated segmentation concentrates on the segmentation of subcortical structures from T_1-weighted MR images of the human brain. Although the method is applied directly to MR brain images, many of the advances could contribute more generally to the field of image segmentation. The method is based on statistical models of shape and intensity for each structure.

In the chapters to follow, the models, their training and their application to new data will be discussed in detail. Chapter 2 addresses the pre-processing of the training data (manually labelled T_1-weighted MR images) that is required to construct our models. The pre-processing includes registration to a standard space, surface parameterization of volumetric labels, and intensity sampling with normalization. Chapter 3 describes the mathematical framework used to model shape and intensity, as well as its fitting to new images. The models are trained from the surfaces and intensity samples that were derived from the pre-processing described in chapter 2. In chapter 4, the models are fit to new image data and the use of classical statistics and discriminant analysis to investigate shape differences is explored. In the final chapter we draw conclusions and describe areas of future research.

In this chapter we will provide a brief introduction to T_1-weighted MR imaging, including bias field effects and partial voluming. We also review neuroanatomy as it relates to the problem of automated segmentation of subcortical structures. This will be followed by a general introduction to the existing image segmentation methodologies that are relevant to the work proposed in this thesis.

1.1 T_1-weighted MR Imaging

Magnetic Resonance Imaging (MRI) is a modality that uses electromagnetic fields to obtain high resolution in-vivo images. MRI measures the radio-frequency (RF) waves emitted by precessing (spinning) protons when exposed to a magnetic field. The measured signal is a bulk effect of all the protons contained within a region. In particular, MRI is sensitive to the protons in hydrogen nuclei within tissue. In the absence of a magnetic field the spins are randomly oriented such that there is no bulk effect. By applying a static magnetic field the spins align with the magnetic field such that there is a bulk alignment in the direction of the magnetic field. The frequency of precession (Larmor frequency) is proportional to the strength of the magnetic field. Applying a spatially varying magnetic field gradient provides spatial encoding by frequency. Then by applying an RF pulse perpendicular to the static magnetic field at the Larmor frequency, the net magnetisation is tilted away from the direction of the magnetic field. The spins will realign to the direction of the magnetic field (this process is called relaxation), emitting their energy as RF signals at the Larmor frequency. The measured signal depends on the proton density, the T_1 relaxation constant, the T_2 relaxation constant, and the relative timing of external RF pulses and field gradients. T_1 is a measure of the time required for the net magnetization to return to equilibrium. T_2 is a measure of the time required for the transverse magnetization to return to zero (due to dephasing of the individual protons). Proton density, T_1 and T_2 are all tissue-specific properties. MRI pulse sequences can be designed to be sensitive to any of these properties, resulting in T_1-weighted, T_2-weighted or proton density images. In this work we only consider T_1-weighted images; a T_1-weighted image of a brain is depicted in figure 1.1.

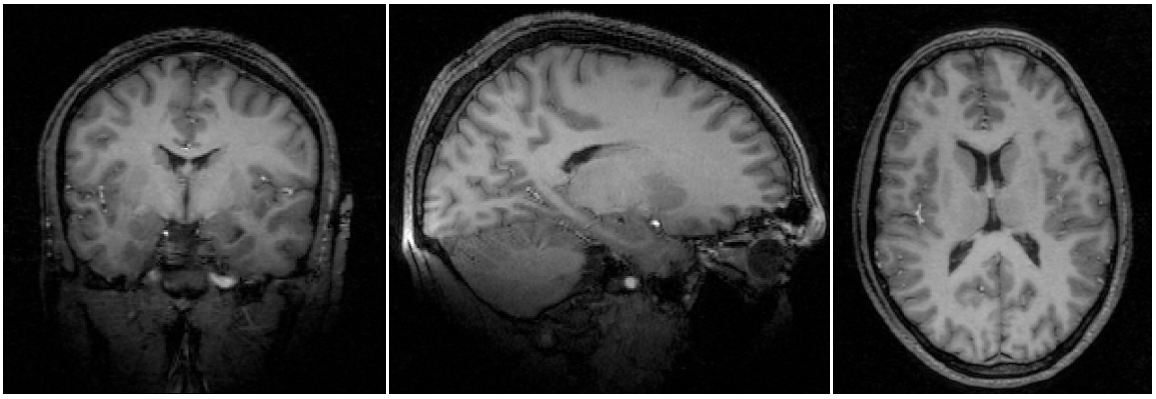


Figure 1.1: A T_1-weighted image of the brain. White matter is the bright tissue in the image and is surrounded (along the outside edge of the brain) by cortical grey matter (dark grey). The subcortical structures are located towards the centre of the image and vary in intensity from as dark as the cortical grey to nearly as bright as the white matter. The very dark area at the centre of the image is CSF.

1.1.1 Bias Field

Bias field effects have the appearance of slow intensity drifts in the image. This artifact is clearly visible in figure 1.1, where the absolute scale of the intensity decreases in the inferior and superior directions from the centre of the image. The intensity drift is an artifact of the imaging rather than a reflection of the underlying tissue. The bias field is principally due to inhomogeneities in the RF field across the image acquisition space. The variation in the RF signal will produce variation in the spin tilts and in the sensitivity to detect the emitted RF signal. The bias field typically has a negative effect on segmentation algorithms since it produces artifactual variations in intensity that, unless modelled appropriately, would be assumed to be variation in the tissue. Segmentation algorithms have been proposed that model and estimate the bias field in an iterative process within the algorithm [62, 2, 52]. Alternatively, the bias field may be estimated and image corrected prior to further processing [55].

1.1.2 Partial Volume Effects

Partial voluming is a result of forming a discrete image representation of a continuous space. The image intensity at a voxel is a reflection of the tissue within the voxel region. Partial voluming is the result of a mixture of tissues within the voxel region. The image intensity resulting from a mixture of tissues within a voxel is dependent on the properties of the tissues as well as the proportion of the voxel volume that is occupied by each tissue. This is of particular relevance since we are training and validating our models using manual segmentations that use a hard labeling (they do not consider partial voluming), particularly since our segmentation output has continuous vertex coordinates. To illustrate the partial voluming effect, figure 1.1.2 shows a synthetic example where the underlying tissue is either black or white (left). The red grid represents the image grid, such that the underlying tissue is on the left and the "imaged" tissue is on the right; note the boxes that contain half black tissue and half white tissue, the resulting image intensity is the mean of the two (grey). For example, in practice, rather than a very distinct boundary between the CSF of the ventricle and the grey matter of the caudate there is a gradual intensity change at the boundary due to partial voluming. Another example of partial voluming is in the posterior horns; at times portions of the horns become so thin that they virtually disappear. Consequently, for some subjects there exists a disconnect in the manual labels for the posterior horns.

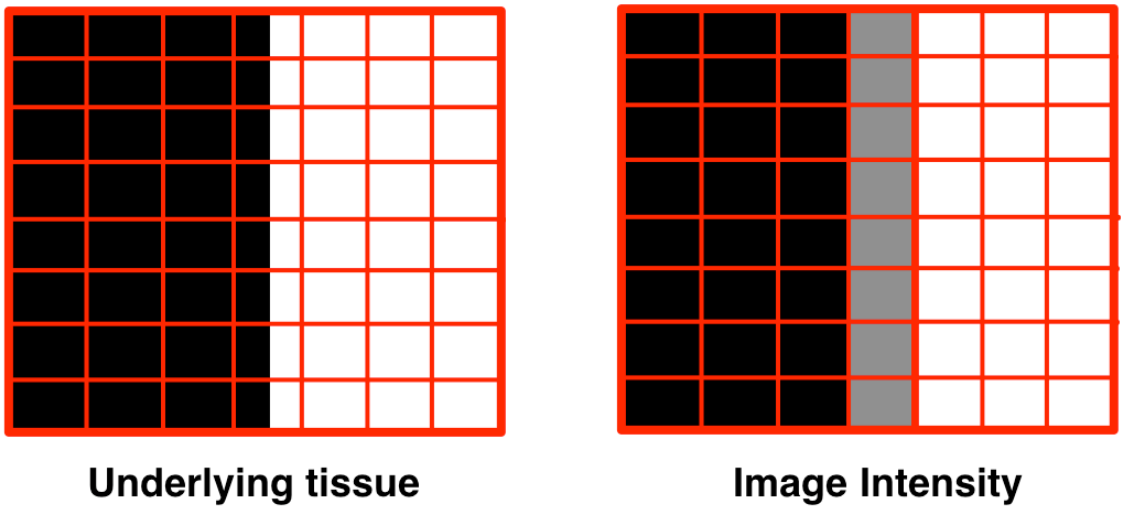


Figure 1.2: An illustrative example of partial voluming. Left: Underlying tissue (black and white) and the image grid (red). Right: Imaged tissue and the image grid (red). For the centre column of the image grid, the half and half mixture of black and white produces grey.

1.2 Neuroanatomy

The focus of this section is on providing a background in the neuroanatomy that underlies the images that we are attempting to segment. We will briefly review the basic cell types that make up the central nervous system (CNS); and then follow up with a discussion of the structural divisions of the brain that we are modelling/segmenting (subcortical structures). In particular, we will focus on their structural boundaries and how they pertain to segmentation.

The brain and spinal cord make up the central nervous system (CNS). The cells within the CNS are categorized into two major types, neurons and glial cells. The neurons are the information-processing units of the CNS, to do so they make use of chemical and electrical signalling mechanisms. Figure 1.3 is an illustration of a

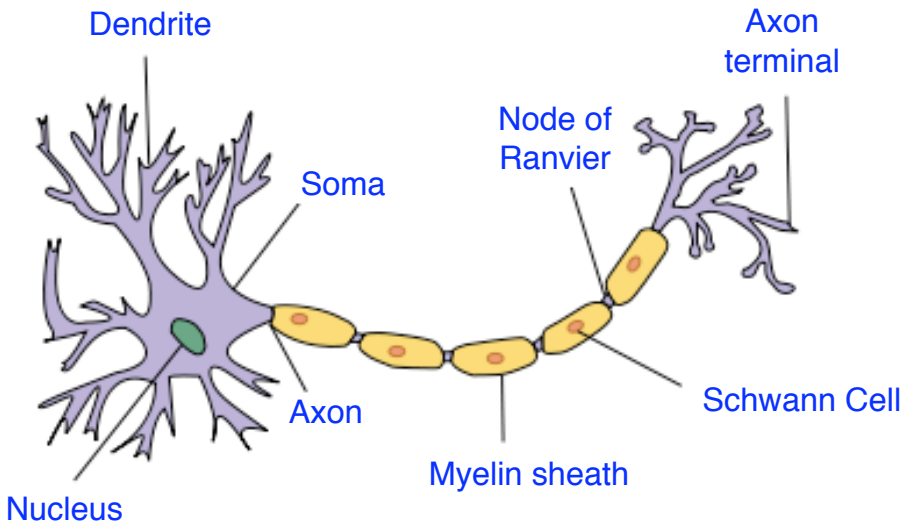


Figure 1.3: Illustration of a typical neuron. (Image courtesy of Wikimedia Commons).

typical neuron. All neurons have a cell body (soma) which handles the metabolic and synthetic needs of the entire neuron. The dendrites are used to receive signals from other neurons (via chemical signalling). The information received from the dendrite is then transmitted electrically from the cell body, along the axon, to the axon terminals. The axon terminals are used to transmit the received signal to other neurons (via chemical signalling). Some axons are covered by a myelin sheath that is comprised mostly of lipids, giving it a fatty white appearance. The number of myelinated axons within a voxel-region serves to modulate the image intensity in T_1-weighted MR images. Using a T_1-weighting, the number of the myelinated axons is positively correlated with the image intensity; this gives a white appearance to tissues with high concentrations of myelinated axons, such as white matter.

The glial cells form a large proportion of the cells in the CNS and in fact exceed the number of neurons. The glial cells surround the neurons in the CNS and serve to

hold them in place, to insulate them from each other, as well as to provide nutrients and oxygen to the neurons. The glial cells also destroy pathogens and remove dead neurons.

At the resolutions available with T_1-weighted MR imaging, we are unable to resolve individual neurons or glial cells, instead the measured signal is that from large bundles of cells. The CNS may be roughly divided into white matter and grey matter. The grey matter is comprised mostly of neuronal cell bodies and dendrites, whereas white matter is comprised mostly of axons (the myelin sheath that covers many of the axons gives the white appearance). Also contained within the brain is cerebrospinal fluid (CSF), which is circulated via the ventricular system and mechanically supports the brain, preventing it from collapsing under its own weight. Most methods for tissue segmentation of MR images divide the brain into white matter, grey matter and CSF. Figure 1.4 shows manually labelled white matter, grey matter and CSF on a T_1-weighted image. Included in the CSF is the choroid plexus which has the appearance of white tissue within the ventricles. The choroid plexus is tissue that is responsible for the production of CSF in the brain.

We will draw a distinction between the cortex, white matter and subcortical structures. The cortex is a grey matter layer that covers the outer brain surface, over the white matter tracts in brain, and is responsible for high-level functions. The white matter comprises the connective pathways between the various processing regions (i.e., grey matter). The terminology "subcortical structures" as used in this document is referring to subcortical grey matter structures. The subcortical structures are groupings of grey matter within the brain that are not included as part of the cortex. The exception to this is our inclusion of the lateral ventricles (CSF and

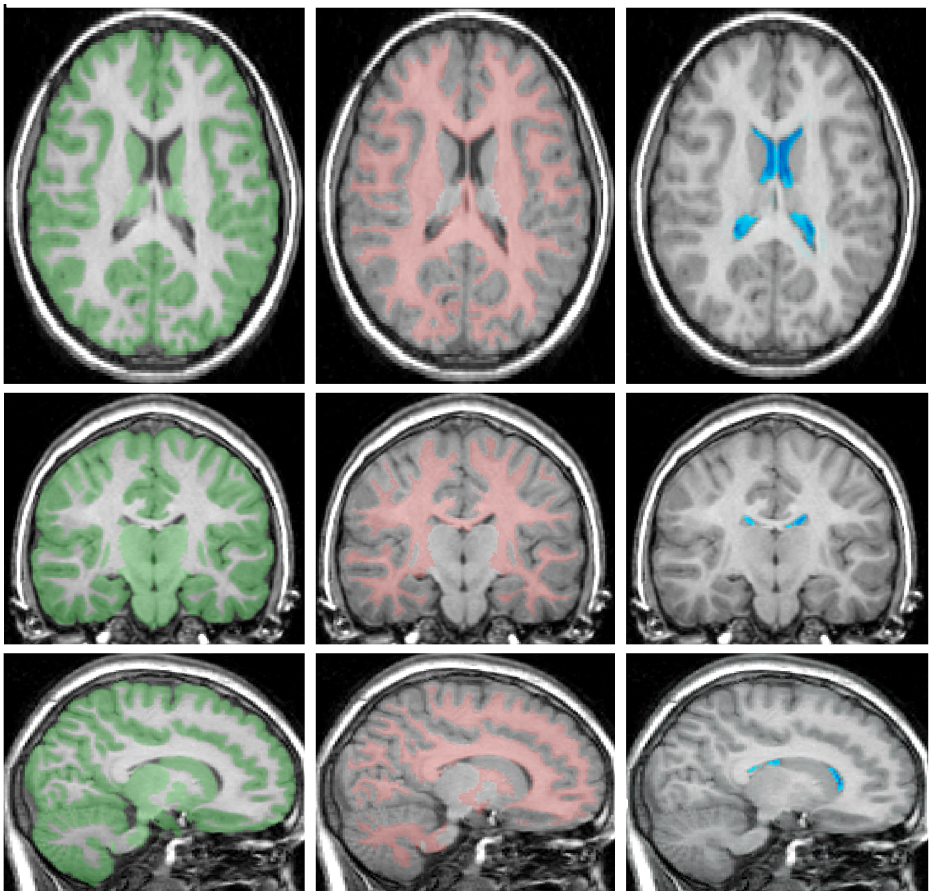


Figure 1.4: Manually labelled grey matter, white matter, CSF (as well as choroid plexus) of a T_1-weighted MR image.

choroid plexus) and the hippocampus in our definition of subcortical structures; as the hippocampus is normally considered a cortical structure. The subcortical structural divisions also correspond to functional groupings of the grey matter. For the purpose of this discussion we divide the subcortical structures being modelled into the brainstem, thalamus, ventricles, basal ganglia and limbic system.

To convey the shape of each structure and their surrounding structures, a coronal, sagittal and axial slice from a single subject of our training set will be shown for each structure. In addition, the mean surface for each structure will be shown. The mean surfaces are derived from the models that were constructed using the methodology that will be described in chapters 2 and 3. Figure 1.6 is the colour map for all structures in the images shown below; "the ventral DC" refers to an amalgamation of several structures[1] that are difficult to discern from each other in T_1-weighted image. In each image the cross-hair will be located within the structure of interest.

For reference when discussing the subcortical structures, figure 1.5 depicts three slices in each of the superior-inferior, medial-lateral, and posterior-anterior directions. The left column for figures 1.5(a), 1.5(b), and 1.5(c) contains a stationary slice, in which the cross-hair indicates the location of the corresponding slice in the right column. The figure also defines the superior, inferior, medial, lateral, posterior, and anterior directions with respect to the brain. Prior to discussing the subcortical anatomy we will briefly discuss the manual segmentation protocol that was used.

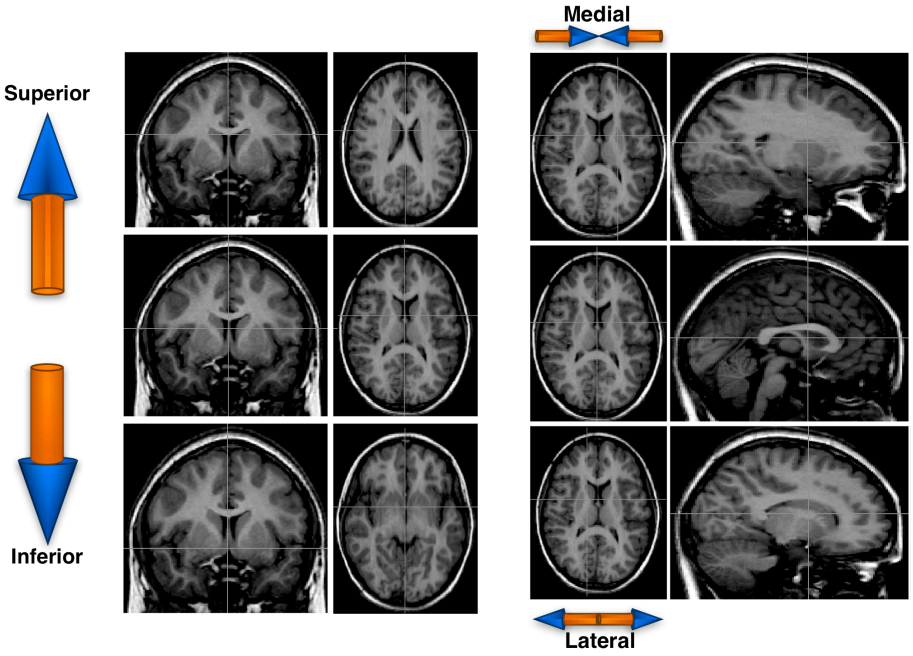
[1]The ventral DC includes the hypothalamus, mammillary body, subthalamic nuclei, substantia nigra, red nucleus, lateral geniculate nucleus (LGN), and medial geniculate nucleus (MGN). Also included in the ventral DC are white matter areas such as the zona incerta, cerebral peduncle (crus cerebri), and the lenticular fasciculus. The optic tract is also included in the most anterior extent of the ventral DC.

1.2.1 Manual Segmentation Protocol

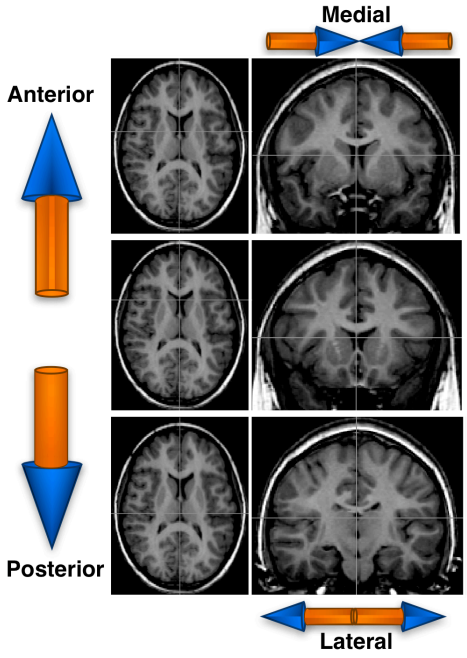
The manual segmentations were performed by a highly trained set of operators at the Centre for Morphometric Analysis (CMA). A brain image requires approximately two to three days to be segmented by a human operator. Operators are trained until they have reached a defined reproducibility; the training process typically lasts about three months. The brain is segment in sequential coronal slices, and as a result the boundary in the posterior-anterior direction is rougher than for the other directions. A semi-automated intensity-based contouring tool is used to aid and quicken the segmentation process; intensity gradients are used to initialize specific boundaries. For regions where there is poor contrast in the coronal slice, guidelines are placed in the posterior-anterior direction. The segmentation protocol is based on intensity boundaries, as well as geometrical rules based on neuroanatomy. The protocol will draw on geometrical rules for boundaries that lack contrast (e.g., the caudate/accumbens boundary).

1.2.2 The Brainstem

The brainstem may be subdivided into the medulla, pons and midbrain, however the delineations between the three divisions is typically poor in T_1-weighted MR images. Consequently, the manual labels that were provided as training data consider the brainstem as a whole and thus our model considers the brainstem as a whole. Generally our structural models are dictated by the definition of the structure used by the manual segmentation protocol. Figure 1.2.2 depicts three views of the manually



(a) Inferior-Superior direction. (b) Medial-Lateral direction



(c) Posterior-Anterior direction

Figure 1.5: From a T_1-weighted image, three slices in each of the superior-inferior, medial-lateral, and posterior-anterior directions.

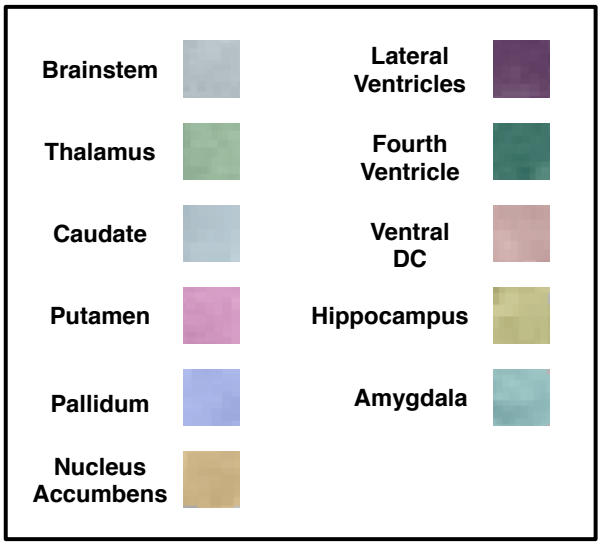
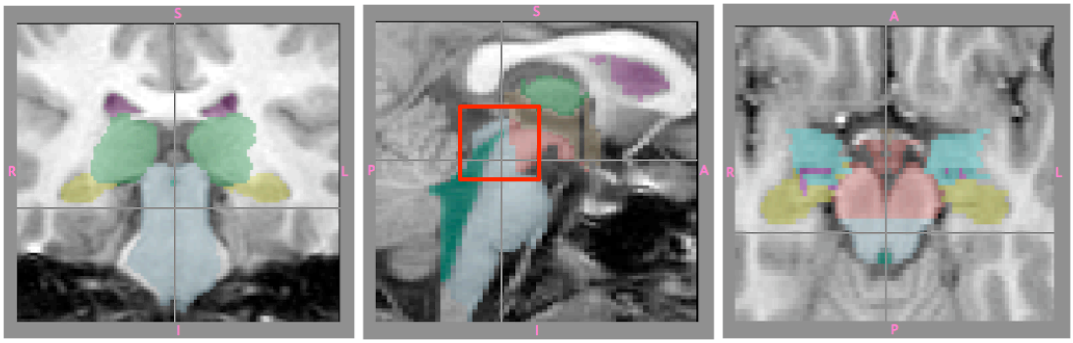


Figure 1.6: Colour map used for the subcortical structures shown in the figures to follow.

labelled brainstem, and figure 1.7 is the mean surface. The mean surface depicted is actually the combined fourth ventricle and brainstem, the reason for which will be discussed briefly in the next paragraph.

The brainstem borders the ventral DC, the fourth ventricle, and the cerebellum (both the ventral DC and the cerebellum are not modelled). Since the fourth ventricle is combined with the brainstem, the combined model does not border any of the other subcortical structures that are being modelled. For the most part the brainstem has clearly defined boundaries in T_1-weighted images. The main challenge from a shape model/segmentation perspective is the fourth ventricle. The fact that the fourth ventricles passes directly through the brainstem (see figure 1.7(b)) makes the topology more complex than for other structures, this poses challenges in training the models (this is the reason for combining the structures). The other problematic region is



(a)

(b)

Figure 1.7: a) Coronal, sagittal and axial slices of the brainstem from a single subject's T_1-weighted image showing the manual labelling. b) Depicts the fourth ventricle passing through the brainstem; the enlarged area of the image is defined by the red box in a).

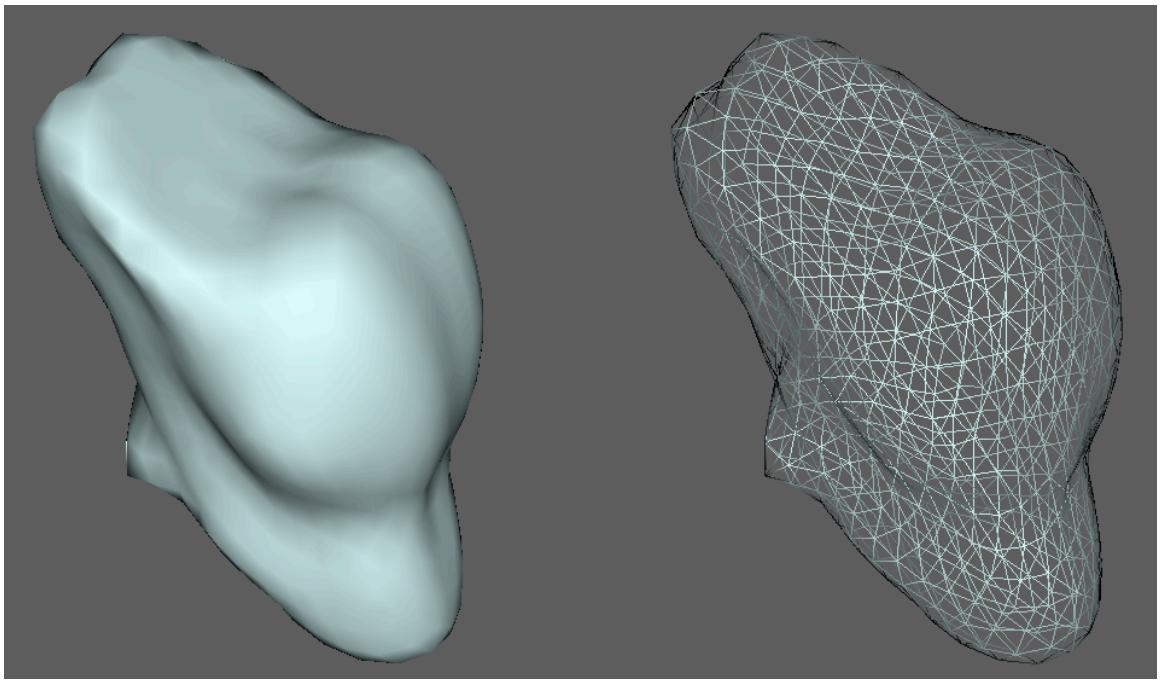


Figure 1.8: Mean surface for the brainstem. Left: Surface rendering. Right: Mesh view of surface.

at the superior end of the brainstem, the boundary with the ventral DC has very poor contrast and as a result was artificially created by observing geometrical rules in the manual segmentation protocol. The brainstem/ventral DC boundary would be virtually impossible to define based on intensity alone.

1.2.3 The Thalamus

The thalamus may be thought of as a relay station for the brain; it serves as a communication hub between various regions of the brain. Figure 1.9 depicts the left thalamus (the right thalamus is also viewable in the coronal and axial slices) for a single subject, and figure 1.10 shows the mean surface. Of the modelled structures,

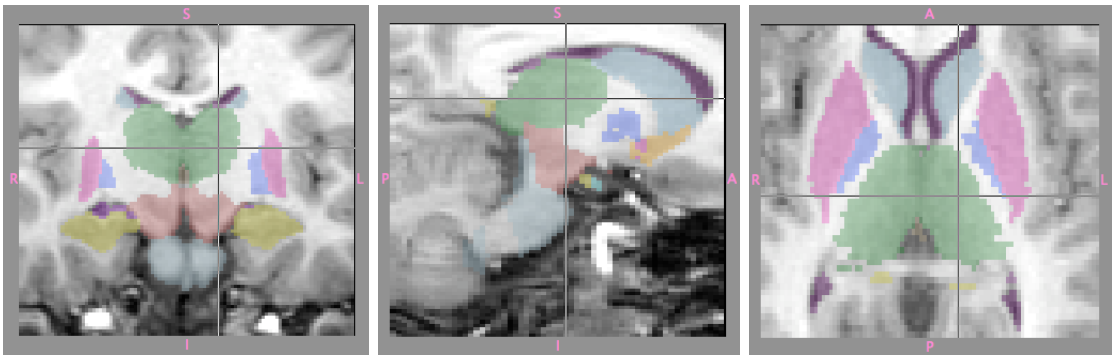


Figure 1.9: Coronal, sagittal and axial slices of the left thalamus from a single subject's T_1-weighted image showing the manual labelling.

the thalamus borders the lateral ventricle, the hippocampus and the caudate. There is a large contrast on the medial boundary with the lateral ventricles. In addition, there is reasonable contrast with the caudate and hippocampus. The tissue contrast between the lateral portion of the thalamus and the neighbouring white matter is usually quite poor in T_1-weighted images and thus poses a challenge to automated segmentation methods. This is an ideal example of where, by modelling shape, we may be able to draw on the well-defined boundaries (and prior shape training) to aid in determining the low-contrast lateral border of the thalamus.

1.2.4 The Basal Ganglia

The basal ganglia refers to a set of structures whose damage results in "extrapyramidal" syndromes (movement disorders) [47]. The term basal ganglia typically includes the caudate nucleus, putamen, pallidum, nucleus accumbens, subthalamic nuclei and substantia nigra. Of these, we model the caudate nucleus, putamen, pallidum, nucleus accumbens, and the thalamus as a whole (we do not provide divisions of the

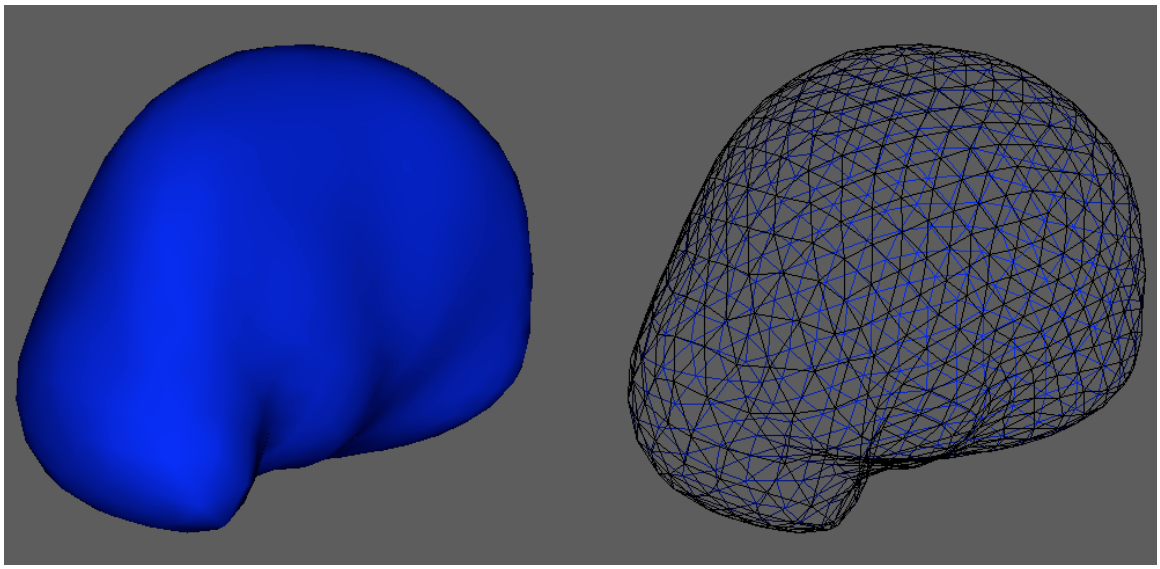


Figure 1.10: Mean surface for thalamus. Left: Surface rendering. Right: Mesh view of surface.

thalamus). The structural boundaries of the caudate, putamen, pallidum and nucleus accumbens will now each be described separately.

The Caudate

The manual segmentation protocol for the caudate truncates the further reaches of the tail. The tail is the more posterior end of the caudate. The tail becomes so thin that, given the resolution, it cannot be readily identified because of partial voluming. The truncation point is typically in a region just superior to the thalamus; this corresponds to the point at which the partial voluming makes the labelling unreliable. The tail is a long thin region of the caudate, which for our purpose is terminated in a region just superior to the thalamus. In actuality the caudate is a C-shaped structure and extends much further, wrapping around in the brain. The manually labelled caudate

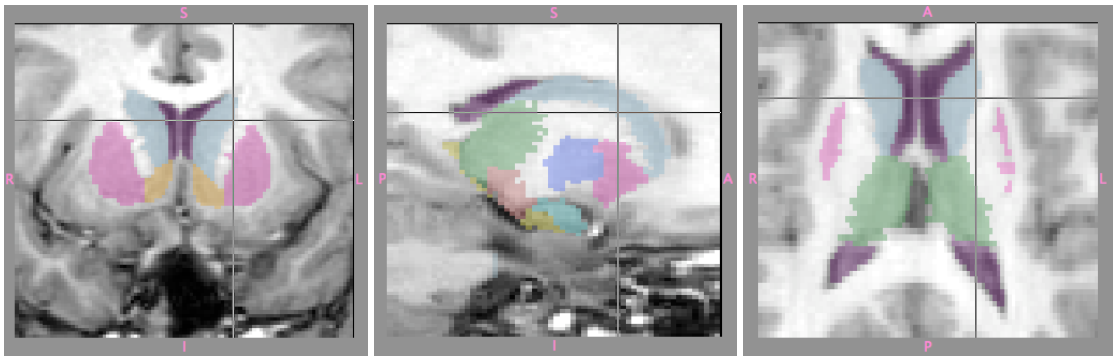


Figure 1.11: Coronal, sagittal and axial slices of the left caudate from a single subject's T_1-weighted image showing the manual labelling.

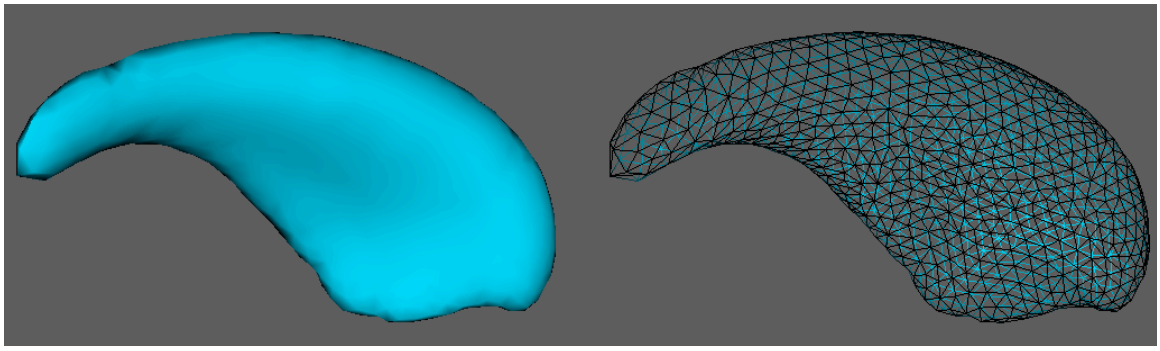


Figure 1.12: Mean surface for the left caudate. Left: Surface rendering. Right: Mesh view of surface.

is viewable in figure 1.11.

Of the structures modelled, the caudate borders the lateral ventricles, the thalamus and the nucleus accumbens. As seen in figure 1.11 the portions of the superior region of the caudate appears to border white matter; this is because the lateral ventricles are so small that the thin layer of CSF that lies between the caudate and white matter is partial volumed. However, it is also most frequently the case that there is clearly CSF between the two. From a segmentation perspective, this means that the expected intensities in that region would vary with size and shape of the caudate. This

is a good example of how a shape-and-appearance model may benefit a segmentation algorithm; the relationship between shape and intensity may be modelled *a priori* from the data. For this example, we wish to learn the fact that as the ventricles expand, we expect the bordering intensities to reflect the increase in bordering CSF.

A problematic region of the caudate is its inferior boundary with the accumbens, as there is very little contrast there in T_1-weighted MR images. In fact, the manual segmentation relies partly on geometrical rules based on anatomical knowledge. This boundary is another example where a good shape model should rely on the high-contrast boundaries to help define the low-contrast ones.

Putamen

The putamen is part of the lenticular nucleus which is composed of the putamen and the pallidum. The putamen generally has good contrast with its neighbouring tissue. Of the structures modelled, the putamen borders the nucleus accumbens and the pallidum. Its contrast with the accumbens is poor; however that boundary is a small portion of the entire boundary for the putamen. Frequently cortical grey matter comes in close proximity to the posterior/lateral border of the putamen; as a result the boundary between the putamen and the cortex may be difficult to determine without considering shape. Figure 1.13 depicts the putamen and its neighbouring structures, and figure 1.14 depicts the mean surface.

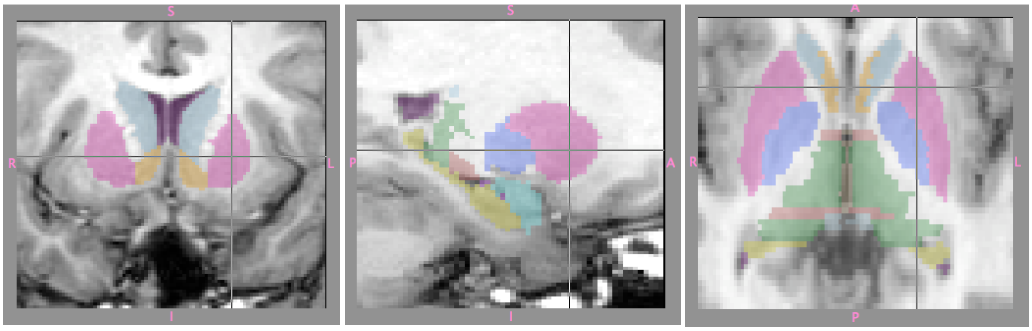


Figure 1.13: Coronal, sagittal and axial slices of the left putamen from a single subject's T_1-weighted image showing the manual labelling.

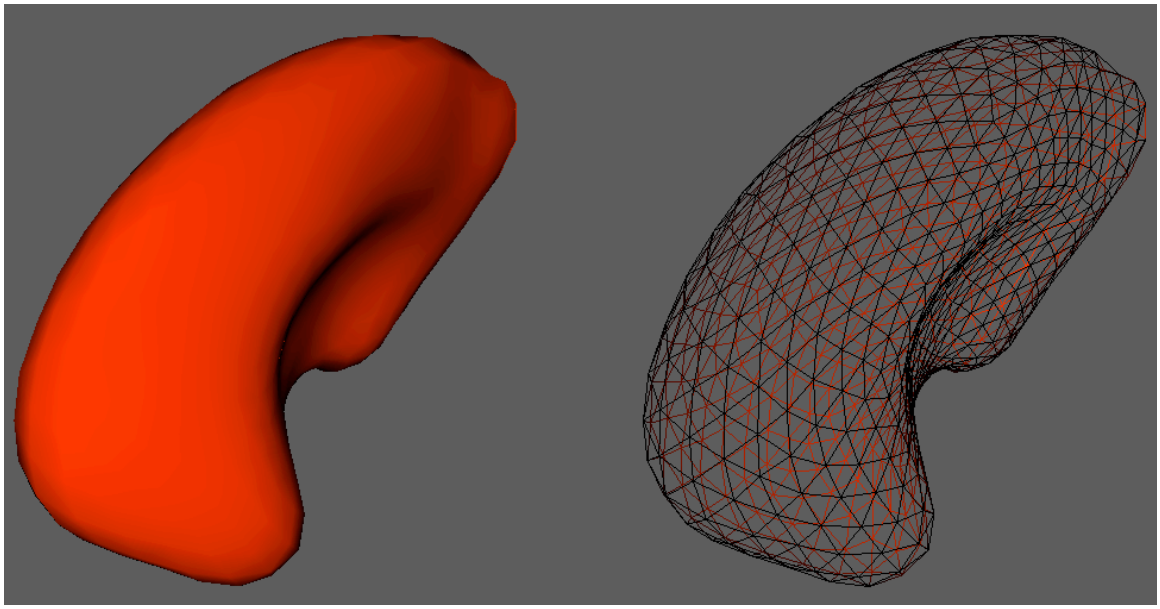


Figure 1.14: Mean surface for the left putamen. Left: Surface rendering. Right: Mesh view of surface.

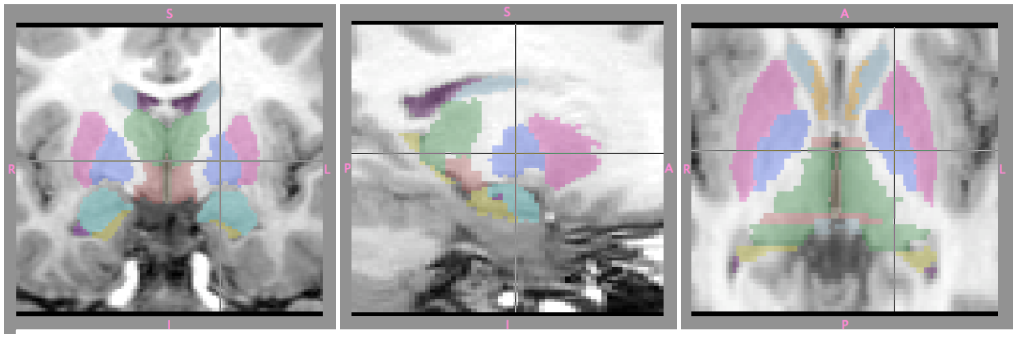


Figure 1.15: Coronal, sagittal and axial slices of the left pallidum from a single subject's T_1-weighted image showing the manual labelling.

Pallidum

Of the structures modelled, the pallidum only borders the putamen. The remaining boundary is with white matter tracts, although the thalamus is located medially across the white matter. Its contrast with the putamen is good, however, its contrast with the white matter is often very poor in T_1-weighted images. As a result the medial border poses a significant challenge to automated segmentation methods. Figure 1.15 depicts the pallidum and its neighbouring structures, and figure 1.16 depicts the mean surface.

To further illustrate the difficulties associated with purely intensity-based segmentation, figure 1.17 depicts an axial slice (containing the lenticular nucleus and the thalamus) of a T_1-weighted image and the grey matter tissue segmentation produced by FMRIB's Automated Segmentation Tool (FAST) [62]; the T_1-weighted image is the same as that depicted in figure 1.1. FAST has clearly misclassified the entire lenticular nucleus and thalamus. In addition to having difficulty in differentiating

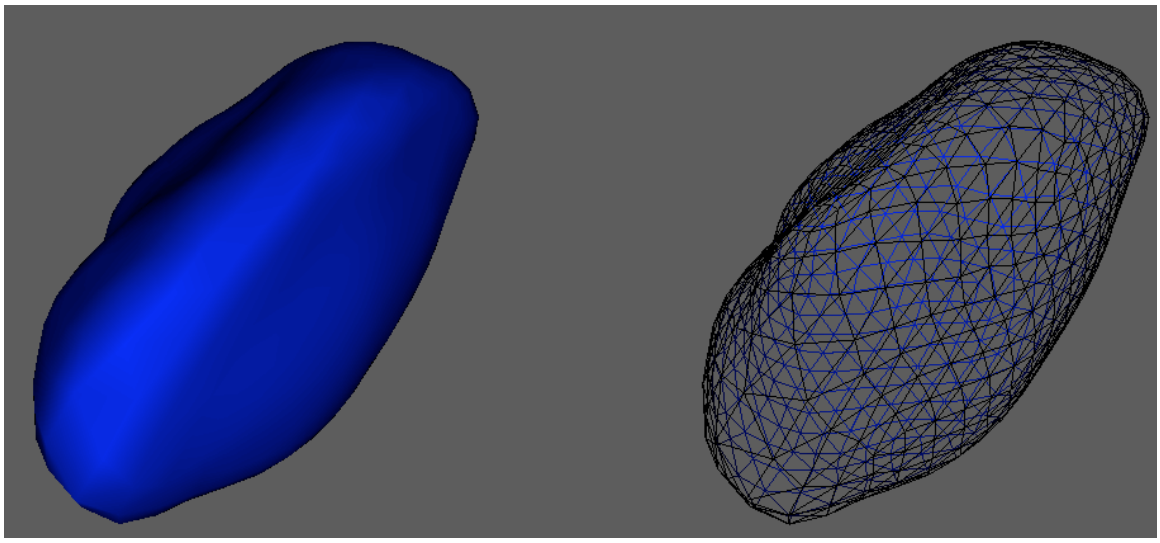


Figure 1.16: Mean surface for the left pallidum. Left: Surface rendering. Right: Mesh view of surface.

tissues with weak contrast (e.g. the pallidum and white matter), the intensity-based segmentation is not capable of differentiating between the bias field and the gradients in grey matter density that exist in subcortical regions. Furthermore, the use of a static prior such as is used in Statistical Parametric Mapping's (SPM[2]) tissue segmentation [1] would merely bias the results towards the tissue prior; the tissue prior is the mean of a sample set taken from a specific demographic. Static priors do not account for normal variation in the population within or across demographics, nor do they consider variation due to pathology. Our shape-and-appearance models preserve the variance information through the use of eigenvectors and eigenvalues. Figure 1.17 also depicts the segmentation output from our shape-and-appearance model, the results of which are reasonable, despite not explicitly modelling or correcting for bias field effects.

[2]SPM is a widely used software package for the analysis of neuroimaging data.

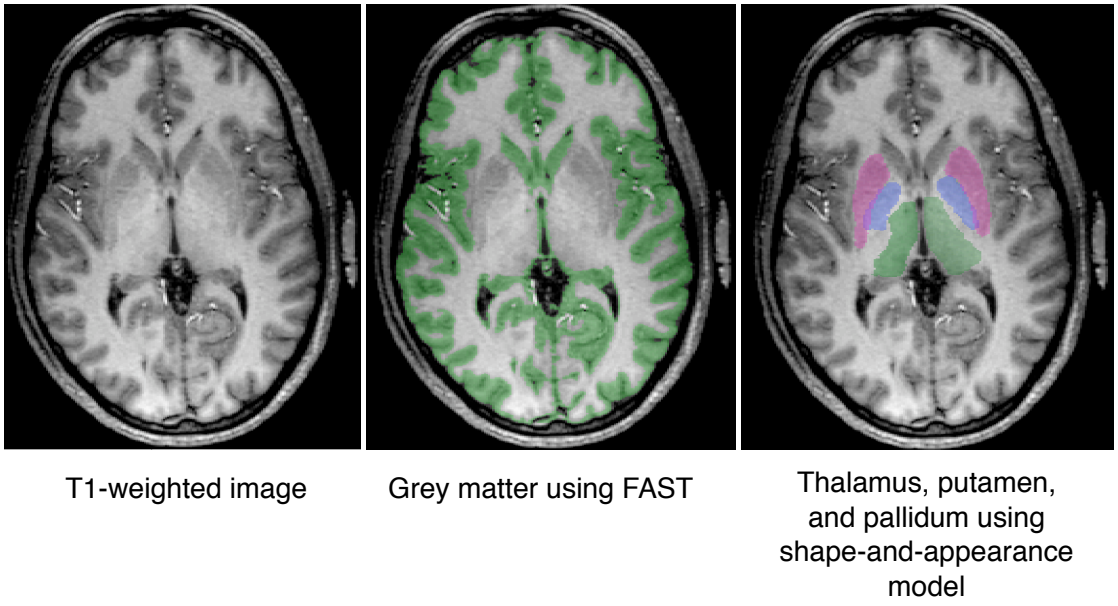


Figure 1.17: Grey matter segmentation for a T_1-weighted image; a) An axial slice of a T_1-weighted image containing the thalamus, putamen, and pallidum; b) FAST grey matter segmentation (green); c) Shape-and-appearance model-based segmentation.

Nucleus Accumbens

The nucleus accumbens is a small structure that is situated between the caudate and the putamen; it is located at the head of the caudate and at the anterior end of the putamen. There is minimal contrast with the caudate and the putamen; given that these boundaries make up a good portion of the accumbens' total boundary, segmentation of this structure is difficult. Its white matter boundaries do provide good contrast and could potentially be used in combination with a shape model to achieve an accurate and robust automated segmentation. One caveat is that the accumbens may be difficult to manually segment and thus the shape model is subject to the manual definition and its variability. Figure 1.18 depicts the nucleus accumbens and its neighbouring structures, and figure 1.19 depicts the mean surface.

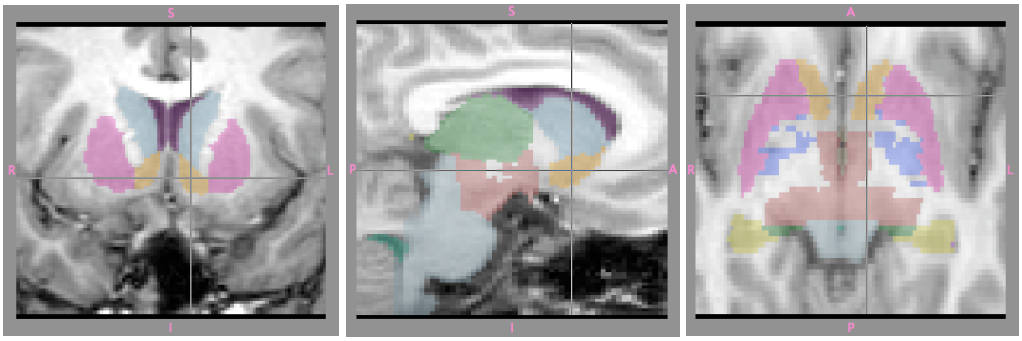


Figure 1.18: Coronal, sagittal and axial slices of the left nucleus accumbens from a single subject's T_1-weighted image showing the manual labelling.

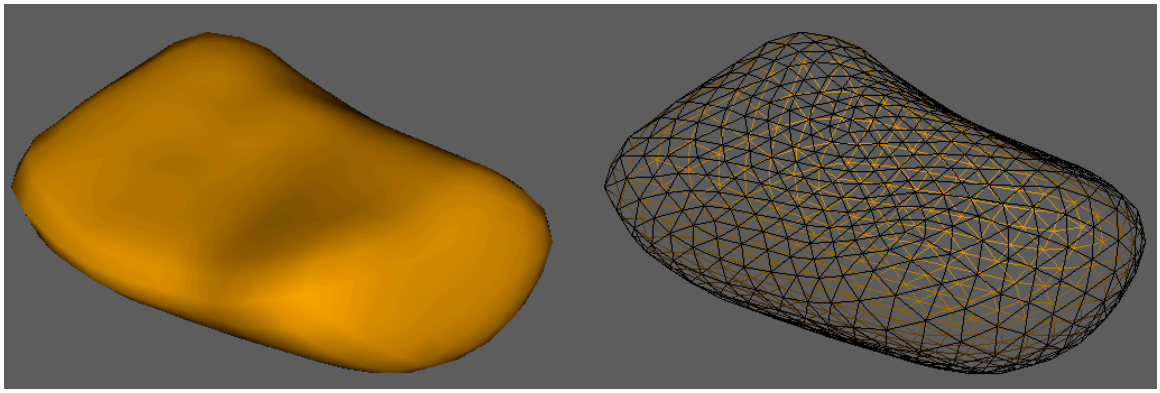


Figure 1.19: Mean surface for the left nucleus accumbens. Left: Surface rendering. Right: Mesh view of surface.

1.2.5 Limbic System

The limbic system is associated with emotion, memory and motivation and is comprised of cortical and subcortical structures. Of the structures belonging to the limbic system, we model the hippocampus, amygdala, and nucleus accumbens. The nucleus accumbens was discussed previously as it also belongs to the basal ganglia.

Amygdala

The posterior border of the amygdala is within the hippocampus. The contrast may be quite poor at this border, and there is typically a very small amount of CSF that lies between the amygdala and the hippocampus. Unfortunately, the CSF is not very easily distinguished due to partial voluming, however it is still frequently used as a marker of the boundary. In cases of severe atrophy the amount of CSF between the hippocampus and amygdala may become quite significant. This is another good example where a shape-and-appearance model may be useful to learn the shape changes that will correlate with the presence of CSF at the posterior border.

The anterior border of the amygdala is in close proximity with cortical grey matter and it may be difficult to determine the boundary between the two. The superior border does have good contrast with the neighbouring white matter. Figure 1.20 depicts the left amygdala and its neighbouring structures, and figure 1.21 depicts the mean surface.

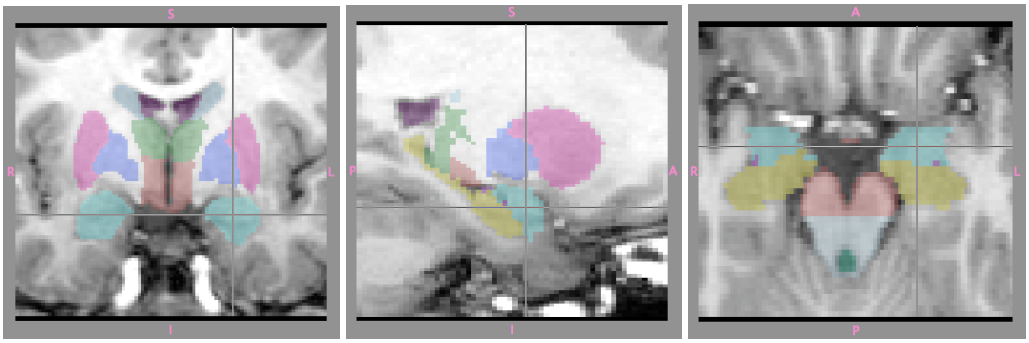


Figure 1.20: Coronal, sagittal and axial slices of the left amygdala from a single subject's T_1-weighted image showing the manual labelling.

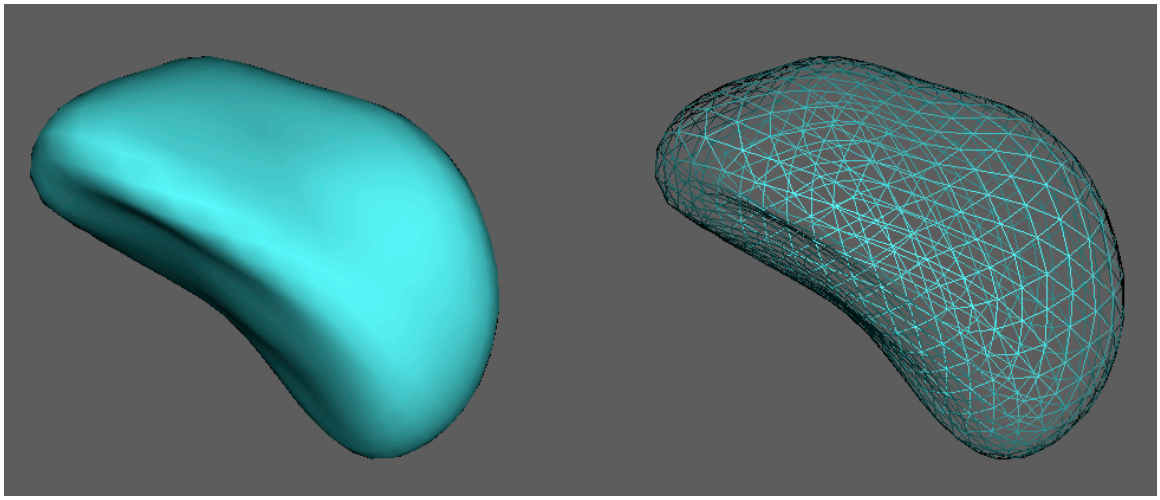


Figure 1.21: Mean surface for the left amygdala. Left: Surface rendering. Right: Mesh view of surface.

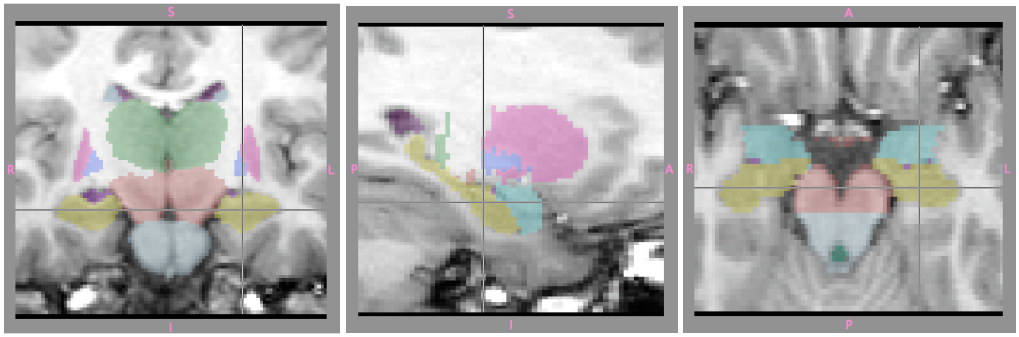


Figure 1.22: Coronal, sagittal and axial slices of the left hippocampus from a single subject's T_1-weighted image showing the manual labelling.

Hippocampus

Of the structures modelled, the hippocampus borders the lateral ventricles, the thalamus and the amygdala. The contrast between the ventricles and the hippocampus is good, however the boundary is small (the superior end of the tail of the hippocampus). There is not much contrast with the thalamus, but again the boundary is small. The posterior boundary of the hippocampus has good contrast with the neighbouring white matter. Overall the borders of the hippocampus are fairly well defined, and so the most significant challenge with segmenting the hippocampus is its complex shape and the large amount of variation therein.

1.2.6 Ventricular System

The ventricular system mechanically supports the brain, preventing it from collapsing on itself due to its own weight. It also allows for changing brain size across the cardiac

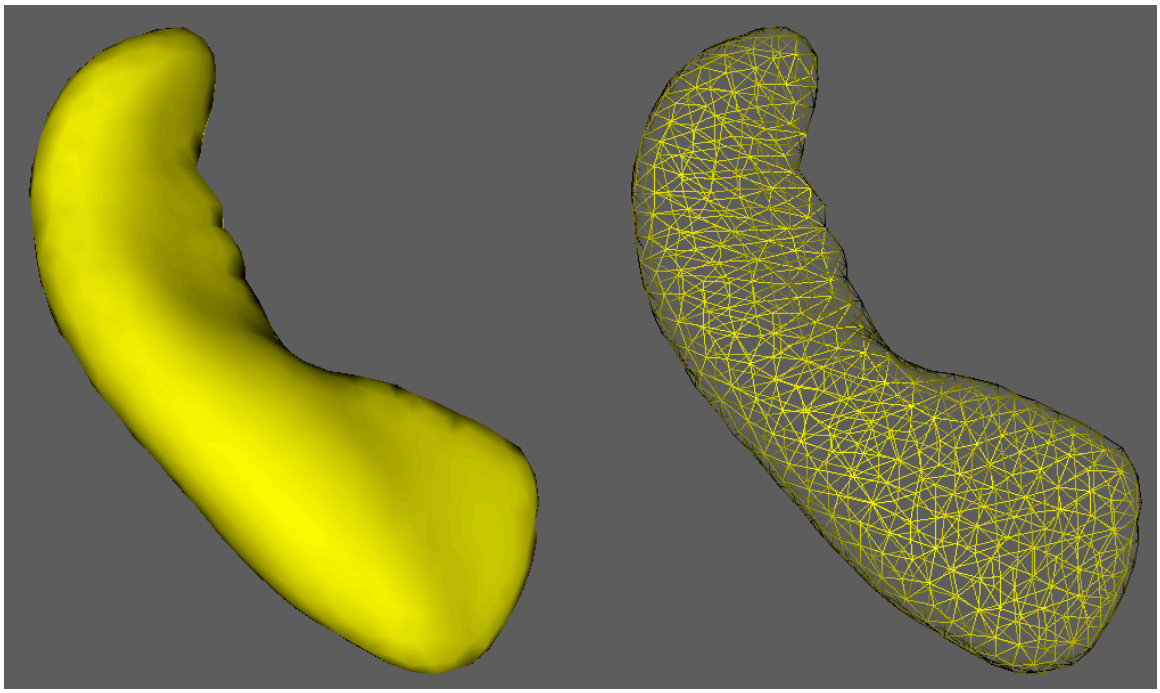


Figure 1.23: Mean surface for the left hippocampus. Left: Surface rendering. Right: Mesh view of surface.

cycle due to vascular flow. As blood flows in, the brain expands and the CSF flows out. Of the ventricular system we model the lateral ventricles (the fourth ventricle was combined with the brainstem). There is a large variation in size of the ventricles across the population, and they also increase in size with age. The lateral ventricles contrast well with the other subcortical structures since the ventricles are made up of CSF.

Contained within the ventricles is the choroid plexus; the choroid plexus is responsible for the production of CSF. The inclusion of a heterogeneous tissue (it is made up of glial cells) within the CSF may pose a challenge to many automated segmentation schemes because of the large difference in intensity between the CSF and choroid plexus. Despite its high-contrast boundaries, the lateral ventricles are challenging

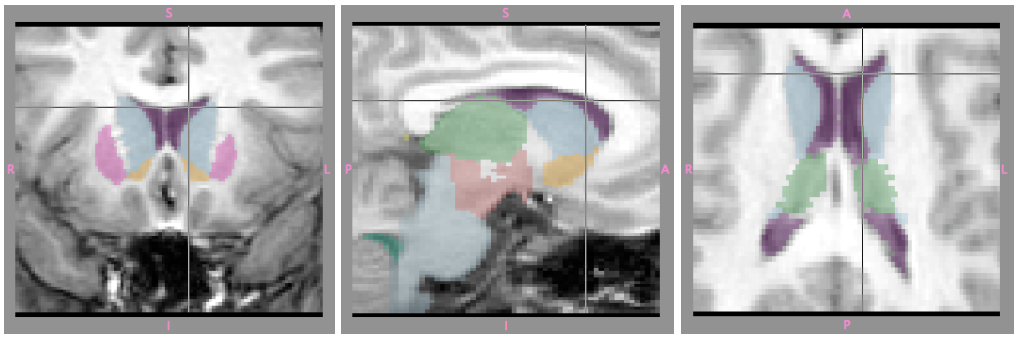


Figure 1.24: Coronal, sagittal and axial slices of the left lateral ventricles from a single subject's T_1-weighted image showing the manual labelling.

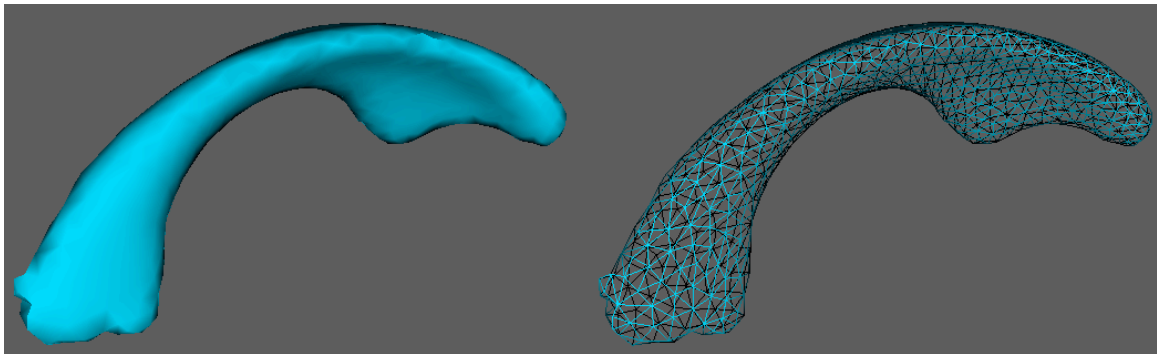


Figure 1.25: Mean surface for the left lateral ventricles. Left: Surface rendering. Right: Mesh view of surface.

structures to fit due to their complex geometry and the large variation in size and shape. Of the structures modelled, the lateral ventricles border the caudate, and the thalamus, as well as a small superior portion of the hippocampus. Figure 1.24 depicts the lateral ventricles and its neighbouring structures, and figure 1.25 depicts the mean surface.

1.3 Registration

We will briefly discuss registration since it is referred to periodically in the chapters to follow. Registration is the process of determining voxel-wise correspondence between images. In medical imaging, registration is essential as it allows us to fuse information across subjects, time points, and/or modalities. Furthermore it is necessary for performing voxel-wise statistics across subjects. Registration methods are classified into either being linear or non-linear. In addition, registration is characterized by the number of degrees of freedom (dof) which reflects the amount of constraint in the deformation field (the lower the dof, the more constrained the deformation). For 3D images, linear registrations may contain up to 12 degrees of freedom; which include translations (3 dof), rotations (3 dof), scaling (3 dof), and shears (3dof). Higher dof is attributed to non-linear registration methods, and these result in deformation/vector fields that describe the displacement of the image grid at each point.

Registration may be applied to the problem of segmentation by registering the image to a template image that has corresponding segmentation labels. The deformation fields can then be used to propagate the labels to the desired image. Linear registration is not adequate for such an approach to segmentation because the low degrees of freedom cannot accommodate the structural variation in the population. A limitation of such an approach, using non-linear registration, is bias towards the choice of template.

1.4 Segmentation

This section will provide a general overview of the segmentation methods that pertain to the methods that will be presented in the chapters to follow. The aim of this section is not to provide a rigorous mathematical background, but rather to provide a conceptual understanding of the methods. The main concepts that will be discussed will be the concepts of the deformable model, the active shape model (ASM), and the active appearance model (AAM). A more thorough explanation of deformable models will be provided in chapter 2 since they are the basis of the majority of the methods proposed in that chapter. The ASM and AAM are the conceptual basis from which our model was formulated. A more in-depth review of current segmentation techniques will also be provided in chapter 3.

In the late 1980s, snakes (dynamic contour models) were introduced for 2D image segmentation [35]. The deformable contours model a boundary using a series of connected vertices along the edge of a structure. The deformable contour is analogous to a rubber band where it may stretch and deform to new configurations, and is constrained by the "mechanical" properties of the material. Each vertex is driven by the image intensities, typically being attracted to voxels with large intensity gradients. Using the prior belief that structures are relatively smooth, constraints are placed on the model to limit the curvature of the contour. By using a model for the boundary and enforcing smoothness along the contour, the deformable contour achieves robustness against noise. The discrete model of the boundary also allows easy calculation of boundary metrics such as curvature. The deformation process is an iterative procedure such that the vertices take small steps towards the boundary

at each iteration.

The deformable contour was originally a 2D method. The idea of the 2D deformable contour extends to three dimensions as a deformable surface. Where the deformable contour can be analogized to a rubber band, the deformable surface is analogous to a balloon. The idea behind the deformable surface is the same as that for the contour except that the vertices move in three dimensions and the metrics that govern the deformation are based on 3D quantities. An example of a deformable surface is shown in figure 1.26, where it depicts the shrinking of a sphere onto a manually labelled left amygdala. It shows the progression of the surface during the deformation process. On the left is the initial sphere, the middle is a midpoint (300 iterations), and the right is the final result after 600 iterations. This process will be revisited in chapter 2.

Deformable models rely on local intensity contrast for segmentation and therefore require good initialization. They are also susceptible to having portions of the mesh getting stuck on local gradients that do not correspond to the structure of interest. The flexibility in their deformation combined with local intensity optimization make them susceptible to noise and artifacts. Many of these problems could potentially be overcome using a more explicit model of shape to constrain the deformation; this was the motivation behind the seminal work by Cootes et al. [11] that proposed the ASM.

The ASM introduced the notion of using a Point Distribution Model (PDM) to model shape. As with the deformable model, a shape is represented by a surface which is composed of a set of connected vertices. A PDM models the variation in vertex location over a population by a multivariate statistical distribution. The ASM,

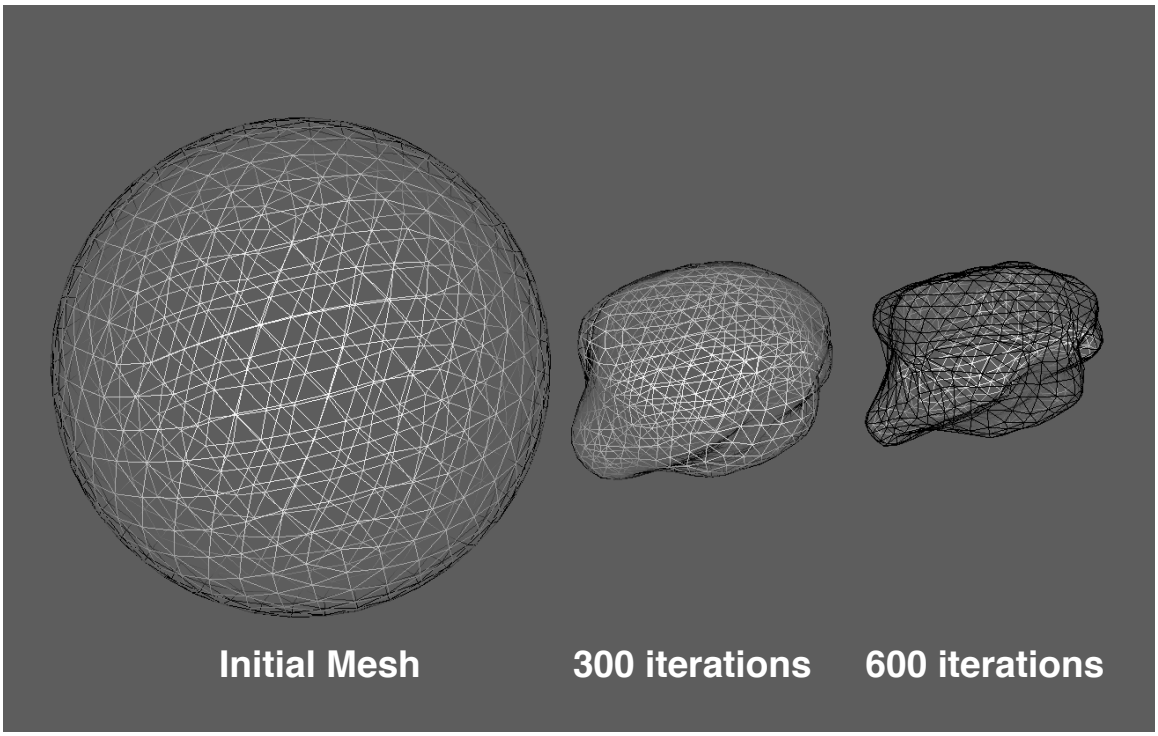


Figure 1.26: Deformation of a sphere to the label image of the left amygdala. Left: Initial surface, middle: 300 iterations, right: end point (600 iterations).

AAM as well as our model assume an underlying multivariate Gaussian distribution for the location of vertices and we will thus restrict our discussion to this case. In order to model vertex location as a multivariate Gaussian distribution, the parameters of the distribution need to be estimated. To estimate the mean and covariance matrix that characterizes the multivariate Gaussian PDM, we need a set of sample surfaces (training data). The PDM assumes that vertices correspond across samples from the population, therefore the training surfaces must have an equal number of vertices with correspondence across the sample set. Figure 1.27 shows four of the 317 training surfaces that were used to construct our model of the left hippocampus.

Based on the PDM that we have estimated from a set of training surfaces, we now have a model for a mean shape (i.e. mean vertex location) and the spatial variation of vertices. Since the vertices were all modelled together using a multivariate distribution, the model retains the information regarding how vertices vary with respect to each other (covariance). A PCA or eigenvector decomposition is a mathematical tool to find the directions of maximum variation. A single eigenvector is a series of 3D vectors (one associated with each vertex) that vary in magnitude from vertex to vertex and describes the displacement direction for each vertex. Along with each eigenvector is a scalar value that is equal to the total variance from the data that is explained by the vertex variation given by the eigenvector. When we attempt to fit the model to a new image we start with the mean shape and only allow the vertices to move along the eigenvectors. In slightly more mathematical terms, we parameterize shape by a linear combination of the mean and eigenvectors (this is the ASM formulation). Given our model, the further out along each eigenvector, the more unlikely it is to find that shape in the population. Therefore, if we only allow vertices to move

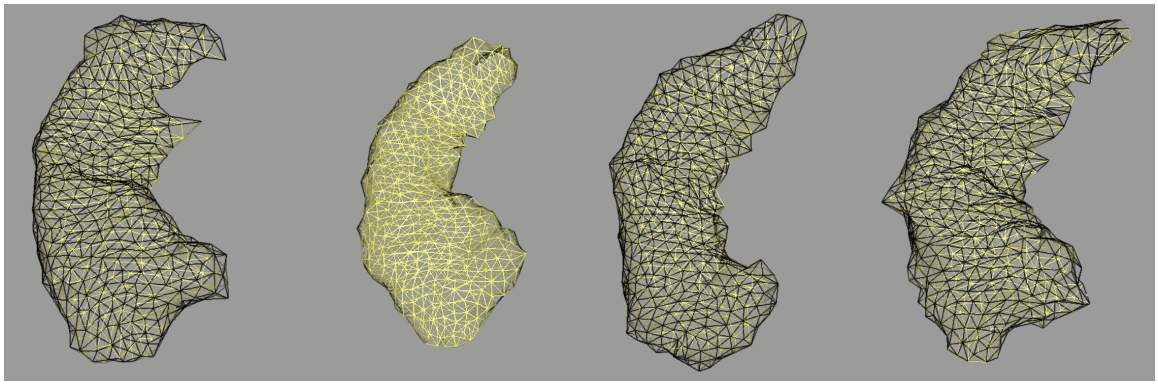
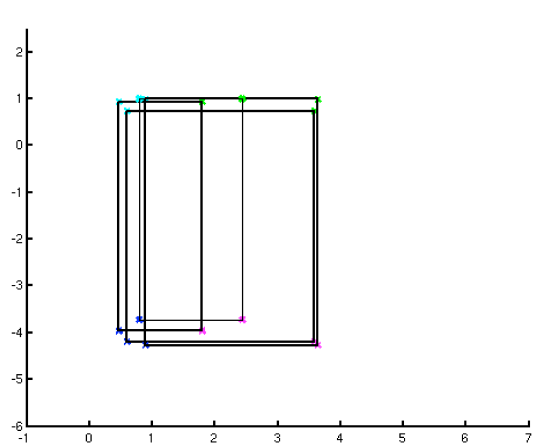


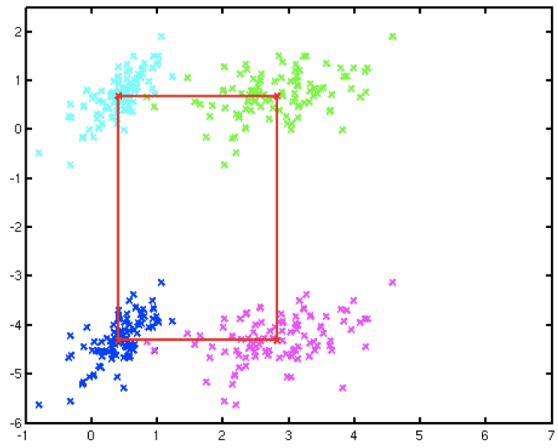
Figure 1.27: Surface representations of the left hippocampus for four subjects from our training set. Each surface has an equal number of vertices, each of which correspond across subjects.

along the eigenvectors and restrict how far the vertices move in that direction, we only allow plausible shapes.

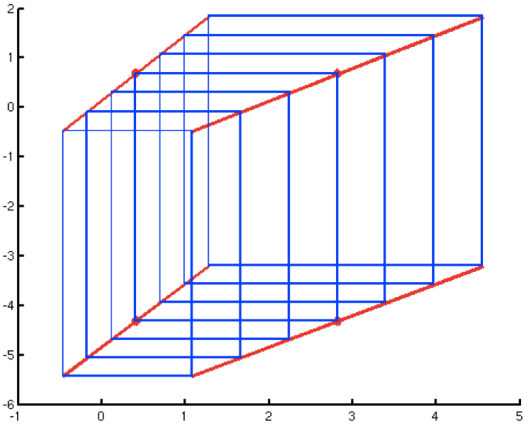
To illustrate the concept we generate a synthetic example of a hundred rectangles that were randomly generated for the purpose of training an ASM. The random rectangles were each parameterized using four vertices, corresponding to the four corners of the rectangle. Furthermore, the rectangles were generated such that vertices correspond across the samples, for example the first vertex is the bottom-left corner for each rectangle. The random sample set and mean rectangle is depicted in figure 1.28(b); the first four random rectangles are shown in figure 1.28(a). Figures 1.28(c) and 1.28(d) respectively show the shape variation modelled by the first two eigenvectors, the series of rectangles (in blue) is generated by displacing the vertices along the eigenvectors within ± 3 standard deviations. It is clear from the figure, that for this example the ASM may only synthesize new rectangles. The restriction to only rectangles is a reflection of the fact that all the training samples were rectangles, whereas in general, four points would not be constrained to form rectangles.



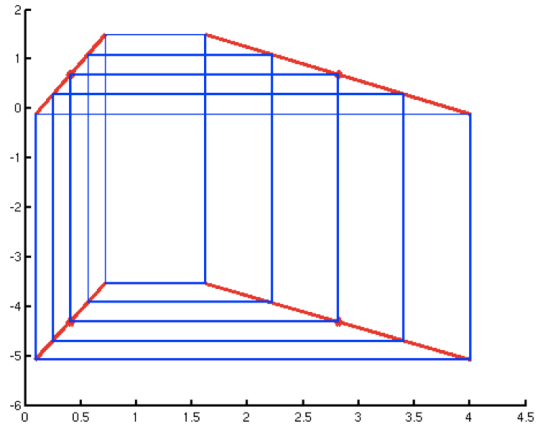
(a) First four training samples (rectangles)



(b) Training Set with Mean Shape



(c) First mode of variation.

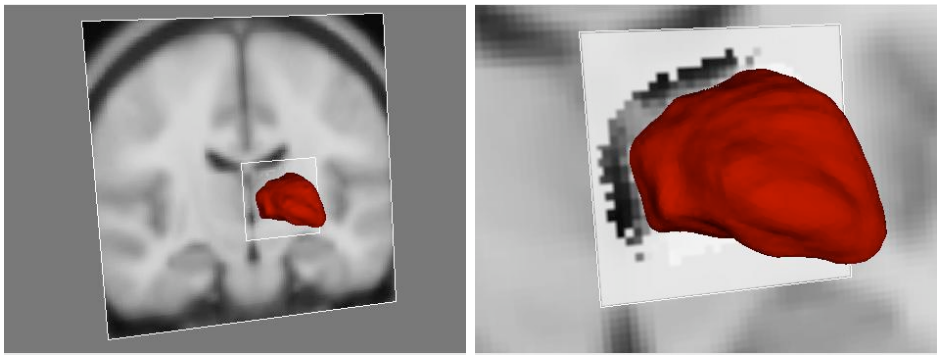


(d) Second mode of variation.

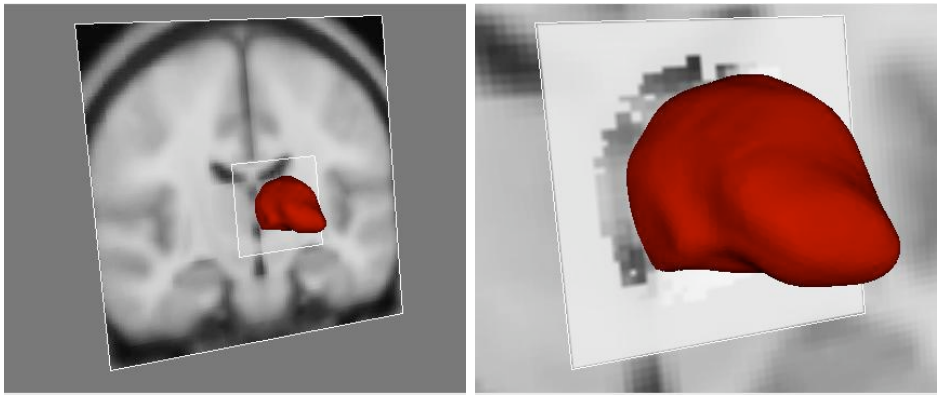
Figure 1.28: a) Four randomly generated rectangle. b) The corner vertices of a randomly generated training set of 100 rectangles with the mean rectangle in red. c) New shape instances (blue rectangles) generated by displacing the vertices ± 3 standard deviations along the first eigenvector(mode of variation). The first eigenvector is depicted in red. d) New shape instances (blue rectangles) generated by displacing the vertices ± 3 standard deviations along the second mode of variation. The second eigenvector is depicted in red.

To model the relationship between shape and intensity in addition to shape, the AAM was developed as an extension to the ASM [12]. The model for shape remains the same as that described for the ASM, however intensity is also modelled by its mean and eigenvectors. The AAM estimates the relationship between the shape and intensity models by investigating how intensity changes with variation in shape within the training data. Our method for modelling the relationship between shape and intensity differs from that of the AAM; the details for both will be given in chapter 3. To illustrate the value in learning combined shape and intensity information we will draw on our shape-and-appearance model for the left thalamus (discussed in chapter 3). Figure 1.29 depicts the shape variation of the left thalamus along the first eigenvector. In the right column the intensity band around the thalamus is the expected intensity for the given shape instance. The darkening of the band as the thalamus descends reflects the correlation between expanding ventricles (increased CSF) and the descending of the thalamus.

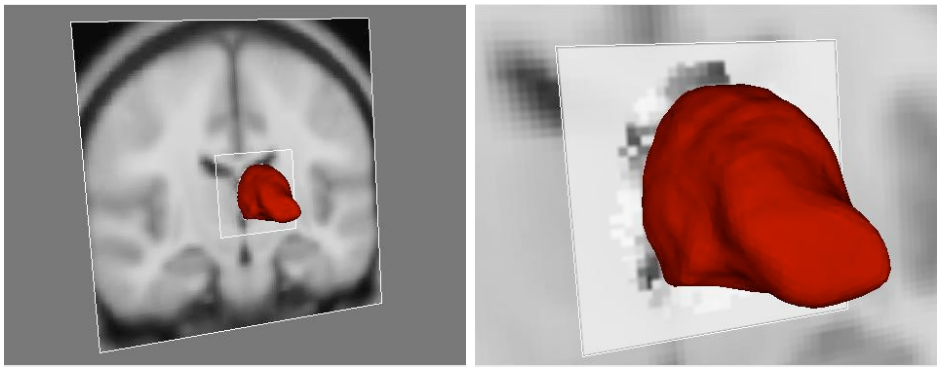
The following chapter will deal with the generation of the surface parameterizations that are necessary to construct our models (as shown in figure 1.27). Chapter 3 will then discuss the model for shape and appearance (as shown in figure 1.29) as well as their fitting to new images. Chapter 4 will apply the models to two clinical datasets and will investigate the use of classical statistics and discriminant analysis to explore shape differences. Chapter 5 will end with conclusions and a discussion of future directions of research.



(a) -3σ



(b) mean



(c) $+3\sigma$

Figure 1.29: First mode of variation for the left thalamus. The first column shows the thalamus surface overlaid on the MNI152 template (an average template created by registering and averaging 152 MR images of the brain). The second column is a zoomed-in view, with the conditional mean overlaid in the square patch. The enlarging dark band of intensities at the thalamus border represent the enlarging ventricle that correlates with the translation and shape change seen in the thalamus.

Chapter 2

Training Data, Pre-Processing and Surface Parameterization

2.1 Introduction

Our main objective is to develop an automated segmentation technique for subcortical structures. To achieve this goal we construct models of shape and appearance (intensity) for each structure. We use manually labelled T_1-weighted MRI images to train our models of shape and intensity for a particular structure. The point distribution model (PDM) is the central idea underlying the statistical shape model that is proposed in the following chapter. The PDM represents shape as a distribution of vertices on the surface of a structure and is trained from a set of surfaces with vertices corresponding across subjects. Therefore prior to training our model (chapter 3) we require surface representations for each structure in the manually labelled image such

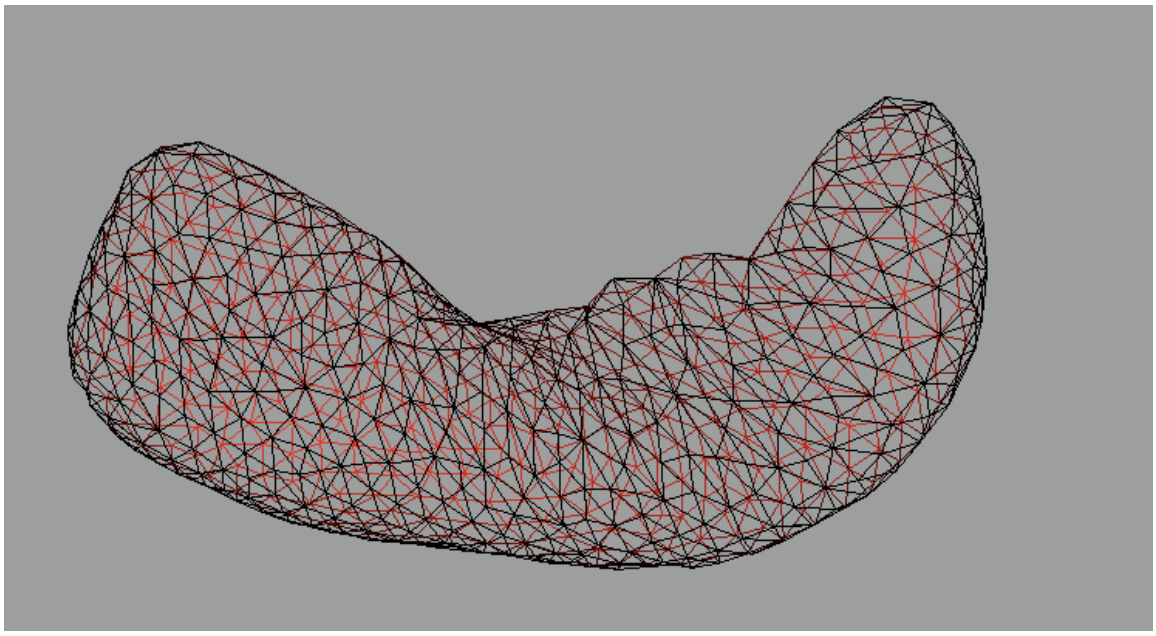


Figure 2.1: Mean surface for PDM of the left hippocampus.

that vertices correspond across the subjects. The main focus of this chapter will be obtaining these surface parameterizations of the manually labelled volumetric data. Figure 2.1 shows the mean of the PDM constructed for the left hippocampus.

In addition to shape, we will extract intensity information to include in our model of shape and appearance. The intensities are sampled along the surface normals from the T_1-weighted image that underlies the manual labels. Figure 2.2 is an illustrative example of sampling along the surface normals. Ultimately the intensities will be used to fit the joint shape and appearance model to new images; since absolute intensity scale and contrast may vary in T_1-weighted images we normalize the intensity samples. The normalization procedure is discussed in section 2.5 of this chapter.

Keeping in mind the goal of fitting the shape and appearance models to new T_1-weighted images, we must define a reference image space in which to construct the

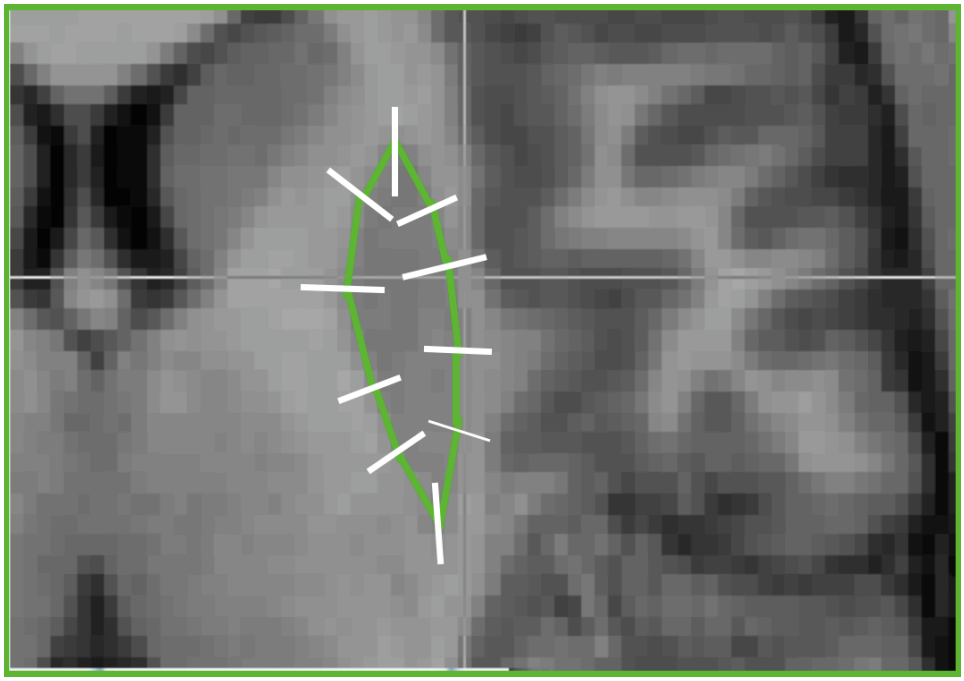


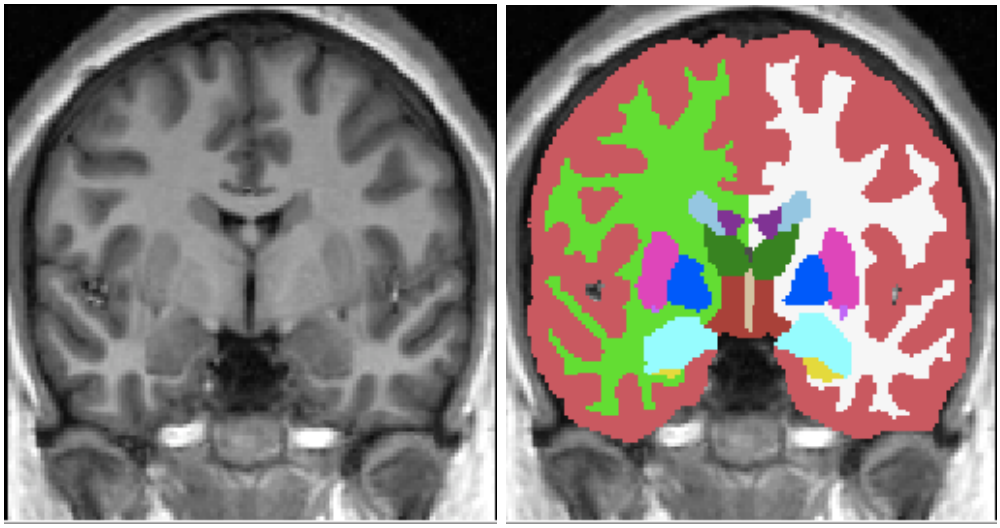
Figure 2.2: A 2D illustration of the intensity profiles for the left putamen. The profiles (white lines) are centred at each vertex and aligned along the local surface normal.

shape models. In addition to a reference space we must define the normalization procedure used to align the intensity image to the reference space. Provided that the models are constructed in the reference image space, a new image may be registered to the shape model by registering the new image to that space. In order for the variation being modelled to be representative of the variation after registering new images to the reference space, the normalization procedure ought to be common to both the training images and to the new images. Consequently, all the training images will undergo the same normalization procedure to the reference image space. The normalization procedure will be discussed in section 2.6 of this chapter.

We begin by discussing the PDM and point correspondence, followed by a brief review of methods for generating surface meshes from manually labelled images, and

then by a review of existing methods to surmount the correspondence problem of mesh vertices across subjects. Our method for surface parameterization which embeds the point correspondence criterion using deformable models, will then be presented in full detail. The methodology contained in this chapter was used to generate the surface parameterizations that are utilized in the next chapter to train our statistical models. Finally, we evaluate the accuracy of the method and discuss its advantages and pitfalls with regard to other methods.

In the interest of clarity, we will now draw a distinction between the various forms of training data throughout this document. The *training data* that we wish to parameterize in this chapter is a set of T_1-weighted MR images with corresponding manual labels (figure 2.3). We will thus refer to these as *training images*, which may subsequently be divided into *training label images* and *training intensity images*. In the context of training shape and appearance models (chapter 3), the training data are the vertex coordinates of the surface parameterizations and their corresponding intensity samples. We will refer to the surfaces derived from the *training label images* as *training surfaces* whereas the corresponding intensity samples will be referred to as the *training surface intensities*. When discussing discriminants in chapter 4, the training data refers to shape and size metrics derived from surfaces that were created by fitting the model to new data; training data is with reference to training the model of the discriminant boundary. We will delay further discussion of this training data until chapter 4.



(a) T_1 image (b) Manual Labels

Figure 2.3: A coronal slice from a single subject of the training images. a) MR T_1-weighted image. b) Manual segmentation overlaying the T_1-weighted image.

2.2 The Point Distribution Model and Point Correspondence

Our aim is to model shape such that it may serve as a prior probability model in an automated Bayesian segmentation/registration algorithm. We adopt a point distribution model (PDM) for shape as proposed in Cootes et al. [11]. The PDM models the distribution of a set of object landmarks (points) that correspond to object features; we use a multivariate Gaussian model. The PDM is trained from a set of examples containing a set of ordered landmarks that correspond across examples. In our application the landmarks are vertices on the surface of the anatomical structures of interest such that the PDM models the spatial variation of the vertices (including inter-vertex covariance). Vertex correspondence across the training surfaces is required when using such a model.

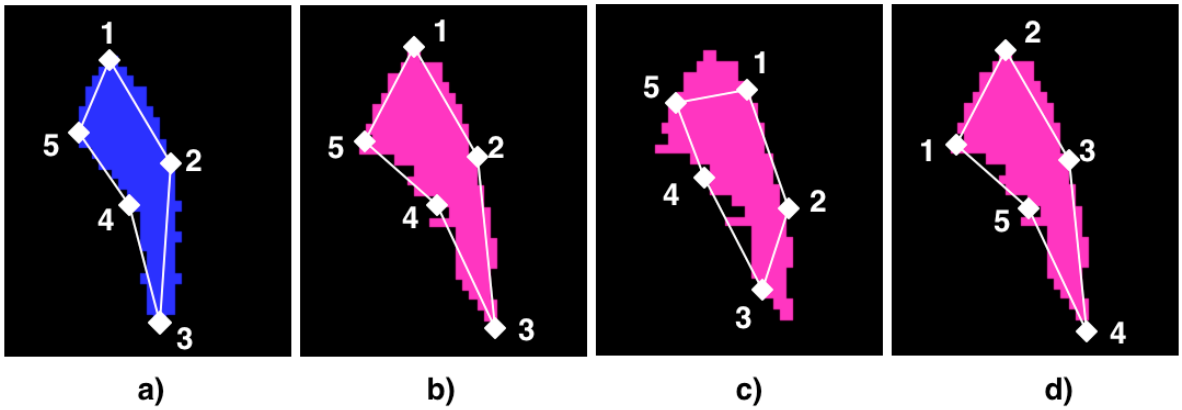


Figure 2.4: 2D illustration of the point correspondence problem for the left putamen - see text for details.

To construct a PDM of subcortical structures that is derived from the *training label images*, we first require a set of mesh parameterizations of the manual labels. In addition, the surfaces must achieve vertex correspondence across subjects. Furthermore, the correspondence criterion restricts the training surfaces for each structure to having an equal number of vertices in each subject. In three-dimensional images, manual assignment of a small number of landmarks typically requires a highly trained operator and may be very difficult and time consuming. The dense point correspondence we require would be nearly impossible to accomplish manually (we typically use in the order of 600 to 700 vertices for a given structure). Consequentially, automated methods of establishing point correspondence is an active field of research as this is both important for PDM-based segmentation (e.g. ASM and AAM) as well as for statistical shape analysis. Strictly speaking, we use mathematical landmarks since they are based on geometrical properties [20] (e.g. curvature) as opposed to anatomical landmarks which are based on some anatomical feature such as the meeting point of two structures; this is typical for automated point correspondence methods. Some of the mathematical landmarks used may correspond with anatomical ones.

To illustrate the problem, an example of a 2D transverse slice of the left putamen is given in figure 2.4. Approximately anatomically equivalent slices of the left putamen from the manual labels of two different subjects are depicted in blue and pink. Five landmarks have been manually placed along the boundary at common points based on curvature. The blue putamen (figure 2.4a) serves as our reference "correct parameterization". Figure 2.4b shows a correct parameterization for the second subject where the landmarks both correspond to the same anatomical location and the same numerical labelling.

Figure 2.4c and 2.4d depict two distinct cases where there is not correspondence. In figure 2.4c the ordering of labels is correct, however the landmarks do not correspond to equivalent anatomical/geometrical points of the left putamen. As shown by the white contour connecting the landmarks, the landmarks do not provide a good representation of the shape. The high surface point-density that we use in practice helps protect ourselves from shape mis-representation since the distance to interpolate between vertices is reduced. It does however still impact the modelling of shape variation since the PDM will view shifts along the boundary as movement of that anatomical/geometrical landmark. In figure 2.4d the landmarks correspond to the proper geometrical points, however the labels do not correspond. As with the previous case this becomes very problematic when attempting to learn variation based on a particular landmark. When training the statistical shape model, the change of label will be interpreted as a displacement of the landmark from the original location to the mis-labelled landmark and thus would not be a reflection of the true shape variation. The amount of influence the landmark mis-alignment has on the PDM is related to the surface point density. The higher the point density, the less influence

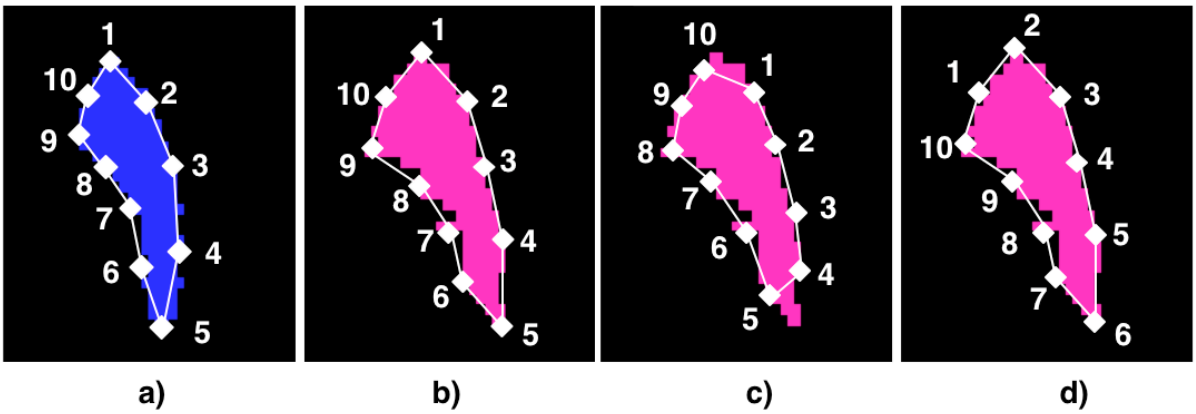


Figure 2.5: 2D illustration of the point correspondence problem for the left putamen with higher point-density - see text for details.

local shifts in the labels have on the overall model. To demonstrate the interaction between point density and the effect of label shifts on the PDM we draw attention to figure 2.5. In figure 2.5 the same left putamen as in figure 2.4 is shown, except with double the point density. As with figure 2.4d the labels in figure 2.5d are rotated one location clock-wise; it is visually evident that the degree of label mismatch (distance from actual "true" landmark) is reduced for the higher density mesh. This of course assumes that in both cases the mis-labellings are local. Furthermore 2.5c shows a shift from the appropriate geometrical point (as with figure 2.4c). Comparing figure 2.5c and 2.4b it is evident that the local shifts from the appropriate geometrical landmark does not affect the shape representation as much when using a higher point density. These properties are important to consider when deciding on an appropriate point density for the surface models.

2.3 Mesh Parameterization

This section gives a brief overview of marching cubes and the deformable model. Marching cubes is an algorithm specifically designed to generate surfaces from 3D volumes. Deformable models will be considered in the context of generating surfaces from *training label images*. They are however traditionally used for segmentation and will be revisited when reviewing segmentation methods in chapter 3.

2.3.1 Marching Cubes

Widely used in the field of computer graphics, machine vision and medical image processing/visualization, marching cubes is an algorithm for the tessellation of volumetric data [42]. Marching cubes associates a distinct triangular primitive to each binary voxel configuration within a cubic neighborhood. Each cubic neighbourhood associated with a boundary is assigned a triangular primitive (three vertices), so that by "marching" through the cubic neighbourhoods, three vertices are assigned to each neighbourhood. Vertices in common between primitives are fused into one, thus giving the mesh its connectivity. Vertex assignment is based on a local region of a binary image and is not dependent on global shape. The number of vertices generated by the marching cubes algorithm will vary depending on the shape and size of the structure. Despite being an efficient and accurate means to parameterize a structure within any given single image, marching cubes provides neither point correspondence, nor does is guarantee equal numbers of vertices across the training surfaces.

Determination of vertex correspondence from a series of meshes generated using

Subject	Volume (mm^3)	Number of Vertices (1mm Isotropic Resolution)	Number of Vertices (3mm Isotropic Resolution)
1	3348.0	2900	448
2	7482.0	6316	710
3	3405.32	3148	434

Table 2.1: Volume and the number of vertices produced for the left lateral ventricle by applying marching cubes to three subjects from the training data.

marching cubes is not a trivial task. To illustrate the variability in vertex topology, the meshes output from applying marching cubes on the left lateral ventricles for three subjects is depicted in figure 2.6. The subjects were chosen such that they represent a reasonable variation (cross-subjects) in ventricular size and shape. The first and second rows of figure 2.6 show the surfaces produced from the same image at 1mm and 3mm isotropic resolution respectively; each column represents a single subject.

As is visible in figure 2.6, for a given image size and resolution, meshes have similar vertex densities and thus the number of vertices is closely correlated with the surface area of the given structure. Table 2.1 shows that the number of vertices is related to gross changes in the volume of the structure (difference between subject 1 and 2, as well as between 2 and 3). In actuality the number of vertices is directly related to surface area (which in turn is related to volume), this is the reason why an increase in volume between subjects one and three results in a decrease in the number of vertices (the surface area increased despite a decrease in volume). The impact of volume on the number of vertices is stronger at higher resolutions.

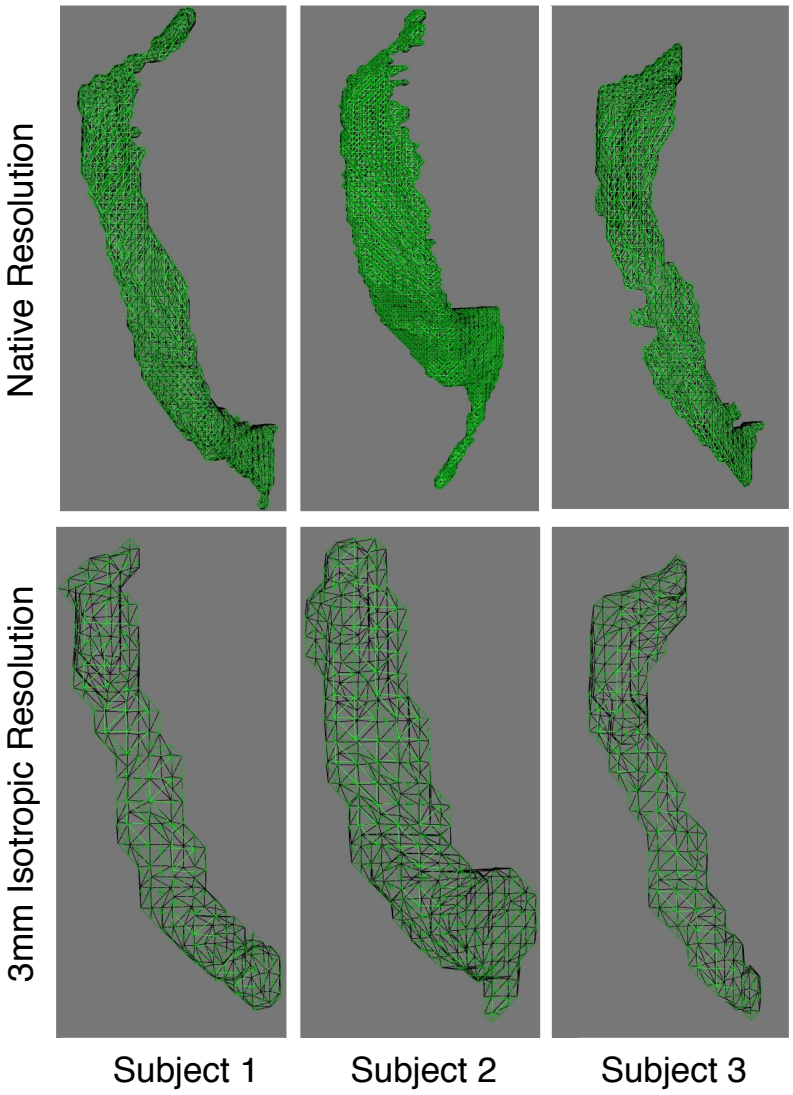


Figure 2.6: Marching cubes applied to the left lateral ventricles for native resolution and down-sampled to 3mm isotropic resolution.

2.3.2 Deformable Models

Deformable models provide an alternative means for obtaining a mesh parameterization from a training label image. We will discuss deformable models in some detail since we use them to parameterize the data whilst retaining correspondence (section 2.7). The advantages of using shape models (such as the ASM) over deformable models will also be highlighted in the following chapter.

Deformable models allow for the iterative displacement of vertices towards a boundary. The vertices are driven by forces acting on each vertex. The surface forces are categorized into internal and external forces. The vertices are displaced as a function of the net force applied to the vertex. The net force $\boldsymbol{f}_{net}$ is expressed as

$$\boldsymbol{f}_{net} = \boldsymbol{f}_{int} + \boldsymbol{f}_{ext}, \tag{2.1}$$

where $\boldsymbol{f}_{int}$ is the internal force, $\boldsymbol{f}_{ext}$ is the external force; each force vector at each vertex is a three-dimensional vector (using a Cartesian basis). The external force, $\boldsymbol{f}_{ext}$, couples the surface to the image and is derived from image metrics such as intensity gradients (or manual label values). The internal force $\boldsymbol{f}_{int}$ is derived from shape metrics (e.g. curvature) and helps to constrain the mesh during deformation. $\boldsymbol{f}_{int}$ and $\boldsymbol{f}_{ext}$ may themselves be comprised of multiple forces as given by

$$\boldsymbol{f}_{int} = \sum_{i=0}^{N} \alpha_i \boldsymbol{f}_{int,i}, \tag{2.2a}$$

$$\boldsymbol{f}_{ext} = \sum_{i=0}^{N} \alpha_i \boldsymbol{f}_{ext,i}, \tag{2.2b}$$

where α_i is the weighting of the i^{th} force component $\boldsymbol{f}_{int,i}$ and $\boldsymbol{f}_{ext,i}$ respectively. The mathematical details of the specific forces used in our application are discussed in section 2.7.2.

The weighting parameters are usually empirically estimated where the optimal value may vary depending on the application and/or image quality. The vertex displacement is then either calculated as being proportional to the net force applied or alternatively a physics model may be used. The physics model attributes a force, acceleration, velocity and displacement to each vertex. In practice the physics model may add stability to the optimization.

The stopping criterion for the deformable model is typically a minimal-displacement criterion such that if the maximum vertex displacement is below a given threshold, the algorithm is deemed to have reached convergence. Alternatively, one may select a fixed number of iterations after which the algorithm stops. Both methods rely on an empirically-chosen threshold, however the latter does not consider the state of the deformation process.

Generally, vertices may be deleted and added from deformable models. The deletion and addition of vertices is done in order to cope with large changes in size and curvature. It is straightforward to keep the number of vertices constant by forbidding the addition and deletion of vertices. This does, however, impact on the stability and flexibility of the deformable model, typically making it more difficult to achieve the stability for the desired amount of flexibility. By forbidding the insertion and deletion of vertices, and provided there is no mesh self-intersection then we are guaranteed to preserve topology (though not necessarily correspondence).

Another aspect of deformable models that needs mentioning is the potential problem of mesh self-intersection. This occurs when part of the mesh crosses another part of the same mesh, thereby turning a portion of the deformable model "inside-out". In the inverted areas of the mesh, forces designed to constrain curvature during the deformation (typically acting inwards) now face the opposite direction, thereby forcing the vertices outwards causing those areas to blow up. The instability due to mesh-intersection is detrimental to the deformation process and violates our shape assumptions, and thus must be carefully controlled for.

When using deformable models to parameterize a training label image, the labels serve as the underlying truth that can be used to drive the vertices to the label surface. In order to use deformable models an initial mesh is required to serve as the starting point for the deformation. Two simple methods of initialization would be: 1) initialization with a sphere (which may be easily created) and 2) initialization using marching cubes. The particulars of the deformable models used, including the choice of forces and mesh initialization will be discussed in section 2.7.

2.3.3 Methods for Establishing Point Correspondence

Methods for establishing correspondence vary in nature. Correspondence may be established based on geometric features, non-linear warp fields or via direct optimization using an objective function.

Wang et al. [60] propose the use of geodesics and local geometry to establish correspondence between surface representations of the cortex. The algorithm begins

with segmented images that may have either been derived from manual or automated methods. The surfaces are then extracted from the segmented images using the marching cubes algorithm. A small set of key landmarks in an atlas space must then be identified and triangulated. The landmarks are divided into four categories that correspond to the anatomical location of the landmarks: 1) inter-hemispheric fissure, 2) creases at the brain stem or cerebellum, 3) sulcal points and 4) gyral points. The manual placement of the key atlas landmarks is time consuming, however it need only be performed once for the atlas. In order to determine correspondence between the key landmarks on the atlas and an individual subject, the subject is aligned to the atlas using translation, rotation and scaling. After alignment an objective function based on local geometry is used to establish the initial correspondence. The objective function is defined as the product of a Euclidean distance function, surface normal matching function and a curvature function. Using the original marching cubes triangulation, the shortest geodesic path between each pair of corresponding points is calculated using a small variation on the extended Fast Marching Method algorithm [37]. The midpoint along the path is then selected as a new landmark such that the mid-points are assumed to correspond across subjects. The process of mid-point selection is re-iterated to produce finer resolution meshes. This iterative mid-point selection process relies purely on geometry, however it is unclear how well shortest paths will correspond for changes in that geometry (i.e. different shaped brains); the mid-point correspondence relies on correspondence between shortest paths.

Kelemen et al. [36] propose a correspondence method based on surface landmarks for the construction of 3D ASMs for neuroanatomical structures. Correspondence is established by mapping the surface to a sphere. Spherical harmonic descriptors

are used to represent the surfaces such that the coefficients of each degree measure the contribution of the spatial frequency. The first degree coefficients represent an ellipsoid and are used to align the subjects to a standard position. The alignment is performed by rotating the parameter space such that the north pole of the long axes of the ellipsoids align. This method provides a fully automated means to achieve correspondence, however the first order ellipsoid may not always be sufficient for establishing correspondence given local shape differences.

Fleute et al. [26] aim to use a statistical shape model of the femur to constrain a deformable model for the segmentation. Ten femurs were digitized with approximately 1500 points whilst the eleventh subject was reserved to serve as a template with approximately 6000 points. The high density template is required for the determination of correspondence with this method. The femurs were registered to the template using a hierarchical free-form deformation with octree splines. For each point in the template mesh an iterative search is used to find a corresponding point. A corresponding point is defined as a point for which the distance is below a given tolerance. Each point on the template mesh is assigned a point for each subject, the points are not contained on the original mesh but are interpolated using the octree spline. This method is closely related to the method proposed by Frangi et al. [28] and to that proposed in this chapter. All three methods establish correspondence based on structural deformation; our method differs from the others in that we use a deformable model rather than elastic registration. In this paper the deformation between surfaces as well as point correspondence needs to be determined; the regularization on the deformation leads to inaccuracies in the correspondence for large shape differences. Our deformable model method embeds the correspondence prior to defor-

mation and relies on the volumetric label image rather than a surface representation to determine the deformation.

Frangi et al. [28] propose the use of a free-form non-linear volumetric registration using a multi-resolutional B-spline warp field to automatically construct statistical shape models. The transformation method comprises of a global affine transformation to remove pose and scale in addition to a free-form deformation field. To accommodate manual labels the *labelconsistency* and κ*statistic* metrics are introduced to drive the deformation. The *labelconsistency* metric is the joint probability of the label from each image. The κ*statistic* measures the agreement between two labels whilst correcting for chance occurrences. Similar to this method we use a free-form deformation, however it is surface based. Our preference is based on the idea that to create a statistical model for surface deformation, one should use a surface-based parameterization method that mimics the same process that we are attempting to model.

Brett et al. [8] propose a polyhedral-based method for pair-wise correspondence that they later extended [9] for the purpose of automated 3D PDM construction. Correspondence is established from sparse polyhedral representations of the high resolution mesh using a global Euclidean distance metric. The sparse surface representation is obtained using a triangular decimation of the original high resolution mesh. Correspondence across a population is performed by matching the sparse polyhedra at different levels of detail. The matching process is represented as a tree structure with the mean shape at the top. Each parent node is the mean of the two sparse polyhedra that descend from it. At the bottom of the tree are the original meshes with each vertex corresponding to a vertex on the average mesh at the top of the tree. The

mean shape defines the surface topology for the shape model; it is decimated until the desired number of landmarks is achieved. Because of the decimation and sparse representations of the surfaces, the process may be less accurate for sharp edges and thin structures.

Kotcheff et al. [38] propose an alternative approach to determining correspondence for the automated construction of PDMs. An objective function is used to optimize for model compactness and specificity. Model compactness can be measured by the number of eigenvectors that can explain a given amount of variance. A model that requires fewer eigenvectors to explain the equivalent amount of shape variance is more compact. Model specificity is a measure of the model's ability to synthesize shapes similar to those in the training data. The objective function used to capture compactness and specificity is the determinant of the covariance matrix. Shape pose and parametrization are optimized with a genetic algorithm using the objective function as a criterion. For each new parameterization and correspondence generated in the search, the covariance matrix is estimated and its determinant evaluated. The determinant of the covariance matrix is equal to the product of the eigenvalues of the covariance matrix. By selecting the parameterizations/correspondence that minimizes its determinant, the model selected balances the minimization of total variance and compactness. The idea is that a shape model built from training surfaces that lack correspondence are less compact and contain more variance than for shape models constructed from training surfaces with correspondence. The advantage of this approach is that it provides an objective function for determining an optimal model, however, it is not clear that an optimal model as defined here necessarily corresponds to an optimal model in terms of segmentation performance.

Davies et al. [18] also treat the correspondence problem as a direct optimization problem where Minimum Description Length (MDL) is used as the objective function. MDL is based on the idea that learning may be viewed as finding regularities in the data and that they in turn may be measured by the data's compressibility. MDL is rooted in information theory and aims to serve as an objective function that attempts to balance the trade-off between goodness-of-fit and complexity. MDL is often used to achieve model selection; it casts the problem as finding the most efficient statistical model for transmitting the information. The MDL criteria can be broken down into the code length and description length. In this case the code length is a function of the dimensionality of the shape vectors and the number of training samples, these quantities are constant for a given training set and thus need not be optimized. Therefore only the description length remains to be optimized. The description length is a function of sample size and model variance such that the present terms are similar to those for the minimum determinant of covariance matrix criteria. The optimization methods provide a mathematically sound framework and have been shown to produce reasonable 3D statistical shape models [17]. Despite the fact that it provides an optimization criterion for a linear eigenvector model, as with the determinant of the covariance matrix criterion, it is not clear that the optimal parameterizations in terms of segmentation performance correspond to the most compact models, particularly when both shape and appearance are being modelled.

Rather than focussing on using compactness or specificity of the resultant model, we centred our focus on developing a method that does not consider the resultant model but rather aims to produce an appropriate deformation model between subjects such that it retains physical correspondence. We propose a method for automated model

construction using deformable models such that point correspondence is implicit to the deformation process. We use three-dimensional deformable models to parameterize a structure's surfaces from a volumetric binary image. This is accomplished by iteratively displacing the vertices of an initial mesh until they lie on the desired boundary. The vertex displacements are determined via external and internal forces that are based on image information and vertex geometry. By using this method we hope to produce a representative model of the physical shape deformation that exists between subjects despite a possibly less compact model. The model was designed and optimized based on accuracy of the fit, and visual inspection of the deformation process and the resultant model.

2.4 Training Images

This section discusses the image data available for training our shape and appearance models. Prior to training our shape models, we need to generate mesh parameterizations of the *training label images*. The training images used in this work consist of 317 manually-labelled T_1-weighted magnetic resonance images of the brain. The 317 datasets are made up of six distinct groups of data. The sample population spans both normal and pathological brains (including cases of schizophrenia and Alzheimer's disease). The size, age group, and resolution for each group is given in table 2.2. The T_1-weighted image and manual labels of a single subject from the training images are depicted in figure 2.3.

All the *training images* were linearly registered to the MNI152[1]standard space using

Group	Size	Age Range	Resolution (mm)	Patient Group
1	37	aprox. 16 to 72	1.0 x 1.5 x 1.0	NC & SZ
2	42	Adults	1.0 x 1.0 x 1.0	NC & AD
3	17	65 to 83	0.9375 x 1.5 x 0.9375	NC & AD
4	87	23 to 66	0.9375 x 3.0 x 0.9375	NC & SA
5	14	9 to 11	1.0 x 1.0 x 1.0	NC & PC
6	120	4.2 to 16.9	0.9375 x 1.5 x 0.9375	NC & ADHD & SZ

Table 2.2: Groups within the training data and their respective size, age group and resolutions. NC indicates normal controls, SZ indicates schizophrenia, AD indicates Alzheimer's disease, ADHD indicates attention deficit disorder and PC indicates prenatal cocaine exposure. For the first group we do not have demographics for the entire dataset.

the procedure described in section 2.6. We are modelling 17 structures: brainstem, the left/right amygdala, caudate nucleus, hippocampus, lateral ventricles, nucleus accumbens, putamen, pallidum and thalamus. We model appearance from normalized intensities that are sampled from the original *training intensity image* along the surface normal at each vertex of the training mesh (both inside and outside the surface); 13 samples per vertex at an interval of 0.5 mm was used. An illustrative example of profile sampling is depicted in figure 2.2, where we show a 2D contour with the profiles drawn for each vertex; the intensity samples are taken along these profiles. The sampling interval was chosen empirically such that it is half the resolution of the reference space. The sampling extent was chosen empirically based on the observed performance of the shape and appearance model.

[1]This is the standard template created by the MNI from affine-aligning 152 subjects before averaging.

2.5 Intensity Normalization

In this section we describe the intensity normalization performed prior to constructing our shape and appearance model. The same intensity normalization will be performed when dealing with a new image. The global scale of intensity values in MRI is variable and arbitrary and thus should be normalized prior to modelling. Furthermore, bias field inhomogeneities may cause slow drifts in intensities across the image that are not tied to the underlying anatomy.

In our work, two stages of intensity normalization are performed prior to constructing the intensity model. The first is a global normalization that scales and shifts the intensities such that the robust minimum and maximum of the image correspond to 0 and 255 respectively; the normalized intensity values are floating point values. The robust minimum and maximum correspond to the minimum and maximum intensity after discarding the intensities in the upper and lower two percentiles. They are used to obtain a more robust estimate of the minimum and maximum intensities of the head region found in the images, otherwise hypo/hyper-intensities due to imaging artifact could easily bias the estimation.

The second normalization is performed on a per-structure basis. Assuming that the distribution of intensities within a structure is of a consistent form, we normalize by subtracting the mode of the distribution. This normalization will need to be estimated at run time from the image region defined by the mesh, at which point there may be a mixture of tissues. Therefore it is important to choose a normalization metric that may be robustly estimated when fitting to a new image. In practice, when estimating the mode intensity, the mixture of tissue types is dominated by the tissue contained

within the structure of interest. When the contribution of the uninteresting tissues is small, it will mostly impact the tails of the distribution. In our experience the distribution of the tissue of interest is reasonably Gaussian. To robustly estimate the mode, an intensity histogram is constructed and the centre of the largest bin is selected; this avoids the confounds due to corrupted tails.

Unfortunately the assumptions mentioned above do not always hold. For example, the shape of the probability distribution function (pdf) within the lateral ventricles tends to vary with structural size/shape. This is primarily due to large regions that are affected by partial voluming and the inclusion of the choroid plexus. When fitting to a new image, the assumption that the pdf is dominated by the tissue of interest does not always hold for the ventricles, caudate, hippocampus, and amygdala. For example, at initialization when the true ventricles are small, the sampled intensities will contain significant amounts of surrounding white matter. The caudate, hippocampus and amygdala encounter problems when there is severe atrophy such as seen in Alzheimer's disease (AD) where the pdf estimate is confounded by the presence of cerebrospinal fluid (CSF). Figure 2.7 shows the outline of the initial mean hippocampal shape for a single subject with AD. It is clearly visible that there will be a significant contribution from CSF. To obtain a robust normalization estimate in these structures we normalize by the mode of a nearby stable structure, in our case the thalamus. The same normalization is used for the training as well as for fitting. We will compare results from the two cases in the next chapter.

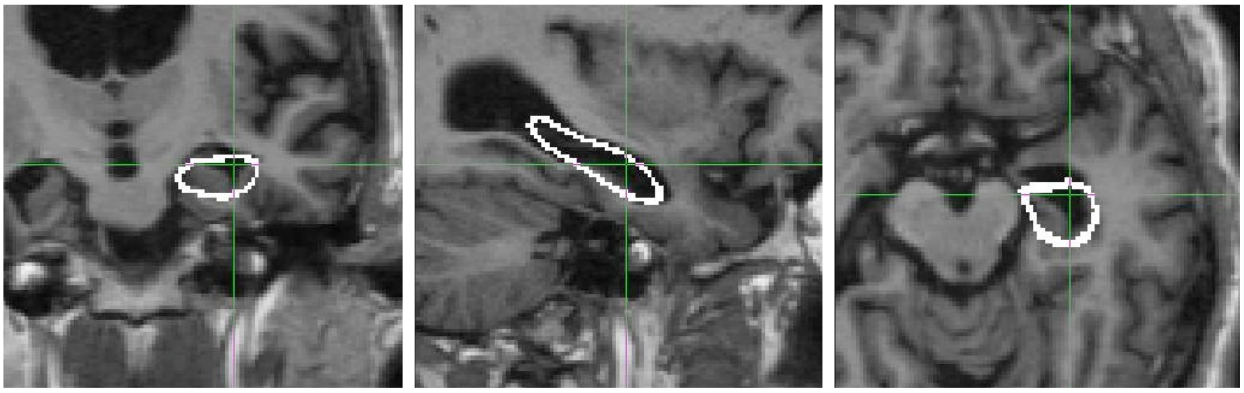


Figure 2.7: Initial hippocampus for a single subject from the training images with AD.

2.6 Linear Subcortical Registration

This section discusses the (initial) spatial normalization that is performed on both the training images and new images alike. The normalization procedure defines the common space in which the models are built and to which new images are aligned. Prior to modelling shape and intensity, a common space is chosen to which all data may be accurately, robustly and automatically aligned. The linear MNI152; template, re-sampled to 1mm isotropic resolution, was chosen as our reference space. The shape variance we are modelling by the shape models described in chapter 3 is in fact the residual shape variance that exists after normalization to this template. The model is therefore specific to the normalization procedure.

Prior to surface parameterization, the normalization process was applied to all the training images. In order to achieve robust subcortical alignment, normalization is performed using a two-stage linear registration where all linear registrations were performed using FMRIB's Linear Image Registration Tool (FLIRT) [33]. When seg-

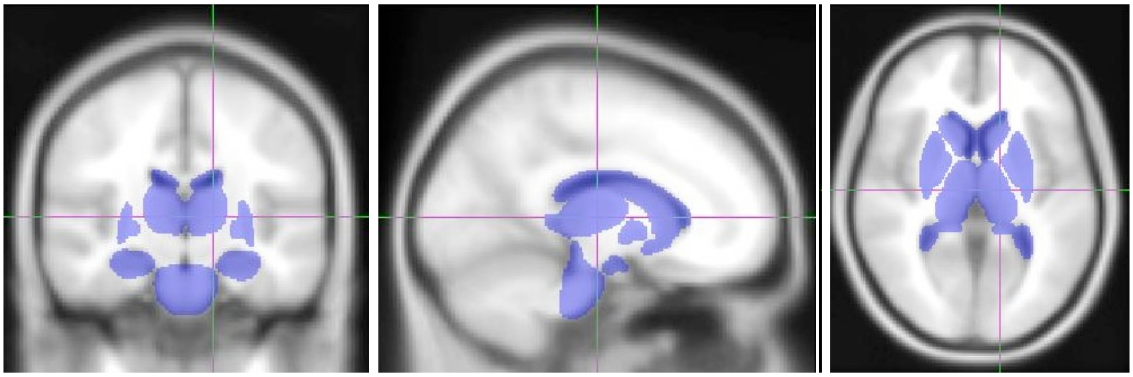


Figure 2.8: MNI152 template at 1mm isotropic resolution with the subcortical mask overlay (blue).

menting a new image, this image is also registered to the template, with the model then brought into the image's native space using the inverse transform. By transforming the model rather than the image we eliminate the need to interpolate the image.

The first stage in the alignment process is a standard whole-head affine registration to the MNI152 template. Using a subcortical mask defined in MNI space, an affine registration to the MNI152 template is then applied to the output image of the first stage. Figure 2.8 depicts a common slice from the MNI152 template and the subcortical mask used in the registration. The mask excludes any voxels outside the masked region when computing the similarity function (correlation ratio) within the registration optimization method. By doing so the registration algorithm will ignore any dissimilarities between the template and the aligned image in the region outside the mask. The subcortical mask was generated from the average surface shape of the 17 structures being modelled from a subset of the 317 training subjects (using only the standard affine registration).

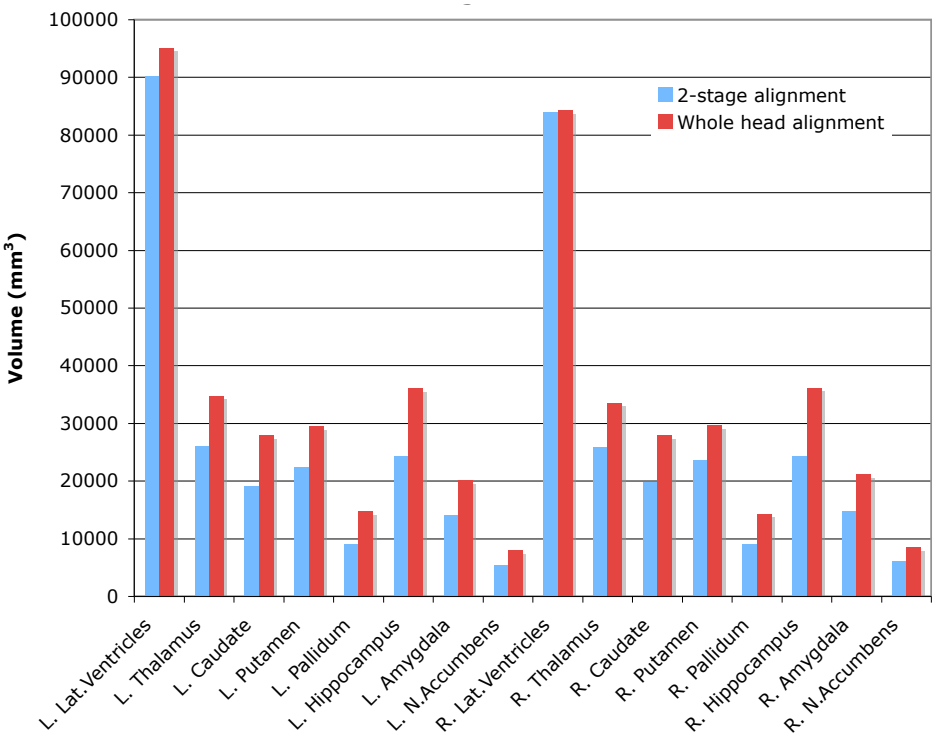
The aim of the subcortical registration is to reduce the residual shape variance that exists after linear registration. The maximum amount of residual shape variance is assumed to be proportional to the spatial extent of the probability map of the structure. The probability map is constructed by applying the linear transformations to the *training label images* and then evaluating the mean at each voxel. The expression for the probability map is given by,

$$P\left(x,y,x\right) = \frac{1}{N}\sum_{i=1}^{N} T_i\left(L_i\left(x,y,z\right)\right), \tag{2.3}$$

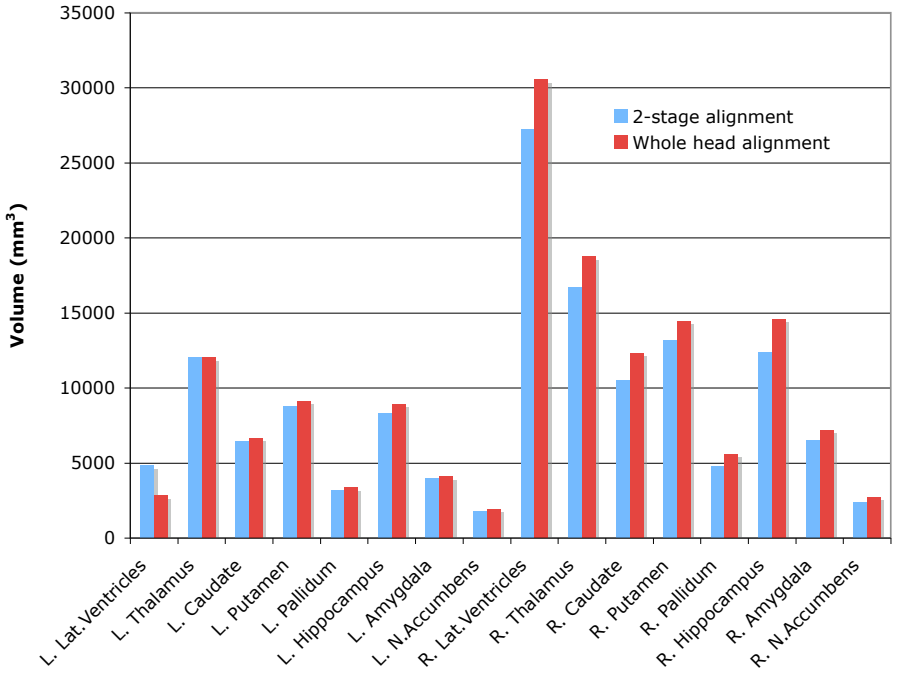
where $L_i(x,y,x)$ is manual label image (binary) for the i^{th} subject for a given structure, $T_i()$ is the linear transformation operator representing the transformation to standard space, N is the number of images (317 in our case). The label image is binarised prior to transformation and tri-linear interpolation is used when evaluating $T_i\left(L_i\left(x,y,z\right)\right)$.

We use the spatial extent of the probability map to compare the whole-head registration to the proposed two-stage subcortical alignment. FLIRT was used to register each of the training intensity training images to the MNI template using both the whole-head registration and the two-stage alignment. The spatial extent was measured by the number of voxels (volume) above zero-shape probability. To investigate whether the differences were simply due to outliers we also calculated the spatial extent for the probability of above 30%.

In figure 2.9, for each modelled structure, the volumetric extent of the probability maps that were generated using the two-stage alignment is contrasted to that generated using a standard whole-head registration. As expected the spatial extent, at a



(a) Spatial Extent ($p > 0$)



(b) Spatial Extent ($p > 0.3$)

Figure 2.9: Spatial extent of probabilistic atlas constructed using a standard whole-head affine registration and the two-stage subcortical alignment.

probability threshold of 0, is reduced for all structures when using the two-stage alignment. This held true for all structures except for the left lateral ventricles and left thalamus when excluding all voxels under a probability of 30%. At a 30% threshold the two-stage alignment showed greater spatial variability for the left lateral ventricles and negligible difference for left thalamus when compared against a standard whole-head registration. The general reduction of spatial variability (with the exception of the left lateral ventricles) after normalization serves as a strong motivation for using the proposed subcortical alignment method. By reducing the amount of variability we facilitate the mesh parameterization using deformable models as well as reduce the extent of our search space when fitting. Furthermore, the method was shown to be adequately robust since there were no observable failures across the 317 training images. The differences observed in spatial extent for the lateral ventricle is negligible compared to that of the other structures (and in the opposite direction for the left ventricle at a probability threshold of 0.3). This was expected since we are placing less emphasis on matching the white matter superior to the ventricles by using the subcortical mask. This is of particular importance for cases with large ventricles where in order to align the other subcortical structure (e.g. caudate) there will need to be disagreement between the superior borders of the ventricles.

2.7 Surface Parameterization with Embedded Correspondence Using 3D Deformable Models

This section describes the proposed parameterization method by which we form surface representations of the *training label images*. Deformable models are used to model the deformation of a structure between subjects, with constraints on the deformation process that preserve vertex correspondence. The surfaces generated via this method along with the corresponding intensities will be used to train our statistical model of shape and appearance that is described in the following chapter.

By considering a training label image as an image to be segmented, the application of a deformable model will provide a surface representation of the underlying structure. Because the underlying images are manual labels, the deformable models do not suffer from the many problems associated with imaging noise and artifacts. For the purpose of constructing a PDM, 3D deformable models must be carefully constrained to retain point correspondence. One constraint is that the number of vertices must remain constant across the training surfaces. Therefore the deformation process must be designed such that it is robust across large variations whilst disallowing vertex insertion/deletion. Smoothness and inter-vertex distance forces are typically included to provide robustness and prevent mesh self-intersection (see section 2.7.2 for details). Use of a smoothing force when fitting to a training label image may sacrifice the accuracy of the parameterization and is thus not appropriate. The effect of the smoothing force is especially noticeable with subcortical structures as they typically have areas of high curvature. An inter-vertex distance force acts tangential to the surface and induces within-surface motion; unwarranted within-surface motion will

violate our point correspondence criteria. To retain point correspondence the within-surface motion contained in the deformation process must be kept to a minimum.

To aid point correspondence we introduce a local image-based structural alignment that serves to initialize the deformable model with a first approximation to point correspondence. We then proceed to the deformation process where we introduce a novel deformation force that allows us to achieve accurate fits whilst preserving correspondence.

2.7.1 Structural Pre-Alignment

Prior to deformation, the label image of each structure for each subject is registered (separately to the other structures) to the binary image formed from the average mesh (we will discuss the selection of the average mesh later in this section). Figure 2.10 shows the before and after image of this affine registration. It is clear that for a single structure there remains affine differences after the global registration to standard space but prior to this local affine transformation. This difference largely disappears after the local affine transformation.

The reference binary image used to remove the local affine components is created by filling the average mesh. The linear registration (transformation $T(x, y, x)$) was performed using twelve degrees of freedom with a least-squares cost function. In practice, the inverse transformation $T^{-1}(x, y, z)$ is applied to the average mesh such that it is transformed into the native image space. The local affine alignment of structural features (using the transformation $T(x, y, x)$) is used to initialize the mesh

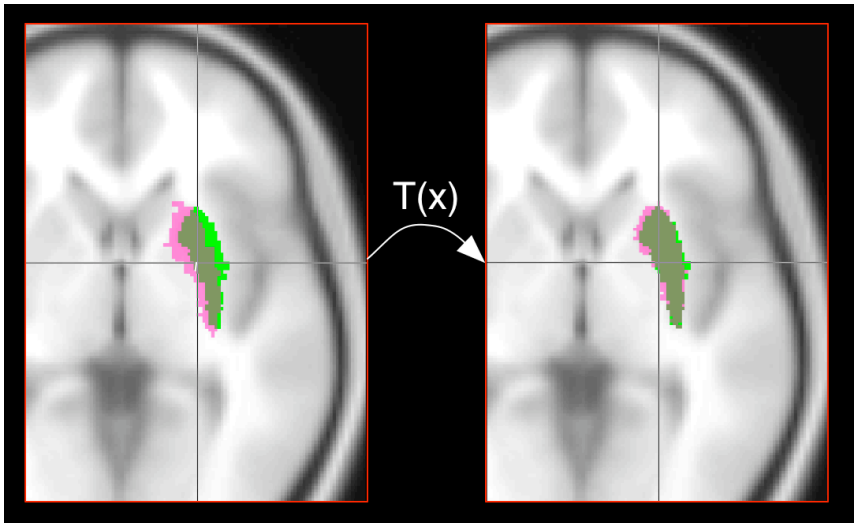


Figure 2.10: The left depicts the left putamen of a single subject overlaid on the "average left putamen", it has been aligned to the MNI152 template using the 2-stage linear registration described in section 2.6 (a global whole-head registration). The right side is the left putamen after an affine registration (local putamen-only registration) between the left putamen aligned to the MNI152 template (as in the left image) and the average left putamen. The local affine transformation, $T(x)$, is calculated using FLIRT on the label images with a least-squares cost function.

prior to the deformation process described in the following section. It serves to preserve point correspondence and reduce the vertex displacement required of the deformable model. Since the structures have a prior alignment (the global affine registration to standard space) and are sufficiently non-spherical, the transformation $T(x, y, x)$ proves to be robust.

2.7.2 The Deformation Process

The deformation process is an iterative update of vertex location. The vertices are displaced according to a weighted sum of forces that are applied to each vertex. The surface forces are categorized into internal and external forces. The external forces serve to displace the vertices towards the intensity boundary in the image whereas the internal forces serve to impose shape constraints.

The external image force acts along the inward/outward pointing vertex normal and is expressed as a step function based on the binary image value (equation 2.4). The force magnitude is greater in the interior than the exterior of the structure to prevent the mesh from clipping thin regions.

$$\boldsymbol{f}_{ext} = \begin{cases} +50\frac{\boldsymbol{s}_n}{|\boldsymbol{s}_n|} & \text{if } label = 1 \ , \\ -5\frac{\boldsymbol{s}_n}{|\boldsymbol{s_n}|} & \text{if } label = 0, \end{cases} \tag{2.4}$$

where $\boldsymbol{s}_n$ is the local surface normal (see figure 2.11) and $|...|$ indicates the L_2 norm. Three internal forces are considered in our deformation model. The first two internal forces (identical to those used in BET [56]) are: 1) A smoothing force, $\boldsymbol{f}_n$, that

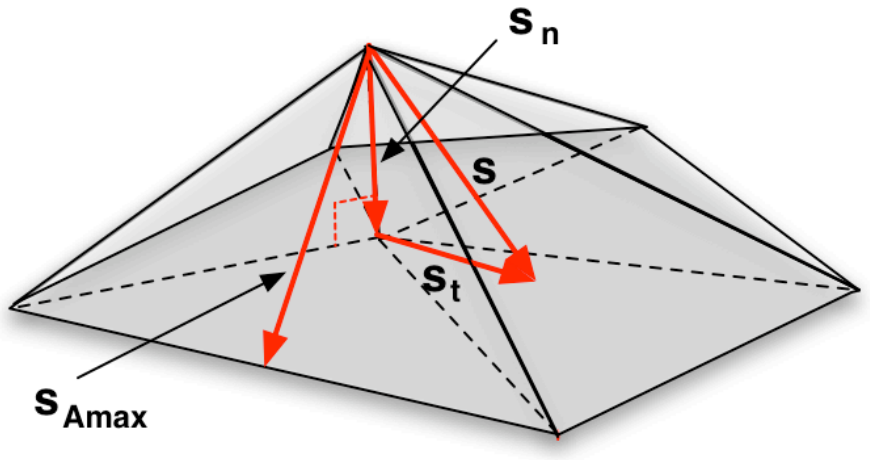


Figure 2.11: Forces acting on vertices of mesh during deformation process.

is proportional to the local curvature and acts inwards along the surface normal (equation 2.5a). 2) An inter-vertex distance regularization force, $\boldsymbol{f}_t$, that favours even point density across the surface (equation 2.5b). We propose a third internal force, $\boldsymbol{f}_{Amax}$, that favours equal area triangular patches. The area-based force acts in the direction that bisects the largest adjacent triangle and is proportional to the area of that triangle (equation 2.5c). In the case where there is more than one triangle with equal and maximum area the algorithm will select the first of the largest triangles encountered. In practice the displacements are small and such a scenario is unlikely to occur and should have little effect on the outcome. The force $\boldsymbol{f}_{Amax}$ was introduced to increase the stability and versatility of the deformation process without excessive smoothing or within-surface motion. The force has the added benefit of being able to attract vertices into long thin regions. With the addition of this force the need for the smoothing force to achieve stability was eliminated (with the exception of an intermediary step for the hippocampus and lateral ventricles). The equations for the three internal forces

$$\boldsymbol{f}_n = \boldsymbol{s_n}, \tag{2.5a}$$

$$\boldsymbol{f}_t = \boldsymbol{s}_t, \tag{2.5b}$$

$$\boldsymbol{f}_{Amax} = A_{max}\frac{\boldsymbol{s}_{Amax}}{|\boldsymbol{s}_{Amax}|}, \tag{2.5c}$$

where $\boldsymbol{s}_n$ is the surface normal vector, $\boldsymbol{s}_t$ is the surface tangent vector, A_{max} is the largest area of the adjacent triangular patches and $\boldsymbol{s}_{Amax}$ is the vector that bisects the largest adjacent triangle. $\boldsymbol{s}_n$ and $\boldsymbol{s}_t$ are projections of the difference vector between a given vertex and the mean neighbour coordinates (equation 2.6a) onto the local surface normal and tangent vector respectively (equations 2.6b and 2.6c). The surface forces, $\boldsymbol{f}_n$, $\boldsymbol{f}_t$ and $\boldsymbol{f}_{Amax}$ that act on each vertex respectively act along the same direction as $\boldsymbol{s}_n, \boldsymbol{s}_t$ and $\boldsymbol{s}_{Amax}$, which are given by

$$\boldsymbol{s} = \boldsymbol{v}_0 - \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{v}_i, \tag{2.6a}$$

$$\boldsymbol{s}_n = (\boldsymbol{s}\cdot\hat{\boldsymbol{n}})\hat{\boldsymbol{n}}, \tag{2.6b}$$

$$\boldsymbol{s}_t = \boldsymbol{s} - \boldsymbol{s}_n, \tag{2.6c}$$

where $\hat{\boldsymbol{n}}$ is the local surface normal (unit vector) for the vertex $\boldsymbol{v}_0$, N is the number of neighbouring vertices and $\boldsymbol{v}_i$ is the i^{th} neighbour.

The net force applied to each vertex is the weighted sum of all forces. The weightings are empirically determined and are generally consistent across structures with the exception of the hippocampus and lateral ventricles. The hippocampus and lateral ventricles are exceptions due to their more complex shape and the relatively

large amount of shape variation that exists in the training images. To construct the statistical shape model, the mesh parameterization process is performed only once. The mesh deforms for a set number of iterations such that the deformation process reaches a steady-state where the vertices maintain only a small oscillation about a stationary location in space.

The number of iterations is chosen empirically based on careful studying of the deformation process and the underlying training images. Across the training images the number of iterations is typically constant for a given structure and is consistent across structures. In particular, the lateral ventricles require tuning of this parameter in problematic cases; this is due to the large amount of shape-and-size variation that exists for the lateral ventricles. The tuning of this parameter is typically required because of the tuning of the other parameters; this is because the other parameters affect the number of iterations required to reach steady-state. For a limited number of cases of the lateral ventricles, the small oscillations at steady-state caused self-intersection of the mesh in the very thin horns; in this case the number of iterations was chosen such the mesh was in steady-state but had not yet reached the point of self-intersection.

The deformation process used to fit to the lateral ventricles and hippocampus differed from that of the other structures by the inclusion of an additional deformation step. The additional step was designed to cope with the more extreme variation in size and shape of these structures. Prior to the deformation process used for the other structures, the mesh is first deformed using a positive weighting on $\boldsymbol{f}_n$ thereby enforcing smoothness in the mesh over the deformation process. The smooth mesh representation is then deformed using a zero weighting on $\boldsymbol{f}_n$. The idea is that in the

first step the general shape is captured whilst taking advantage of the stability that $\boldsymbol{f}_n$ provides for large deformations. The second step, the refinement deformation, then captures the fine details.

Periodically the algorithm tests for self-intersection in the mesh. In the case where self-intersection occurs, the process is reset and the weight on the vertex-distance regularization force ($\boldsymbol{f_t}$) is incremented. In this way the deformation process only allows the within-surface motion required to reach a steady-state. The deformation process is depicted in figure 2.12.

2.7.3 Mesh Initialization

When parameterizing each structure, an "average mesh" is used for each subject to initialize the deformation. The choice of average mesh is critical as it defines the vertex topography used for the model. Smoothness of the mesh, vertex density and the uniformity of vertex spacing are major factors to consider when creating an "average" mesh. These characteristics impact the deformable model's ability to adapt to new subjects and to capture detail. The aim is to have a smooth mesh with uniform vertex spacing. The number of vertices chosen also defines the dimensionality of the PDM and hence the number of parameters we need to estimate. We will describe three methods that are used to generate the average mesh.

The first method is to shrink an initial sphere onto the most representative subject for that structure. Since we do not desire very fine detail (because of the roughness due to voxel labelling), the smoothness force described in equation 2.5a was given

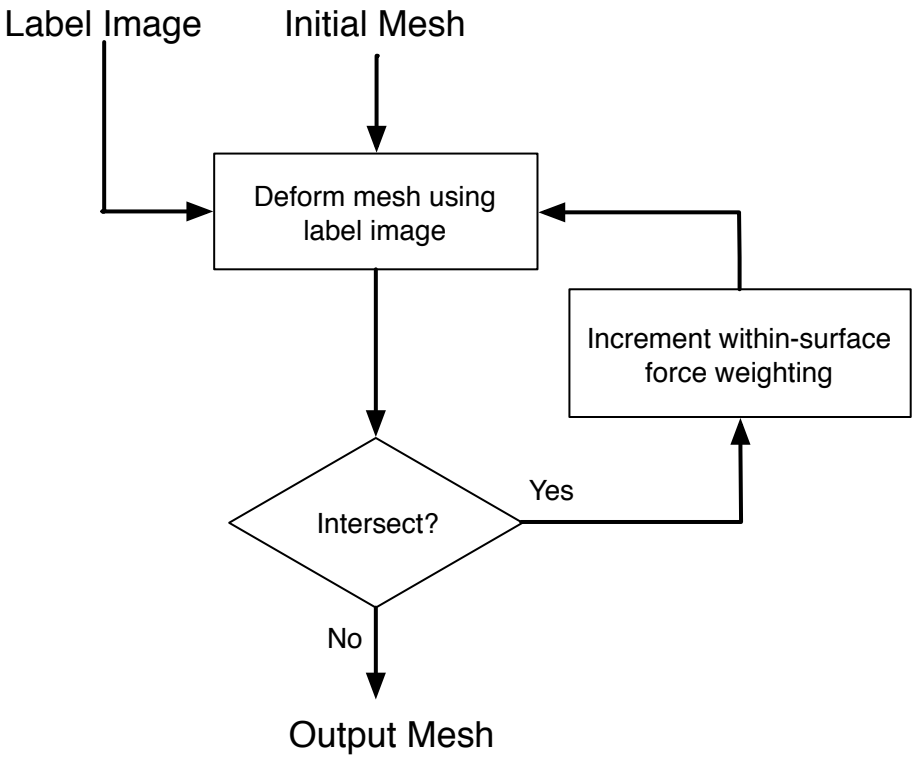
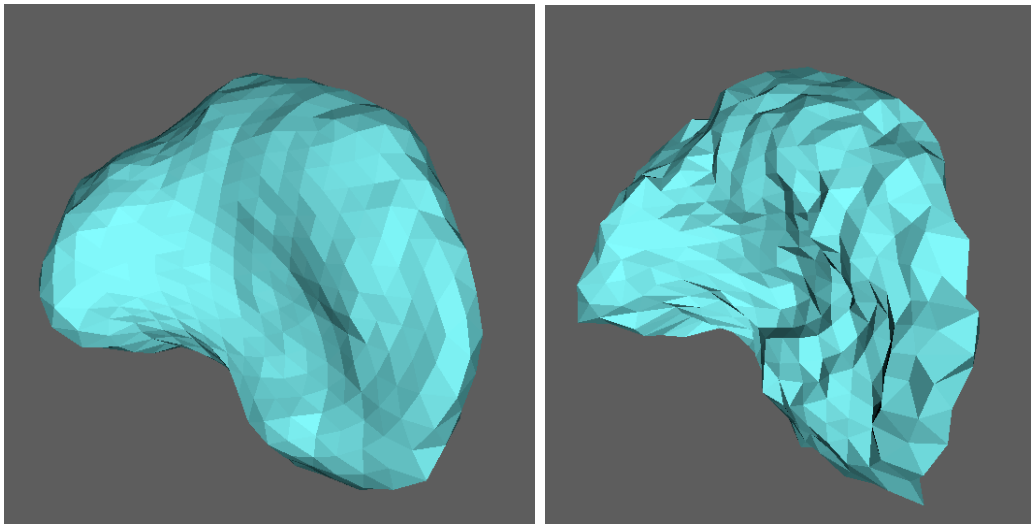


Figure 2.12: Deformation process to parameterize a labelled image.



(a) "Average" Mesh (b) After Deformation

Figure 2.13: a)"Average" mesh for the left putamen used to initialize the deformation. b) Mesh representation of the left putamen for a single training subject after deformation.

a large weighting. By enforcing smoothness we are allowing the deformable model greater robustness to larger deviations in shape from the training data. Figure 2.13 shows the average left putamen mesh and the putamen mesh for a particular individual. The mesh after deformation is visibly rougher than the average mesh; the roughness reflects the coarseness of the voxel boundary in the manual labels. The most representative subject for a given structure onto which the sphere in shrunk was defined by the maximum cumulative Dice overlap (based on the manual labels). The cumulative Dice overlap is given by

$$CDO = \sum_{i \epsilon N} \frac{2TP_i}{2TP_i + FP_i + FN_i}, \tag{2.7}$$

where i is the subject index, N is the number of subjects, TP, FP, and FN are the respective true positive, false positive, and false negative voxel counts.

The second method for generating an "average mesh" is to use marching cubes on a thresholded mean label image. This is necessary for complex structures where concavities result in a non-uniform vertex density across the mesh when shrinking a sphere. For our purposes marching cubes captures too much detail from an individual subject mask, thus a thresholded mean (across subjects) label image is used instead of the most representative subject. Marching cubes was implemented using the VTK library [53]. Vertex density of the mesh was controlled by sub-sampling the label image which results in a reduction in the number of mesh vertices produced by marching cubes (as discussed in section 2.3.1). For additional smoothness the marching cubes output was deformed to the label image using the deformation process described in section 2.7.2.

The final method uses the mean mesh from previously created models. The use of this technique is partially a result of the fact that the data was obtained in parts over a long span of time. Given a model constructed from a finite data set, an average mesh was formed by averaging corresponding vertices across subjects; this produces a visually smooth mesh. It is therefore sensible to use this estimate of the true mean as the starting point for future deformations. Using the final method, the average meshes currently used were constructed from a 127 subject subset of the total 317 training images.

2.8 Evaluation of Parameterization Accuracy

To evaluate the accuracy of the parameterization method, the distance from the mesh vertices to the nearest labelled voxel was measured. For all evaluations the *training label images* are regarded as the gold standard. In addition, volumetric overlap between the filled mesh and original manual labels was used. The *training label images* were parameterized only after their transformation to MNI152 space at 1mm resolution. Therefore for this chapter all overlap/distance measures are reported for the images at 1mm isotropic resolution.

A limitation of the deformable model approach was that it was required to truncate the posterior horn of the lateral ventricles. The posterior horns proved difficult to fit to because of the large variation in length (partly due to partial volume effects). Furthermore the brain stem and fourth ventricle were combined for fitting as the fourth ventricle is very small and passes through the brain stem. For evaluation the manual labels were used as supplied, the posterior horns are included when evaluating the ventricle fit and the fourth ventricle is excluded when evaluating the brain stem fit.

Two vertex-to-voxel summary distance metrics were used: 1) maximum distance and 2) root-mean-square distance. The Euclidean distance from the vertex to the centre of the nearest voxel was used; when interpreting the distance results it is important to bear in mind that the vertex location is a continuous quantity that is being fit to discrete data. For example, a vertex within a distance of less than 0.5mm (for 1mm isotropic resolution) of the voxel centre is guaranteed to be within that voxel. To give a better sense of the quality of fit for a given distance, figure 2.14 depicts a

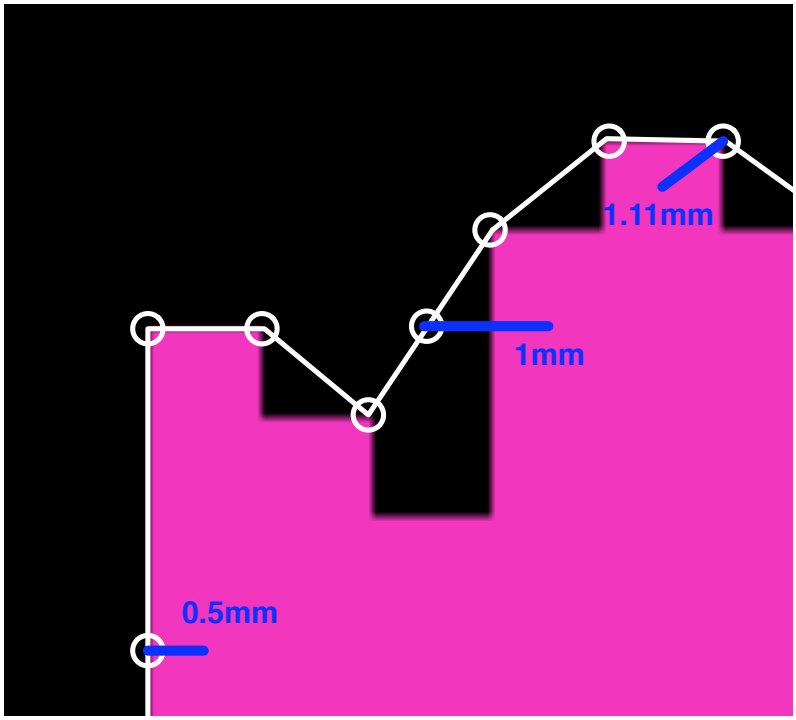


Figure 2.14: Illustrative example of vertex-to-voxel distances for a contour fit to label data at 1mm resolution. The vertex-to-voxel distance is shown for three vertices.

synthetic example of a continuous contour fit to discrete voxels with the vertex-to-voxel distance labelled for selected vertices. The distance was measured to the centre of the voxel instead of its edge since it should be a close approximation to the distance to the surface and is simpler to compute.

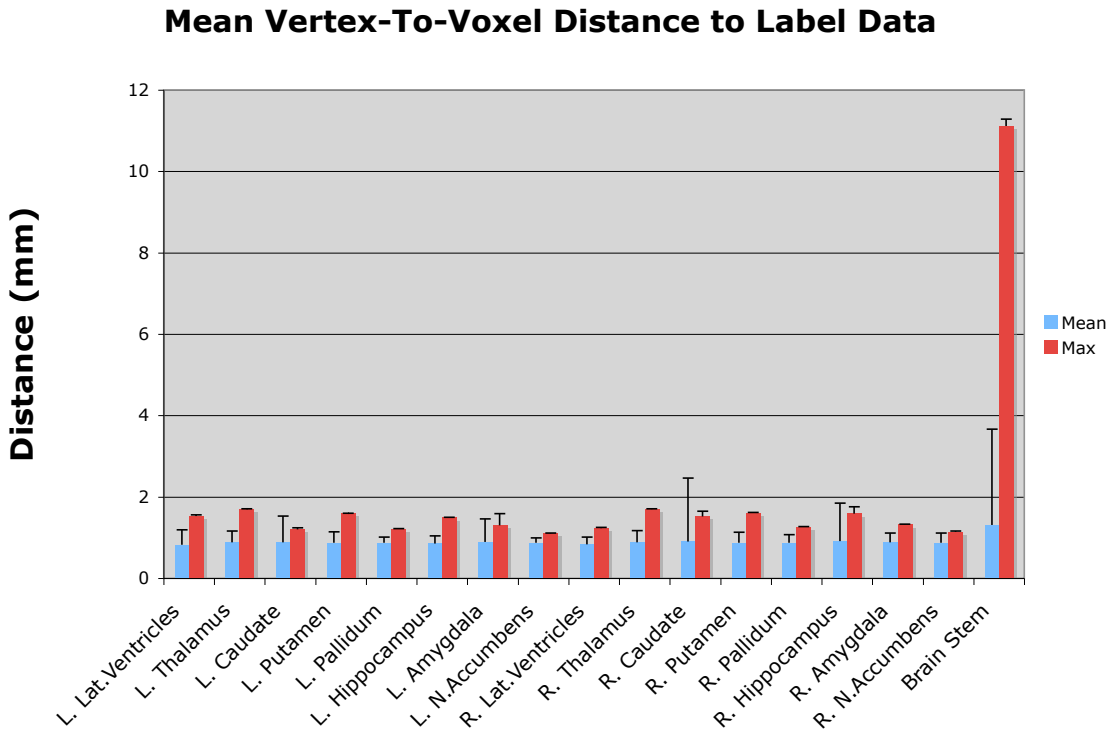
For each structure figure 2.15a summarises the vertex-to-voxel distance for each structure and were evaluated over the 317 training datasets. The mean mesh distances reported are less than the voxel resolution and the maximum is only just greater. Given the continuous to discrete nature of the boundary fit we deem this to be sufficiently accurate. The large maximum distance observed for the brain stem is due to the inclusion of the fourth ventricle for the fitting but not for the evaluation.

Our statistical shape models are mesh-based models and hence the output is inherently a mesh. In practice, users generally desire volumetric representations for visualization or as input into other applications (such as ROI selection for fMRI). Furthermore, many segmentation algorithm evaluations report volumetric overlap for validation. If volumetric output is to be used, volumetric overlap ought to be provided for both the surface parameterizations as well as for the segmentation results. The volumetric overlap is measured using the Dice overlap metric as given by

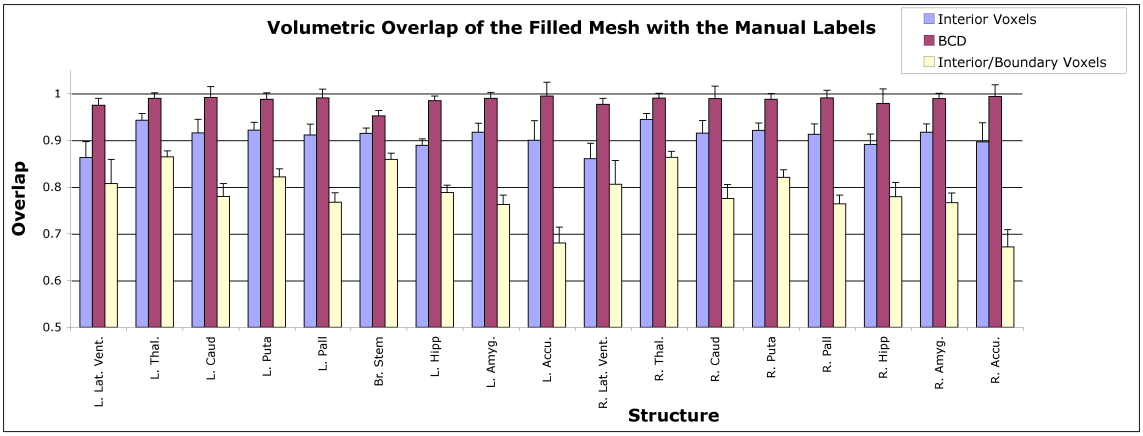
$$D = \frac{2TP}{2TP + FP + FN}, \tag{2.8}$$

where TP is the true positive voxel volume, FP is the false positive volume, and FN is the false negative volume.

The volumetric output used to compute the Dice metric is the result of filling the output mesh. The mesh filling process consists of two steps: 1) drawing the mesh outline, and 2) filling the interior. As a consequence of these two steps, we are able to classify an output voxel as belonging to the boundary or the interior. In practice the mesh boundary is a continuous boundary passing through a discrete voxel; consequently we are often uncertain whether a voxel "belongs" to the structure or not. This is exemplified in figure 2.15b where we use the Dice overlap metric based on the interior voxels, the union of the interior and boundary voxels as well as a boundary corrected dice (BCD). The BCD is calculated by replacing the boundary voxels with the underlying manual label, then recalculating the Dice metric based on the corrected boundary. In practice when correcting boundary voxels, membership is decided based on image statistics (this will be discussed in further detail in chapter 3). By evaluating the mesh fits using BCD we obtain an upper bound on our fitting



(a) Mesh distance



(b) Volumetric Overlap

Figure 2.15: a) Vertex-to-voxel distance for the fitted mesh to the manual labels. b) Dice overlap between the filled mesh and the manual labels. "Interior Voxels" refers to the overlap of only the interior voxels and the manual labels. "BCD" refers to the overlap of the filled mesh (with all the boundary voxels corrected) and the manual labels. "Interior/Boundary Voxels" refers to the overlap of the combined interior and boundary voxels with the manual labels.

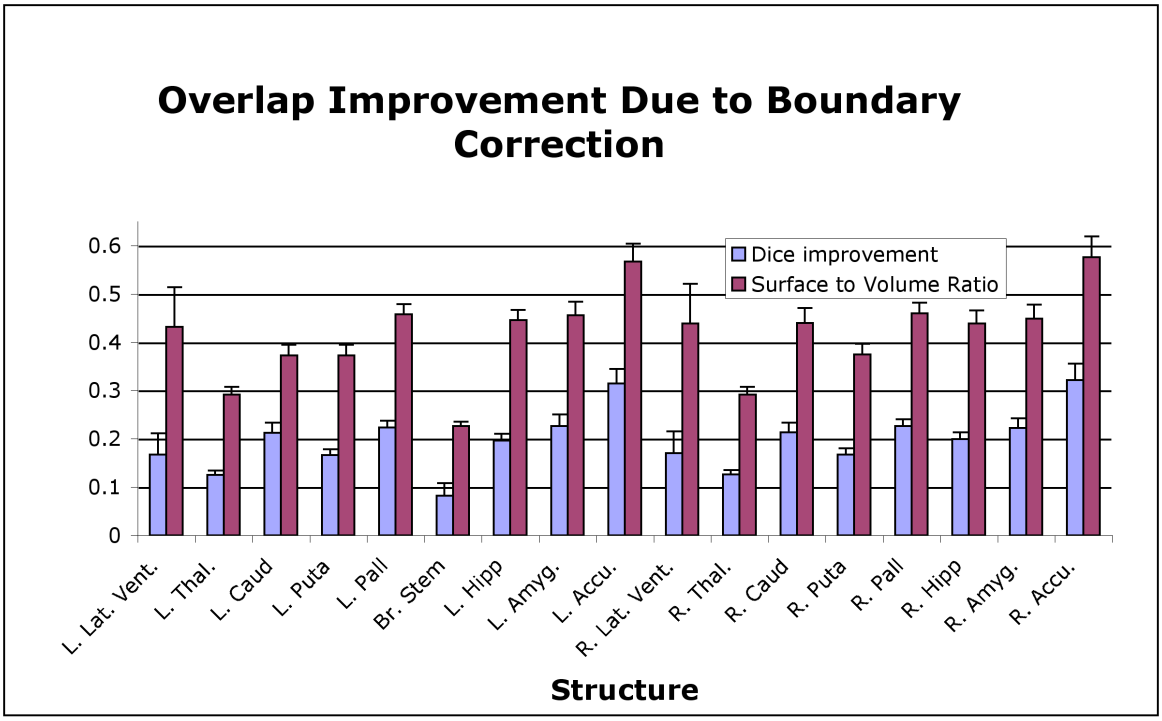


Figure 2.16: Improvement in Dice overlap due to boundary correction. The mean and one standard deviation for each of the parameterized structures is shown. In addition, the surface to volume ratio for each structure is shown to demonstrate the strong correlation with Dice improvement.

performance over all possible boundary correction methods. From figure 2.15b it is clear that an uncorrected boundary produces sub-optimal results. The impact of boundary correction is most prominent in the accumbens and least so in the thalamus. The impact is related to the ratio of surface area to total volume. We approximate the ratio by the boundary voxel volume divided by the total volume (including the boundary). The improvement in Dice overlap due to boundary correction is plotted in figure 2.16 along with the surface to volume ratio for each structure. The correlation between the mean overlap improvement and the surface to volume ratio across the 17 structures is 0.9002. The t-statistic for the correlation coefficient is 8.0 indicating that the correlation between the two metrics is significant ($p < 0.00001$). The t-statistic for the correlation coefficient is given by the standard formula

$$t = \frac{r}{\sqrt{\frac{1-r^2}{N-2}}}, \tag{2.9}$$

where r is the correlation coefficient and N is the number of samples.

2.9 Conclusions

By using a common reference space and applying the identical alignment procedure across the training images the surface parameterizations will reflect the residual variation over and above the normalization procedure. The normalization to a common reference space is important in order to automate the segmentation procedure. Provided that the new image is aligned using the same protocol as that used for the training images, the statistical shape models will reflect the variation in new images

after alignment. Consequently, the more robust and accurate the alignment procedure, the less vertex variation should be seen in the parameterizations. By reducing the structural variation after normalization our search space will be reduced when applying the shape and appearance model to new images. When designing the alignment procedure for the training images it was crucial that it relied solely on the T_1-weighted intensity images and not the label images. When considering new images, corresponding label images are not available and thus if the label images were used in any way the shape model would not necessarily reflect the same variation to the normalization procedure as would the new images. Thus the models are specific to the normalization procedure, and strictly speaking the *training label images* should be re-parameterized for any changes to the alignment process.

As opposed to using a mathematical criterion based on the statistical shape model (e.g. model compactness) as a means of parameterizing a structure, the deformable model provides a physical model for the structural deformation of structures between subjects. The deformable model does not explicitly optimize for a specific shape model except in that it defines correspondence across subjects as being inherent to the physical deformation process. In other words the correspondence is a function of the underlying label image rather than the resultant shape model. By using this approach a statistical model of vertex location should reflect the same deformation process. A possible pitfall is that poor design of the deformation process will lead to inefficient models and may lack meaningful interpretation, thus resulting in poor segmentation performance.

The deformable models do require the setting of several weighting parameters. The parameters are set based on carefully studying the data and the deformation process,

and they are typically constant for a single structure. In fact the same parameters were used for all the structures except for the hippocampus and lateral ventricles. Despite this drawback, the parameterization need only be performed once as it will be used to construct a shape and appearance model. The surface-based deformation method for obtaining correspondence is similar in nature to the non-linear registration method as proposed by Frangi et al. [28] except that ours is a surface-based deformation model rather than a volumetric-based one.

One novel contribution to the deformable model was the area-based surface force. The force was found to be necessary to produce the accurate parameterizations whilst preserving the point correspondence that was necessary. The force provides stability to the mesh without excessive within-surface motion. This was necessary to prevent self-intersection whilst preserving vertex correspondence. In addition, the area-based force eliminates the need for surface smoothing such that we may achieve more accurate representations of the underlying manual segmentation. Apart from general improved stability, the force has the added benefit that it aids in the propagation of vertices into long narrow regions that pose difficulties when using only a tangential and normal force. Furthermore the force facilitates the local contraction/expansion of the mesh which may propagate down the mesh, providing flexibility to the deformation process. The forces cause the vertices to jitter or oscillate a small amount as they displace towards the boundary.

The distance results reported for the mesh parameterizations indicate that the procedure produces an accurate mesh representation of the underlying label images. The RMS distance was below the voxel resolution whilst the maximum distance, being just over a voxel dimension, was considered to be in good agreement given that the

distance was measured to the centre of the voxel. Furthermore some of the larger inaccuracies may be due to the smoothing of the jagged boundary that can be seen in the posterior-anterior direction in the label images. The rough boundary is effectively noise in the manual segmentation process.

The volumetric overlap provides equally persuasive results for the accuracy of parameterization, particularly when considering boundary correction. The use of boundary correction relates to the conversion between discrete volumetric images and surface representations. Where this chapter mostly concentrates on moving from volumes to surfaces, the boundary correction addresses the imperfection in converting back from surfaces to voxels. We will revisit the issue of boundary correction in the following chapter when discussing the conversion of the surface output to volumetric output.

The deformable model approach allows for large variation in size and shape as exemplified by the lateral ventricles. It does not allow for topological changes in the way that 'level-set based shape models' do, which is viewed as an advantage given that we do not expect variations in the topology of subcortical structures. The deformation process was carefully studied and visualized to ensure correspondence in the deformation process. Point correspondence is difficult to validate when not considering the shape model and its performance, thus the validation of point correspondence will be postponed to the following chapters when discussing the shape and appearance model and its applications. Because no ground truth exists for point correspondence we rely on segmentation performance and meaningful shape statistics to serve as validation of the correspondence.

Chapter 3

Bayesian Models of Shape and Appearance

3.1 Introduction

The previous chapter presented a parameterization method for volumetric label images that preserves point correspondence. The motivation having been to construct a point distribution model (PDM) of shape and intensity, so that they may be applied to automated subcortical segmentation. This chapter focusses on the mathematical model that is used for shape and intensity, their fitting to new images as well as the validation of the segmentation.

The concept of the PDM and active shape model (ASM) was introduced in chapter 1, and further elaborated on the PDM with regard to point correspondence in chapter

2. We will now revisit the concept of the PDM and ASM since they are core concepts underlying the model presented in this chapter. The PDM models the spatial variability of vertices as a multivariate distribution. The multivariate distribution considers all vertices simultaneously, thereby modelling the covariation in vertex location that was present in the training data. The PDM requires correspondence between vertices across the surface parameterizations of each subject.

The ASM, proposed by Cootes et al. [11], uses a multivariate Gaussian PDM to model shape. The ASM parameterizes shape as the linear combination of the mean shape and the eigenvectors of the spatial coordinates. The shape space defined by the ASM is given by

$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{U}\boldsymbol{D}\boldsymbol{b}^t, \tag{3.1}$$

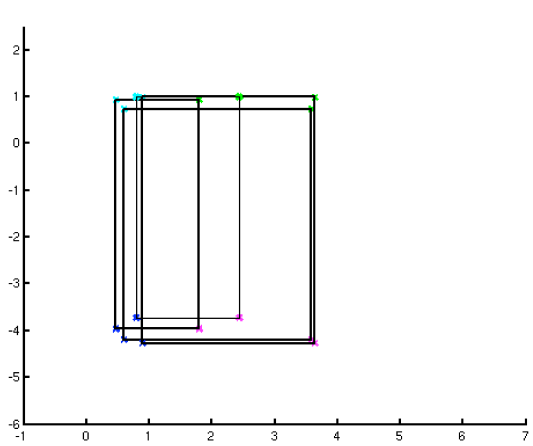
where $\boldsymbol{x}$ is a shape vector containing the spatial coordinates for a single shape instance, $\boldsymbol{\mu}$ is the mean shape, $\boldsymbol{U}$ is a matrix of the eigenvectors, $\boldsymbol{D}$ is a diagonal matrix of the singular values, and $\boldsymbol{b}$ is a vector of the mode parameters.

New shape instances may be synthesized by varying the mode parameters b_i (the i^{th} component of $\boldsymbol{b}$). The singular values are in fact the standard deviations associated with each eigenvector (mode of variation), the mode parameters thus indicate the number of standard deviations along each eigenvector the shape lies. Therefore, under the multivariate Gaussian model, the log-probability of an estimated shape instance is proportional to the sum of the squared mode parameters. Consequently, the closer the mode parameters are to zero, the more probable the shape. The ASM is fit to new images by optimizing the mode parameters, typically using the image's intensity gradient at the vertices as the feature being maximized. Restricting the search space

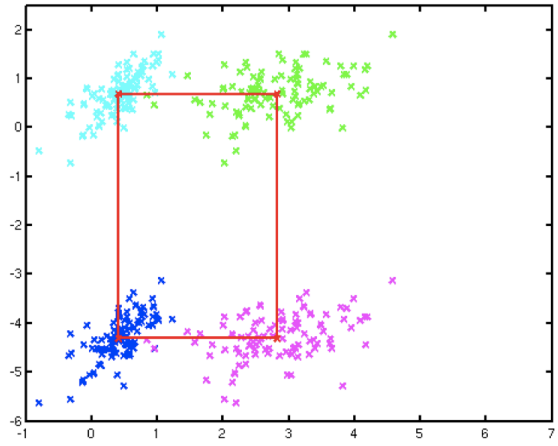
to linear combinations of the eigenvectors imposes strict shape constraints, penalising unlikely shapes.

To illustrate the concept we will revisit the synthetic example presented in chapter 1. The hundred synthetic rectangles that were randomly generated were each parameterized using four vertices, each corresponding to a corner of the rectangle. Furthermore, the rectangles were generated such that vertex correspondence was preserved. The random sample set and mean rectangle is depicted in figure 3.1(b). Figures 3.1(c) and 3.1(d) respectively show the shape variation modelled by the first two modes of variation. The series of rectangles (in blue) is generated by varying the respective mode parameter (b_i) between -3 and 3. Therefore the shape variation depicted represents ± 3 standard deviations along each mode of variation. It is clear from the figure that for this example the ASM may only synthesize new rectangles. The fact that the shape space spans only rectangles is a reflection of the fact that all the training samples were rectangles. This is a desirable property if we believe that the true shape space does indeed consist of only rectangles (as the sample set would indicate). The first and second modes of variation correspond to the major and minor axes of ellipsoids fit to the vertex point clouds.

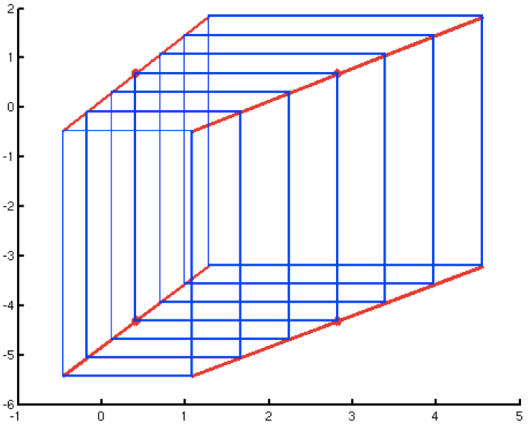
In our work, the shape parameterization is the same as that used for the ASM, however the probabilistic model is slightly altered by taking into account the finite sample set. The multivariate Gaussian assumptions are retained, however by posing the problem through a Bayesian perspective; given a finite training set, we obtain a multivariate Student distribution for the estimated model. Within the Bayesian derivation, a prior is added to the covariance estimate (this helps cope with the low sample density). The chosen prior does not alter the eigenvectors, however it increases



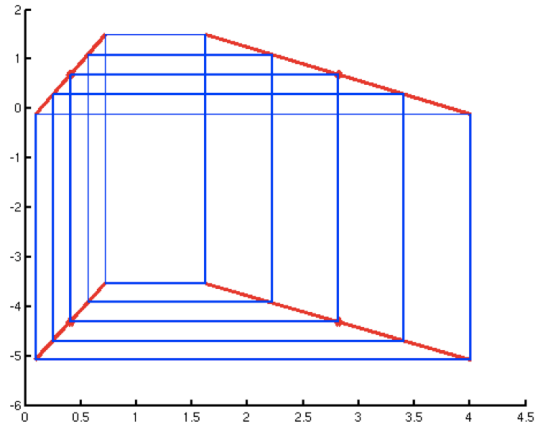
(a) First four training samples (rectangles)



(b) Training Set with Mean Shape



(c) First mode of variation.



(d) Second mode of variation.

Figure 3.1: a) Four randomly generated rectangle. b) The corner vertices of a randomly generated training set of 100 rectangles with the mean rectangle in red. c) New shape instances (blue rectangles) generated by displacing the vertices ± 3 standard deviations along the first eigenvector(mode of variation). The eigenvector is depicted in red. d) New shape instances (blue rectangles) generated by displacing the vertices ± 3 standard deviations along the second mode of variation. The eigenvector is depicted in red.

the variance (eigenvalues). The addition of the prior has some important practical implications that will be discussed later in the chapter.

The ASM was later extended to incorporate appearance (AAM [12]) such that variation in shape is associated with variation in intensity. The details of the AAM will be discussed in section 3.1.2. The main difference between the model proposed in this chapter and the AAM is in how the relationship between shape and intensity is modelled. We model the relationship using the conditional distribution of intensity given shape, and by doing so we eliminate the need for arbitrary weighting parameters that relate shape to intensity variance. The conditional distribution provides an analytic expression (with parameters estimated from the training data) for the mean and covariance of intensity given a shape instance. In contrast, the AAM does not model the intensity variance about the predicted intensity. The intensity covariance matrix weights the contributions of the intensity samples by their uncertainty; the AAM considers all intensity samples equally. An additional application of the conditional distribution is to provide a probabilistic framework for incorporating inter-structure information, whereby a shape distribution may be constrained by the known location of another shape. The particulars of our model will be discussed in further detail throughout the chapter.

In this chapter, we will review existing segmentation methods with a particular focus on neuroanatomical segmentation using shape and intensity priors. The incorporation of learnt shape and intensity priors brings robustness in the presence of noise. As with our model, most of the current methods make use of trained priors/models to aid in the segmentation. A limitation of learnt models such as for shape and intensity is their dependency on training data, and this in turn gives rise to the issue of dimensionality.

Dimensionality will be discussed first, followed by the review of existing segmentation techniques, and then by the proposed model, segmentation technique and validation.

3.1.1 The Issue of Dimensionality

Dimensionality becomes important when training models since the sampling density (the sample size relative to dimensionality) impacts on the reliability and accuracy of the parameter estimation. In the case of a parametric model we use a set of samples (training data) to estimate the parameters of the distribution we are interested in. In our application our samples/training data is a set of surface parameterizations with corresponding intensity samples. Thus we are trying to model the statistical variation of cartesian vertex coordinates (in mm) and gray scale values (sampled along the surface normal at each vertex). When training the shape and intensity models, assuming a multivariate Gaussian model, we are estimating the mean and covariance of the mesh vertices and intensity samples respectively.

In practice, particularly in the medical field, the size of the training set is small compared to the dimensionality of the model. To train the models we are dealing with a largely underdetermined inverse problem (and hence rank-deficient covariance matrices); this is especially problematic in 3D since typically many more control points are required to describe a shape. The dimensionality of the multivariate Gaussian used to model shape is equal to the dimensionality of the data multiplied by the number of control points. For example, a 3D mesh representation with N vertices would have a dimensionality of $3N$. Specifically, our 3D model of the left putamen consists of 642 vertices and thus the shape model would have a dimensionality of

Structure	Number of Vertices
L/R Accumbens	642
L/R Amygdala	642
L/R Putamen	642
L/R Pallidum	642
L/R Thalamus	642
L. Hippocampus	732
R. Hippocampus	664
L. Caudate	970
R. Caudate	1068
L. Lat. Ventricles	768
R. Lat. Ventricles	592
Brain Stem	642

Table 3.1: Number of vertices used per structure.

3×642, which by far exceeds our training set of 317 manually labelled images. The dimensionality of appearance models is increased by a factor equal to the number of intensity samples. In the case of the left putamen the intensity samples would add an additional 13×642 dimensions. The number of control points used within each of the structures being modelled is provided in table 3.1.The reason for the asymmetry in the number of vertices for the hippocampus, lateral ventricles and caudate is due to asymmetry in the average shape for each structure combined with the fact that the marching cubes is used to generate the surface topology for the structures. As discussed in the previous chapter, the number of vertices generated using marching cubes is dependent on the size and shape of the image region, therefore asymmetry in the average structure results in an asymmetry in the number of vertices.

For a single shape model we are dealing with a dimensionality ranging from 1776 to 3204 (9472 to 17088 for the joint shape/intensity model). The situation worsens substantially when modelling multiple structures since the dimensionality increases

whilst the number of subjects remains unaltered. Typically, the solution is to apply a singular-valued decomposition (SVD) to determine the eigenvectors of the space spanned by the data (ignoring the null space). The null space includes the eigenvectors that span the unseen variation from the unsampled population. The probabilistic model presented in this report copes with the underdeterminancy through the incorporation of a prior. This problem of low sample sizes and large dimensionality plagues most trained segmentation techniques in the medical imaging field.

3.1.2 Review of Segmentation Methods

Contour and Surface-Based Segmentation

Contours and surfaces respectively refer to the 2D and 3D representations of a structural boundary. The contour and surface are boundary parameterizations such that the shape is represented as a set of connected vertices. By using such a representation, shape constraints may be imposed using vertex-based shape metrics or statistics. The deformable model that was introduced in the previous chapter is an example of a surface-based segmentation method.

For the purposes of segmentation, deformable models require good initialization and may be susceptible to leakage. In the previous chapter the underlying truth was known and the segmentation was used to obtain a surface parameterization. When applied to medical images, driven by image intensities and under shape constraints the model is deformed in the image space to determine the true underlying boundary (segmentation). Good initialization is required because the image (external) forces

are derived from local intensity measurements. Since multiple boundaries may exist within a local neighborhood the deformable model may fall into a local minimum at an undesired boundary. The problem of leakage refers to the surface bleeding outside of the true boundary due to weak contrast and/or image noise. To alleviate these problems the active shape model (ASM) [11] was introduced to constrain the search space to only plausible shape instances (as defined by the trained model).

The ASM has become widely used in the field of machine vision and medical image segmentation over the past decade. ASMs model the vertices (control points) of a structure as a multivariate Gaussian distribution. Shape is then parameterized in terms of the mean and eigenvectors of the vertex coordinates. New shape instances are constrained to the space spanned by the eigenvectors. Consequentially, if the dimensionality of the shape representation exceeds the size of the training data, the only permissible shapes are linear combinations of the original training data. Within a given Mahalabonis distance, the search space is typically restricted to plausible shapes.

The active appearance model (AAM) is an extension of the ASM that models the relationship between shape and appearance [12]. In addition to shape, the AAM models the intensity distribution as a multivariate Gaussian and can thus be parameterized in terms of its mean and eigenvectors. The AAM relates shape and intensity parameterizations by learning a diagonal weighting matrix from the training set. The weighting matrix relates the mode parameters for the shape model to the mode parameters for the intensity model. The weighting parameters are required to relate a unit of shape variance to a unit of intensity variance. To estimate the weighting parameters, the training data is revisited and each mode parameter for the shape model is system-

atically varied and the intensities sampled. The weighting parameters are estimated as the RMS change in intensity per unit change of the given mode parameter. For example, given the optimal shape for a training subject, in order to estimate the first weighting parameter the shape is deviated from the optimal solution by varying the first mode parameter and then sampling the intensities for the sub-optimal shape. The weighting parameter is then calculated by the RMS difference between the intensity samples for the optimal and sub-optimal shape (across all training subjects), divided by the change in shape mode parameter. Using the weighting matrix, the separate intensity and shape parameterizations are combined into a single model. The AAM is fit to new data by minimizing the mean-squared-difference between the predicted intensities (given a shape instance) and the observed image intensities. Many of the surface-based techniques are extensions and/or variations of the ASM and AAM. In this thesis, we describe a novel probabilistic framework that models the relationship between shape and intensity, eliminating the need for empirical weightings between shape and intensity.

Deformable models, ASMs and AAMs have been applied to the segmentation and/or modelling of the heart, spinal cord, prostate, as well as several neuroanatomical structures [4, 29, 51, 54, 63]. Deformable models provide flexibility and do not require explicit training, though they are sensitive to initialization and noise. ASM/AAMs may lead to greater robustness, however are more rigid than deformable models and may be over-constrained and hence not generalize well to the unsampled population (particularly for small amounts of training data relative to the dimensionality). Many methods attempt to find a balance between the flexibility of the deformable model and the strict shape constraints of the ASM by fusing learnt shape constraints with

the deformable model.

An alternative to the deformable model that came about around the same time are level-set techniques. They have become widely used in the field of medical image segmentation and are based on the principle of front propagation [50]. The level-set method uses an implicit representation of a contour by expressing it as the level line of an embedded function; most commonly the zero level line is used (zero level-set). There are various methods for evolving the contour (propagating the front), the most popular of which is through a time-dependent function derived from partial differential equations. Alternatively, a level-set evolution equation may be derived from the minimization of an energy function (similar to that used for deformable models). The most relevant work to ours is the statistical approach to level-set segmentation whereby the contour is propagated such that the probability of the image partitions given the observed intensities is maximized [14]. Bayes' rule may be used to express the probability as a proportionality to the product of the probability of observed intensities given the shape partitions and the probability of the image partitions. Cremers et al. [14] describe the simplification of the intensity likelihood given a shape partition using independent identically distributed (IID) assumptions such that it becomes equal to the product of the intensity probability given a partition over each voxel. Under these assumptions correlations between neighbouring voxels are neglected and it does not model local intensity features. The incorporation of statistical shape priors into a segmentation/registration framework using level-sets is proposed by Leventon et al. [41], Tsai et al. [59] and Pohl et al. [46]. Despite framing the segmentation/registration problem using the joint distribution of shape and intensity they only retain the shape information and not the joint shape and intensity

information present in the training data. The methods proposed by Leventon, Tsai and Pohl will be discussed in further detail later in this section.

Shen et al. [54] use an adaptive-focus deformable (AFDM) and shape (AFDSM) model to segment the ventricles, caudate nucleus, and the lenticular nucleus. Neighbourhood layers are created for each vertex such that the first layer is comprised of direct neighbours, the second layer is the direct neighbours of the first layer and so on. The layers define the spatial extent around a vertex that is incorporated in the computation of the shape metric. The layers provide a means of changing focus between global and local shape metrics, such that a metric becomes more localized as the number of layers is reduced. Therefore, successively reducing the number of layers used amounts to a global to local optimization scheme. An additional criteria for the inclusion of a vertex in a layer (apart from vertex connectivity) is that any vertex within a defined Euclidean distance of a layer-one neighbour becomes a member of that layer; this helps prevent mesh self-intersection as well as intersection between disconnected meshes (i.e. different structures). Each vertex has an attribute vector that describes the geometry from a local (lower layers) to a global scale (upper layers). The volume of the tetrahedron defined by the vertices contained within a given layer was chosen as the shape attribute. The attribute is evaluated for several layers and is used to impose shape constraints on the deformation. To ensure affine-invariance, the layer specific volumes (attribute) are normalized by dividing by the total volume contained within the entire mesh. The model is deformed based on a model and data energy metric. The model energy (shape constraint) is proportional to the difference between the model attribute vectors (attributes for a mean mesh) and those for the deformed mesh. The data energy is derived from the image gradient. Unlike

conventional deformable models, surface segments are deformed rather than isolated vertices. The AFDM was extended to incorporate a shape model like that was used for the ASM. To construct the model, point correspondence was obtained through the application of the AFDM to manually labelled data. An adaptive-focus search strategy is used such that more reliable vertices are fit prior to the remaining vertices. The vertex reliability is learnt *a priori* and is proportional to the difference between the gradient strength at the vertex and the mean neighbourhood gradient. The unequal weighting of vertices based on a learnt measure of vertex reliability provides a means to increase robustness, but unfortunately adds another empirical weighting into the deformable model. In contrast to this, by posing our shape/appearance model in a probabilistic framework the learnt intensity covariance matrix serves to weight the vertices appropriately without the need of an additional empirical weighting.

Pitiot et al. [51] approaches segmentation as a problem of fitting deformable templates to the boundaries of target structures. The models (templates) are constructed in a reference space and non-linear registration is used to transform the template into a native image space. The shape templates are iteratively deformed using internal regularization, external image coupling and global shape constraint energies. The force applied to a vertex is proportional to the derivative of the total energy that is equal to the weighted sum of the three latter energies. The external image coupling energy is proportional to the distance to the strongest gradient. The internal regularization energy is proportional to the mean curvature within a spherical neighbourhood. The global shape constraint energy is imposed by projecting the image coupling energy onto the eigenvectors of the vertex covariance matrix. Inter-structure shape constraints are incorporated via a force that acts along the gradient of the distance map

to a target mesh (the constraining structure). Acting along the negative gradient of the distance map attracts a vertex to the structure, along the positive gradient repels the vertex (avoiding inter-penetration of two surfaces). In addition, texture metrics are used to drive the deformation. A large number of texture metrics are calculated then passed through a feature selection algorithm to determine the most relevant metrics. The texture metrics are used to classify the voxels within a ROI as defined in the standard space, the ROI helps to reduce computation time as well as eliminate confounding structures. Of the tested classifiers, they found that the best classification was achieved by the non-linear classifier (support vector machine). The texture force is derived from distance maps of the classification images. The method requires several empirical parameters whose values vary across structures. It is also unclear whether the values should also vary with signal-to-noise ratio (SNR). In our work we aim to provide a framework that eliminates empirical weighting between image forces and regularization constraints.

Bernard et al. [4] applied statistical shape/appearance models to articulated structures. They cope with the low sampling density by decoupling shape and appearance as well as using hierarchical principal component analysis (PCA). A two-level hierarchical PCA was used to model shape and topology. At the first level, PCA is applied to the vertex coordinates. At the second level PCA is applied to the output parameters from the first level. Rather than a standard AAM appearance model, a shape-free appearance model is created through non-linear elastic registration. The intensity probability distribution is estimated from the training data after a non-linear registration to the mean shape has been applied. Observed intensities are evaluated based on the mean image and the most significant eigenimages. Given the

low sampling densities, the two-level hierarchical PCA provides a means for modelling inter-structure variation, however it does require two levels of optimization, iterating between the two levels of the PCA. The use of a shape-free appearance model is useful for coping with the high dimensionality, however, it may throw away meaningful correlations between shape and appearance that exists within the training data.

Volumetric/Registration-Based Methods

Automated Nonlinear Image Matching and Anatomical Labelling (ANIMAL) [10] treats segmentation as a problem of voxel correspondence between a template and target image. Non-linear registration is used to obtain a displacement field that maps voxel correspondence between the template and target. The registration uses a multi-scale optimization approach with constraints on the smoothness of the deformation field. The inverse transformation is then used to map the voxel label from the template into the native space. This approach is biased towards the anatomy of the particular template of choice, as opposed to the statistical shape model where modelling the variation across a large number of subjects of varying age and pathology aims to reduce such bias.

ASEG [25] proposes a Bayesian formulation of voxel classification and aims to produce a whole brain segmentation that includes white matter, grey matter and various subcortical structures. The classification is made based on the maximum *a posteriori* (MAP) estimate of a label map given observed intensities and a linear transformation to an atlas space. This formulation incorporates learnt shape and intensity priors. The probability distribution of intensity given class membership is assumed

to be Gaussian with the mean and variance estimated from manually labelled training data; the individual voxel statistics are retained for each class. The noise is assumed to be independent and hence the total probability of observing the image intensities given a classification map is calculated as the product of the probabilities at each voxel. The probability of a voxel belonging to a class is defined as the proportion of the voxels (from the training data) that were mapped to that particular voxel and belonged to the given class. In addition, class membership given neighbouring voxels' classification is modelled as an anisotropic Markov Random Field (MRF). The pair-wise neighbour probabilities are stored for each voxel, which is computationally tractable since in practice only a small subset of classifications may exist in the neighbouring voxel. Furthermore, the neighbourhood is restricted to six voxels (each cardinal direction).

The linear transformation attempts to match structures in an image to the atlas. The transform is calculated based on a subset of voxels for each class, as well as class-specific intensity priors. The voxels that are included for a given class are chosen based on two criteria: 1) the maximum class probability for the given voxel is for the class of interest, and 2) the class probability is close to the maximum probability observed for that class. Iterative conditional modes (ICM) [7] is used to maximize the conditional posterior; iterations continue until no voxel classification updates are needed. ASEG models shape implicitly through the voxel-wise probabilities and the MRF. In our work, we pose a more explicit model for shape so that rather than using a static intensity distribution for all shape instances for a given structure, we are able to model the variation in mean and variance of the intensity distribution across changes in shape.

Proposed by Leventon et al. [41], the zero level-set of the signed distance function may be used to parameterize shape whilst circumventing the problem of vertex point correspondence. Given a set of label images in a common reference space, a signed Euclidean distance map is formed for each image. Exterior and interior voxels are assigned positive and negative values respectively. The label images are usually transformed into the same reference space using linear registration. A mean shape is defined by the zero level-set of the mean distance map. The zero level-set refers to the structural boundary as defined by the zero value of the distance map (or transition between positive and negative). After de-meaning the set of distance maps, an eigen-decomposition of the estimated covariance matrix is performed. The eigenvectors represent modes of variation of the distance map which correspond to modes of variation of the zero level-set. A potentially undesirable effect of the distance map approach is that it allows for changes in topology (this may be seen as a desirable property in some contexts, e.g. 2D images). Although the method eliminates the need for vertex correspondence, it does rely on the particular voxel correspondence of the images as defined by the registration to a reference space. Furthermore, the dimensionality of the image space model for shape may be much larger than that for the surface-based model since the number of voxels typically exceeds the number of vertices on a corresponding surface. In practice, the distance map technique is much simpler than determining vertex correspondence, however a surface parameterization provides an explicit model for shape, maintains topology and reduces the dimensionality of the model.

Tsai [59] integrates the ASM shape parameterization into a registration framework. The problem of point correspondence is circumvented by using the zero level-set

of the signed distance function to parameterize shape; PCA is applied to the set of signed distance images. This framework allows for the simultaneous fitting of multiple shapes. Structural covariation is captured by applying PCA on the concatenated distance map representations for each shape. In addition to the level-set models, parameters are added to model changes in pose. The model uses a single mutual information cost function to optimize the model parameters. This method shares the same disadvantages/advantages of Leventon et al. [41], as discussed in the previous paragraph.

Pohl et al. [46] incorporates shape constraints into an expectation-maximization (EM) [19] segmentation algorithm using the distance map representation for shape. The E-step calculates the posterior probabilities (weights) for each structure given the current shape parameter and intensity inhomogeneity estimates. The M-Step estimates the shape parameters and intensity inhomogeneities for the updated weights. The inhomogeneity parameters for the model are treated as nuisance parameters in the Bayesian formulation. The method was later extended to perform simultaneous registration and segmentation such that in addition to inhomogeneity and shape parameters, registration parameters are also estimated within the EM updates [52]. The registration parameters are comprised of global affine as well as local individual structure affine parameters. Instead of explicitly constraining the search space to linear combination of the modes of variation the method uses shape as a probabilistic constraint on the deformation. This differs from our proposed approach where we restrict the search space to the modes of variation as well as constrain the deformation by the shape probability. Also, instead of assuming IID intensities across voxels we model the intensity samples as a multivariate Gaussian distribution. Analytically we

then formulate the conditional distribution of intensity given shape such that we are able to learn the parameters of the distribution from the training data.

Ashburner et al. [2] propose a unifying probabilistic framework for combining tissue classification and template registration approaches to image segmentation and registration. The model for intensity is derived from a mixture of Gaussians (MOG). Bias field (slow varying intensity drifts due to inhomogeneity in the RF field) is modelled as an exponential of a linear combination of low frequency basis functions and is incorporated directly into the model as an additional parameter. Tissue probability maps encode voxel-wise tissue priors and are allowed to deform. The parameters of the model are optimized such that the probability of the observed intensities given the MOG, bias field, and tissue map registration parameters is maximized. ICM is used to maximize this conditional posterior. At each iteration the MOG parameters are estimated using EM whilst keeping the bias field and registration parameters fixed. Bias field and registration parameters are optimized using the Levenberg-Marquardt (LM) algorithm. The framework allows for spatial variation of the mixing parameters, although it does assume independence between voxel intensities. In practice, grey matter, white matter, CSF tissue probability maps as well as a probability map of not belonging to one of the three mentioned tissue types are used. Theoretically the framework generalizes to the inclusion of any number of structural probability maps, however the boundaries for subcortical structures may contain poor contrast and prove difficult to segment. Voxel-based MOG models have difficulties in differentiating tissues when there is no or little intensity contrast. One way that this can be overcome is by using a shape and intensity model that incorporates the learnt statistics into the segmentation such that areas of high intensity precision drive the

fitting and influence the low contrast areas via the shape model. This is the method that we propose in the next section.

3.2 Bayesian Shape and Appearance Model

As discussed briefly in the introduction to this chapter, conditional distributions provide a probabilistic means for modelling the relationship between two quantities, where the quantities are random variables such as shape and intensity. Given assumptions about the particular distributions we may derive an analytic expression for this relationship, the parameters of which may be estimated from training data. The Bayesian appearance model proposed in this section is based on a PDM of shape and intensity that is assumed to be multivariate Gaussian.

A Bayesian framework is proposed for modelling shape and appearance while explicitly accounting for the limited amount of training data. The framework facilitates the calculation of conditional distributions from the data (through the addition of a prior) which is otherwise problematic due to rank deficient covariance estimates. The appearance model is framed as the conditional distribution of intensity for a given shape, this eliminates the need for empirical weighting parameters describing the relationship between intensity and shape variance. As with the ASM we parameterize shape in terms of the mean and eigenvectors. The model is fit to new data by maximizing the posterior probability of shape given intensity; we are searching for a new set of model parameters (eigenvector weightings) given the observed intensities. The conditional distribution may also be used in terms of predicting one shape

distribution given the location of another shape.

We are now going to present in detail the mathematics of our proposed model. For clarity, $N \times 1$ matrices will be denoted using lowercase boldface (consistent with the vector notation from the previous chapter), matrices with both dimensions greater than one will be in uppercase boldface with the exception of the precision matrix given by $\boldsymbol{\lambda}$.

Although the framework generalizes to any measure that fits the distribution assumptions, for our purposes we are modelling mesh vertex coordinates and sampled intensities. The training data used are the mesh parameterizations derived from the application of the method described from chapter 2 to the volumetric training data. We will attempt to insert concrete examples periodically throughout the chapter for clarity.

3.2.1 Mathematical Model

Our model is trained from a finite set of mesh vertices and intensity samples that correspond to the finite set of volumetric training data. Given that we have a finite set of training data $\mathcal{Z} = \{\boldsymbol{x}_1 \dots \boldsymbol{x}_{n_s}\}$, our multivariate Gaussian model of the underlying distribution is given by

$$p(\boldsymbol{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\lambda}) = N_k(\boldsymbol{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\lambda}) = |\boldsymbol{\lambda}|^{\frac{1}{2}} (2\pi)^{\frac{-k}{2}} \exp\left(-\frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu_i})^t \boldsymbol{\lambda} (\boldsymbol{x}_i - \boldsymbol{\mu_i})\right), \quad (3.2)$$

where k is the dimensionality of $\boldsymbol{x}_i$, $\boldsymbol{\mu}$ is the mean and $\boldsymbol{\lambda}$ is a $k \times k$ positive-definite precision matrix. The precision matrix is equal to the inverse of the covariance matrix

$\boldsymbol{\Sigma}$. N_k is the k-dimensional multivariate Gaussian (Normal) distribution.

In our case, a training sample vector $\boldsymbol{x}_i$ for shape will be a column vector of the vertex coordinates. For example, for a 2D rectangle parameterized by the corner vertices $\mathcal{V} = \{(-2, 0), (-2, 3), (3, 3), (3, 0)\}$ the training vector would be $\boldsymbol{x}_i = [-2\ 0\ \ -2\ 3\ 3\ 3\ 3\ 0]^t$. The order of the vertices must be consistent across the training data. For the intensity distribution the training vector would merely be the intensity samples corresponding to the given rectangle. When discussing the joint distribution of shape and intensity the training vector would be the concatenation of the two training vectors.

Using Bayes' theorem, the distribution of new observed data given the training data is given by

$$p(\boldsymbol{x}_{obs} \mid \mathcal{Z}) = \int p(\boldsymbol{x}_{obs} \mid \boldsymbol{\mu}, \boldsymbol{\lambda}) p(\boldsymbol{\mu}, \boldsymbol{\lambda} \mid \mathcal{Z}) d\boldsymbol{\mu} d\boldsymbol{\lambda}, \tag{3.3}$$

where $\boldsymbol{x}_{obs}$ is a new observation sampled from the same distribution as that of $\mathcal{Z}$.

Given the sufficient statistics $t(\mathcal{Z})$ [5], it can be shown that

$$\begin{aligned} p(\boldsymbol{x}_{obs} \mid \mathcal{Z}) &= p(\boldsymbol{x}_{obs} \mid t(\mathcal{Z})) \\ &= \int p(\boldsymbol{x}_{obs} \mid \boldsymbol{\mu}, \boldsymbol{\lambda}) p(\boldsymbol{\mu}, \boldsymbol{\lambda}, \mid t(\mathcal{Z})) d\boldsymbol{\mu} d\boldsymbol{\lambda}. \end{aligned} \tag{3.4}$$

We use the sufficient statistics for the multivariate Gaussian given by

$$t(\mathcal{Z}) = (n_s, \bar{\boldsymbol{x}}, \boldsymbol{S}), \tag{3.5}$$

where

$$\bar{\boldsymbol{x}} = n_s^{-1} \sum_{i=1}^{n_s} \boldsymbol{x}_i, \tag{3.6}$$

$$\boldsymbol{S} = \sum_{i=1}^{n_s} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^t (\boldsymbol{x}_i - \bar{\boldsymbol{x}}). \tag{3.7}$$

The expression for $p(\boldsymbol{x}_{obs} \mid \boldsymbol{\mu}, \boldsymbol{\lambda})$ is given by the predictive model in (3.2). In order to evaluate (3.4) we now need to derive an expression for $p(\boldsymbol{\mu}, \boldsymbol{\lambda}, \mid t(\mathcal{Z}))$ which is the joint distribution for the true mean and variance given the sufficient statistics.

Using Bayes' theorem,

$$p\left(\boldsymbol{\mu}, \boldsymbol{\lambda} \mid t(\mathcal{Z})\right) = \frac{p(t(\mathcal{Z}) \mid \boldsymbol{\lambda}, \boldsymbol{\mu}) p(\boldsymbol{\lambda}, \boldsymbol{\mu})}{\int p(t(\mathcal{Z}) \mid \boldsymbol{\lambda}, \boldsymbol{\mu}) p(\boldsymbol{\lambda}, \boldsymbol{\mu}) d\boldsymbol{\mu} d\boldsymbol{\lambda}}, \tag{3.8}$$

where

$$p(t(\mathcal{Z}) \mid \boldsymbol{\mu}, \boldsymbol{\lambda}) = p(\boldsymbol{S} \mid \bar{\boldsymbol{x}}, \boldsymbol{\mu}, \boldsymbol{\lambda}) p(\bar{\boldsymbol{x}} \mid \boldsymbol{\mu}, \boldsymbol{\lambda}), \tag{3.9}$$

The sampling distributions $p(\boldsymbol{S} \mid \bar{\boldsymbol{x}}, \boldsymbol{\mu}, \boldsymbol{\lambda})$, $p(\bar{\boldsymbol{x}} \mid \boldsymbol{\mu}, \boldsymbol{\lambda})$ are given by

$$p(\bar{\boldsymbol{x}} \mid n_s, \boldsymbol{\mu}, \boldsymbol{\lambda}) = N_k(\bar{\boldsymbol{x}} \mid \boldsymbol{\mu}, n_s \boldsymbol{\lambda}), \tag{3.10}$$

$$p(\boldsymbol{S} \mid \bar{\boldsymbol{x}}, n_s, \boldsymbol{\mu}, \boldsymbol{\lambda}) = \mathrm{Wi_k}\left(\boldsymbol{S} \mid \frac{1}{2}(n_s - 1), \frac{1}{2}\boldsymbol{\lambda}\right). \tag{3.11}$$

$\mathrm{Wi_k}$ is a Wishart distribution with $\frac{1}{2}(n_s - 1)$ degrees of freedom, and a precision matrix of $\frac{1}{2}\boldsymbol{\lambda}$. For this case, to satisfy the requirements of the Wishart distribution, n_s must be greater than k. Substituting (3.10) and (3.11) back into (3.9), we arrive

at

$$p(t(\mathcal{Z}) \mid \boldsymbol{\mu}, \boldsymbol{\lambda}) = N_k(\bar{\boldsymbol{x}} \mid \boldsymbol{\mu}, n_s\boldsymbol{\lambda}) \, \mathrm{Wi_k}\left(\boldsymbol{S} \mid \frac{1}{2}(n_s - 1), \frac{1}{2}\boldsymbol{\lambda}\right). \tag{3.12}$$

To calculate the posterior $p(\boldsymbol{x}_{obs} \mid t(\mathcal{Z}))$, we need to specify the prior $p(\boldsymbol{\mu}, \boldsymbol{\lambda})$. Using the conjugate prior, and introducing the hyperparameters n_0, $\boldsymbol{\mu}_0$, and $\boldsymbol{\beta}$, the prior is given by

$$p(\boldsymbol{\mu}, \boldsymbol{\lambda} \mid \boldsymbol{\mu}_0, n_0, \boldsymbol{\beta}) = N_k(\boldsymbol{\mu} \mid \boldsymbol{\mu}_0, n_0\boldsymbol{\lambda}) \, \mathrm{Wi_k}(\boldsymbol{\lambda} \mid \alpha, \boldsymbol{\beta}). \tag{3.13}$$

By substituting (3.12) and (3.13) into (3.8), followed by (3.8) and (3.2) into (3.4), and then integrating, we obtain (as given in [5]).

$$p(\boldsymbol{x}_{obs} \mid \mathcal{Z}, n_0, \boldsymbol{\mu}_0, \boldsymbol{\beta}, \alpha) = St_k(\boldsymbol{x}_{obs} \mid \boldsymbol{\mu}_n, (n_0 + n_s + 1)^{-1}(n_0 + n_s)\alpha_n\boldsymbol{\beta}_n^{-1}, 2\alpha_n), \tag{3.14}$$

where

$$\begin{aligned}
\boldsymbol{\mu}_n &= (n_0 + n_s)^{-1}(n_0\boldsymbol{\mu}_0 + n_s\bar{\boldsymbol{x}}), \\
\boldsymbol{\beta}_n &= \boldsymbol{\beta} + \frac{1}{2}\boldsymbol{S} + (n_s + n_0)^{-1}n_s n_0(\boldsymbol{\mu}_0 - \bar{\boldsymbol{x}})(\boldsymbol{\mu}_0 - \bar{\boldsymbol{x}})^t, \\
\alpha_n &= \alpha + \frac{1}{2}n_s - \frac{1}{2}(k - 1),
\end{aligned} \tag{3.15}$$

and St_k is a multivariate Student distribution. The full expression for a multivariate Student distribution is given by

$$St_k(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\lambda}, \alpha) = c\left[1 + \frac{1}{\alpha}(\boldsymbol{x} - \boldsymbol{\mu})^t\boldsymbol{\lambda}(\boldsymbol{x} - \boldsymbol{\mu})\right]^{-\frac{\alpha+k}{2}}, \tag{3.16}$$

where $c = \frac{\Gamma(\frac{1}{2}(\alpha+k))}{\Gamma(\frac{1}{2}\alpha)(\alpha\pi)^{\frac{k}{2}}}$, and the variance is given by $\boldsymbol{V}[x] = \boldsymbol{\lambda}^{-1}\frac{\alpha}{\alpha-2}$.

3.2.2 Choice of priors

The first prior chosen is $n_0 = 0$ as this is a flat, non-informative prior on $p(\boldsymbol{\mu})$ and results in $p(\boldsymbol{x}_{obs} \mid \mathcal{Z}, n_0, \boldsymbol{\mu}_0, \boldsymbol{\beta}, \alpha)$ being centered at $\bar{\boldsymbol{x}}$ with no dependence on $\boldsymbol{\mu}_0$. By substituting $n_0 = 0$ back into (3.14) we obtain

$$p(\boldsymbol{x}_{obs} \mid \mathcal{Z}, n_0 = 0, \boldsymbol{\beta}, \alpha) = St_k(\boldsymbol{x}_{obs} \mid \boldsymbol{\mu}_n, \frac{n_s}{n_s + 1}\alpha_n \boldsymbol{\beta}_n^{-1}, 2\alpha_n), \tag{3.17}$$

where

$$\begin{aligned} \boldsymbol{\mu}_n &= \bar{\boldsymbol{x}}, \\ \boldsymbol{\beta}_n &= \boldsymbol{\beta} + \frac{1}{2}\boldsymbol{S}, \\ \alpha_n &= \alpha + \frac{1}{2}n_s - \frac{1}{2}(k-1). \end{aligned} \tag{3.18}$$

Now, expanding (3.17) into the general form for the multivariate Student distribution, and rearranging we obtain

$$p(\boldsymbol{x}_{obs} \mid \mathcal{Z}, n_0 = 0, \boldsymbol{\beta}, \alpha) = c\left[1 + \frac{1}{n_s - \frac{1}{n_s}}(\boldsymbol{x} - \bar{\boldsymbol{x}})^t \left(\frac{\boldsymbol{S} + 2\boldsymbol{\beta}}{n_s - 1}\right)^{-1} (\boldsymbol{x} - \bar{\boldsymbol{x}})\right]^{\frac{-(k+n_s-\frac{1}{n_s})}{2}}. \tag{3.19}$$

In the above, the prior α has been chosen such that the sample covariance is normalized by $n_s - 1$ (which corresponds to the standard unbiased estimate of a covariance matrix). By setting $\boldsymbol{\lambda} = \left(\frac{\boldsymbol{S}+2\boldsymbol{\beta}}{n_s-1}\right)^{-1}$, in order for $p(\boldsymbol{x}_{obs} \mid \mathcal{Z}, n_0 = 0, \boldsymbol{\beta}, \alpha)$ (equation (3.19)) to be consistent with the multivariate Student distribution (equation (3.16))

it follows that

$$2\alpha_n = n_s - \frac{1}{n_s}, \tag{3.20}$$

$$\alpha = \frac{1}{2}\left(k + 1 - \frac{1}{n_s}\right). \tag{3.21}$$

This meets the minimum criteria, $2\alpha > k - 1$, for degrees of freedom (as given by (3.13)). For rotational invariance, $\boldsymbol{\beta}$ is chosen to be a scaled identity matrix $\epsilon^2\boldsymbol{I}$. ϵ^2 may be interpreted as an error variance and can be estimated from the training data (chosen to be a percentage of the total estimated variance). The addition of the scaled identity matrix is analogous to ridge regression where a scaled identity matrix is added to the covariance matrix estimate.

This particular prior broadens the distribution, reflecting the fact that we believe there are types of variation in the larger population that are not observed in the training data.

Our model takes the final form

$$p(\boldsymbol{x}_{obs} \mid \mathcal{Z}, \epsilon) = St_k\left(\boldsymbol{x}_{obs} \mid \bar{\boldsymbol{x}}, \frac{\boldsymbol{S} + 2\epsilon^2\boldsymbol{I}}{n_s - 1}, n_s - \frac{1}{n_s}\right), \tag{3.22}$$

with the variance given by

$$\boldsymbol{V}[\boldsymbol{x}_{obs}] = \left(\frac{\boldsymbol{S} + 2\epsilon^2\boldsymbol{I}}{n_s - 1}\right)\gamma_v, \tag{3.23}$$

where we have defined $\gamma_v = \frac{n_s - \frac{1}{n_s}}{n_s - \frac{1}{n_s} - 2}$.

3.2.3 Conditional distributions

We are interested in conditional distributions across partitions of the joint multivariate Gaussian model. A partition is a subset, $\boldsymbol{x}_j$, of $\boldsymbol{x}$ corresponding to a particular attribute j (e.g. shape, intensity, etc...). In the case of training data, each partition will still have the same number of samples n_s. In our application, we partition the data into either shape and intensity or into different shapes. The shape/intensity partitions are used to estimate the distribution of intensity given a particular shape whereas the shape/shape partitions are used to estimate the distribution of a shape given another. The conditional distribution is essential to model the relationship between shape and intensity.

If $\boldsymbol{x}$ can be partitioned such that,

$$\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2), \tag{3.24a}$$

$$\boldsymbol{\lambda} = \begin{bmatrix} \boldsymbol{\lambda}_{11} & \boldsymbol{\lambda}_{12} \\ \boldsymbol{\lambda}_{21} & \boldsymbol{\lambda}_{22} \end{bmatrix}, \tag{3.24b}$$

$$k = k_1 + k_2, \tag{3.24c}$$

where k_j is the dimensionality of the jth partition, then $\mathcal{Z}$ can be partitioned in the same manner, such that $\mathcal{Z} = (\mathcal{Z}_1, \mathcal{Z}_2)$, where $\mathcal{Z}_j = \{\tilde{\boldsymbol{x}}_{ij} \dots \tilde{\boldsymbol{x}}_{n_s j}\}$. For the joint shape and intensity distribution, the shape partition $\mathcal{Z}_1$ is the unwrapped vertex coordinates and the intensity partition $\mathcal{Z}_2$ is the corresponding intensity samples. For evaluating a shape distribution given another shape (e.g. the distribution of the caudate given the known location of the thalamus), the first partition is the unwrapped vertex

coordinates of the structure of interest and the second partition is the unwrapped vertex coordinates of the predictive structure.

It follows that

$$p(\boldsymbol{x}_1 \mid \boldsymbol{x}_2, \mathcal{Z}) = \frac{p(\boldsymbol{x}_1, \boldsymbol{x}_2 \mid \mathcal{Z})}{p(\boldsymbol{x}_2 \mid \mathcal{Z})} = \frac{St_{k_1+k_2}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\lambda}, \alpha)}{St_{k_2}(\boldsymbol{x}_2 \mid \boldsymbol{\mu}_2, \boldsymbol{\lambda}_2, \alpha_2)}, \tag{3.25a}$$

which simplifies to

$$p(\boldsymbol{x}_1 \mid \boldsymbol{x}_2, \mathcal{Z}) = St_{k_1}(\boldsymbol{x}_1 \mid \boldsymbol{\mu}_{1|2}, \boldsymbol{\lambda}_{1|2}, \alpha_{1|2}), \tag{3.25b}$$

where

$$\begin{aligned} \boldsymbol{\mu}_{1|2} &= \boldsymbol{\mu}_1 - \boldsymbol{\lambda}_{11}^{-1}\boldsymbol{\lambda}_{12}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2) \\ &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \end{aligned} \tag{3.25c}$$

$$\begin{aligned} \boldsymbol{\lambda}_{1|2} &= \boldsymbol{\lambda}_{11}[\frac{\alpha_{1,2} + k_2}{\alpha_{1,2} + (\boldsymbol{x}_2 - \bar{\boldsymbol{x}}_2)^T\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \bar{\boldsymbol{x}}_2}), \\ \boldsymbol{\Sigma}_{1|2} &= \left(\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right)[\frac{\alpha_{1,2} + (\boldsymbol{x}_2 - \bar{\boldsymbol{x}}_2)^T\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \bar{\boldsymbol{x}}_2)}{\alpha_{1,2} + k_2}], \end{aligned} \tag{3.25d}$$

$$\alpha_{1|2} = \alpha_{1,2} + k_2. \tag{3.25e}$$

This is a standard manipulation for the multivariate Student distribution given partitioned matrices [5].

For a partitioned covariance matrix the prior $\boldsymbol{\beta}$ will be defined as a piecewise-scaled identity matrix such that

$$\boldsymbol{\beta} = \begin{bmatrix} \epsilon_1^2\boldsymbol{I_1} & 0 \\ 0 & \epsilon_2^2\boldsymbol{I_2} \end{bmatrix}, \tag{3.26}$$

where ϵ_i^2 is the error variance corresponding to the i^{th} partition. Thus for each partition a different error variance, ϵ_i^2 may be used.

3.2.4 Parameterization of Bayesian Models from Finite Training Data

Our shape model uses the mean and eigenvectors to parameterize shape so as to constrain the search space. Furthermore by expressing the training data in terms of its eigenvectors and singular values, we are able to simplify the evaluation of the model in terms of computational cost as well as to provide a better insight into the mechanisms underlying the probabilistic model.

Defining the matrix $\boldsymbol{Z}$ as the demeaned training set $\tilde{\mathcal{Z}}$, we express $\boldsymbol{Z}$ in terms of its SVD,

$$\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T, \tag{3.27}$$

where $\boldsymbol{U}$ are the eigenvectors of the covariance matrix, $\boldsymbol{D}$ are the singular values, and $\boldsymbol{V}$ is the parameter matrix needed to reconstruct the original data.

The un-normalized sample covariance matrix, S, maybe be expressed in terms of the demeaned training data by

$$S = \boldsymbol{Z}\boldsymbol{Z}^T = \boldsymbol{U}\boldsymbol{D}^2\boldsymbol{U}^T. \tag{3.28}$$

Calculating the inverse of the full covariance matrix is computationally expensive, but the low rank of the estimated covariance matrix may be exploited for computational

simplifications. As mentioned in the previous section, a scaled identity matrix ($\epsilon^2\boldsymbol{I}$) is added to the estimated covariance matrix, $\boldsymbol{S}$ (equation (3.7)), and results in a well-conditioned matrix. Despite the fact that adding the scaled-identity matrix results in a full-rank covariance matrix, simplifications may still be made by using the fact that adding a scaled-identity matrix to a covariance matrix is equivalent to adding the scale factor to each eigenvalue $\boldsymbol{D}_i^2$ (including to the zero eigenvalues which correspond to the null space) [45].

As stated earlier the variance of the multivariate Student distribution is given by $\boldsymbol{V}[\boldsymbol{x}] = \left(\frac{\boldsymbol{S}+2\epsilon^2\boldsymbol{I}}{n_s-1}\right)\gamma_v$. For convenience we now define

$$\boldsymbol{\Sigma}\gamma_v = \boldsymbol{V}[\boldsymbol{x}] = \left(\frac{\boldsymbol{S}+2\epsilon^2\boldsymbol{I}}{n_s-1}\right)\gamma_v. \tag{3.29}$$

For computational purposes we are able to express the eigen-decomposition of equation $\boldsymbol{\Sigma}\gamma_v$ in terms of the eigenvectors and eigenvalues of the matrix S (which is typically sparse) and the scaled identity matrix $\epsilon^2\boldsymbol{I}$. We substitute equation (3.28) into equation (3.29) and simplify such that

$$\begin{aligned}\boldsymbol{\Sigma}\gamma_v &= (\boldsymbol{S}+2\epsilon^2\boldsymbol{I})(n_s-1)^{-1}\gamma_v \\ &= \boldsymbol{U}(\boldsymbol{D}^2+2\epsilon^2\boldsymbol{I})\boldsymbol{U}^T(n_s-1)^{-1}\gamma_v \\ &= \boldsymbol{U}\boldsymbol{D}_\epsilon^2\boldsymbol{U}^T(n_s-1)^{-1}\gamma_v,\end{aligned} \tag{3.30}$$

where $\boldsymbol{D}_\epsilon^2$ is a diagonal matrix consisting of the eigenvalues of $2\epsilon^2\boldsymbol{I}+\boldsymbol{S}$.

Performing an SVD on the $k_j \times n_s$ data matrix provides the first n_s eigenvectors without requiring an eigenvalue decomposition of the full $k_j \times k_j$ covariance matrix.

This has a large computational saving when n_s is much less than k_j. For example, we use $n_s = 317$ (a fairly large training set in medical imaging) and the shape and intensity dimensionality (k_j) for the left putamen is 1926 and 8346 respectively.

As with ASMs, we can now parameterize our data in terms of the mean and eigenvectors, as given by

$$\hat{\boldsymbol{x}} = \bar{\boldsymbol{x}} + \boldsymbol{U}\frac{\boldsymbol{D}_\epsilon\sqrt{\gamma_v}}{\sqrt{(n_s-1)}}\boldsymbol{b}, \tag{3.31}$$

where $\boldsymbol{b}$ is the model parameter vector that weights the linear combination of eigenvectors used to create new shape instances. The magnitude of elements of $\boldsymbol{b}$ indicate the number of standard deviations along each mode.

3.2.5 Bayesian Appearance Models

Our mathematical framework is now applied to appearance models. The joint distribution of shape and intensity is being modelled as a multivariate Gaussian distribution. From our training set, using the model given by (3.22), we learn the joint intensity/shape distribution $p(\boldsymbol{x}_I, \boldsymbol{x}_s)$. $\boldsymbol{x}_s$ is a column vector containing the x, y and z coordinates of all the vertices. $\boldsymbol{x}_I$ is a column vector containing all the intensity samples. Given that $p(\boldsymbol{x}_I, \boldsymbol{x}_s \mid \mathcal{Z})$ is partitionable we can calculate the conditional intensity distribution, $p(\boldsymbol{x}_I \mid \boldsymbol{x}_s, \mathcal{Z})$, given a particular shape and a finite training set. $p(\boldsymbol{x}_I \mid \boldsymbol{x}_s, \mathcal{Z})$ takes the form of equation (3.25) with $\boldsymbol{x}_I$ and $\boldsymbol{x}_s$ corresponding to partitions $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ respectively. The shape partition is modelled using equation (3.31), so for any $\boldsymbol{b}_s$ vector (new shape instance) we can predict the intensity distribution. The conditional distribution between shape and intensity captures the intensity shifts

in conditional mean and covariance that correlate with shape. An example of such a relationship would be the case of hippocampal atrophy when the contraction of the hippocampal surface correlates to the observation of CSF instead of grey or white matter at the border facing the thalamus.

Returning to our 2D rectangle parameterized by the corner vertices $\mathcal{V} = \{(-2,0),(-2,3),(3,3),(3,0)\}$, with a shape training vector would be $\boldsymbol{x}_s = [-2\ 0\ -2\ 3\ 3\ 3\ 3\ 0]^t$, we now wish to model the intensity variation at the vertices. We observe the set of intensities

$\mathcal{I} = \{(1,2,0,-2,-3),(2,1,1,-1,-2),(2,-1,2,-2,3),(0,1,-1,-3,-2)\}$,

such that five intensity samples were taken for each vertex; one at the vertex, two in the x-direction, and two in the y-direction. The intensity training vector then becomes

$\boldsymbol{x}_I = [1\ 2\ 0\ -2\ -3\ 2\ 1\ 1\ -1\ -2\ 2\ -1\ 2\ -2\ 3\ 0\ 1\ -1\ -3\ -2]^t$.

The combined shape-intensity training vector for the joint shape and intensity distribution would be

$\boldsymbol{x}_{s,i} = [\boldsymbol{x}_s\ \boldsymbol{x}_i] =$

$[-2\ 0\ -2\ 3\ 3\ 3\ 3\ 0\ 1\ 2\ 0\ -2\ -3\ 2\ 1\ 1\ -1\ -2\ 2\ -1\ 2\ -2\ 3\ 0\ 1\ -1\ -3\ -2]^t$.

The order of the intensity samples must be consistent across the training data. It is clear from this simple example how quickly the dimensionality increases with the incorporation of intensity samples.

3.3 Model Fitting and Evaluation

In the previous section we proposed a probabilistic model for shape and intensity, in this section we describe the means by which the model is fitted to new data. Because the model was framed probabilistically we may now use Bayes' rule to maximize the posterior probability of shape given the observed intensities. As discussed earlier, when fitting to new data, the model was first registered into the native space using the inverse transform from the two-stage linear subcortical registration. To register the model, the linear transformation matrix need only be applied to the average shape and eigenvectors (see appendix A). One aspect of this shape model that is somewhat unique is that not all pose is removed from the structure and that the remaining pose information is modelled through the eigenvectors. In practice the pose differences are small because of the initial linear registration.

3.3.1 Posterior as a Cost Function

To fit our model to new data we are searching for a new set of model parameters given the observed intensities. Instead of using $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ as was used in the model formulation we now use $\boldsymbol{x}_I$ and $\boldsymbol{x}_s$ respectively to indicate the nature of the partition (i.e. intensity and shape). Hence, when fitting the Bayesian appearance model $p(\boldsymbol{x}_I \mid \boldsymbol{x}_s)$, we aim to maximize $p(\boldsymbol{x}_s \mid \boldsymbol{x}_I)$, as given by

$$p(\boldsymbol{x}_s \mid \boldsymbol{x}_I) \propto p(\boldsymbol{x}_I \mid \boldsymbol{x}_s)p(\boldsymbol{x}_s). \tag{3.32}$$

In our application we limit the search space to the span of the eigenvectors and hence the gradients are taken along each mode of variation. We are effectively maximising equation (3.32) with respect to the shape model parameters $\boldsymbol{b}_s$. Now, simplifying the posterior and expressing it in terms of its logarithm we obtain (see appendix B)

$$\begin{aligned} -\ln p(\boldsymbol{x}_s \mid \boldsymbol{x}_I) \propto C &+ \frac{k_I}{2} \ln \left(\frac{\alpha_{I,s} + k_s}{\alpha_{I,s} + b_s^T b_s \gamma_v} \right) \\ &- \frac{(\alpha_{I,s} + k_s + k_I)}{2} \ln \left(1 + \frac{1}{\alpha_{I,s} + b_s^T b_s \gamma_v} (\boldsymbol{x}_I - \boldsymbol{\mu}_{I\mid s})^t \boldsymbol{\lambda}_{cII} (\boldsymbol{x}_I - \boldsymbol{\mu}_{I\mid s}) \right) \\ &+ \frac{(\alpha_s + k_s)}{2} \ln \left(1 + \frac{1}{\alpha_s} b_s^T b_s \gamma_v \right), \end{aligned} \tag{3.33}$$

where $\alpha_{I,s}$ is the degrees of freedom of $p(\boldsymbol{x}_I \mid \boldsymbol{x}_s)$ (defined by equation (3.21)), $\boldsymbol{\lambda}_{cII}$ is the un-scaled conditional precision matrix, k_I and k_s are respectively the dimensionality of intensity and shape partitions, $\boldsymbol{x}_I$ is the observed intensities and $\boldsymbol{\mu}_{I\mid s}$ is the conditional mean given shape as expressed in (3.43); the conditional mean is a function of $\boldsymbol{b}_s$ (see section 3.3.4 and appendix C). The simplified expression for the conditional mean is given by

$$\boldsymbol{\mu}_{I\mid s} = \boldsymbol{\mu}_I + \boldsymbol{Z_I} \left[\boldsymbol{V}_s \boldsymbol{D}_s \boldsymbol{D}_{\epsilon_s}^{-1} \sqrt{\frac{\gamma_v}{n_s - 1}} \right] \boldsymbol{b_s}, \tag{3.34}$$

where $\boldsymbol{Z}_I$ and $\boldsymbol{Z}_s$ are the demeaned intensity samples and vertex coordinates respectively. $\boldsymbol{Z}_I$ and $\boldsymbol{Z}_s$ are defined in terms of their SVD expansion as $\boldsymbol{Z}_I = \boldsymbol{U}_I \boldsymbol{D}_I \boldsymbol{V}_I^T$ and $\boldsymbol{Z}_s = \boldsymbol{U}_s \boldsymbol{D}_s \boldsymbol{V}_s^T$. $\boldsymbol{D}_{\epsilon_s}$ is defined as $\boldsymbol{D}_{\epsilon_s} = \boldsymbol{D}_s + \epsilon_s^2 \boldsymbol{I}$, where $\epsilon_s^2 \boldsymbol{I}$ is the amount of variance being added to each eigenvalue of the estimated covariance matrix (the prior

mentioned previously). The simplified conditional covariance matrix is given by

$$\boldsymbol{\lambda}_{cII} = [\boldsymbol{U}_{I|s}\boldsymbol{D}_{cIIV}^{-2}\boldsymbol{U}_{I|s}^{T} + \frac{1}{2}\epsilon_{I}^{-2}\boldsymbol{I}](n_s - 1), \tag{3.35}$$

where $\boldsymbol{U}_{I|s} = \boldsymbol{U}_{I}\boldsymbol{U}_{cIIV}$ such that $\boldsymbol{U}_{cIIV}$ and $\boldsymbol{D}_{cIIV}^{2}$ are the eigenvectors and eigenvalues of

$$\boldsymbol{\Sigma}_{cIIV} = \boldsymbol{D}_{\epsilon_I}^{2} - \boldsymbol{D}_{I}\boldsymbol{V}_{I}^{T}\boldsymbol{V}_{s}\boldsymbol{D}_{s}^{T}\boldsymbol{D}_{\epsilon_s}^{-2}\boldsymbol{D}_{s}\boldsymbol{V}_{s}^{T}\boldsymbol{V}_{I}\boldsymbol{D}_{I}^{T}. \tag{3.36}$$

The expression for these are discussed in section 3.3.4 and derived in appendix C. All the matrices in equation 3.36 are of size $n_s \times n_s$ (n_s is the number of subjects), the subscript ",s" that is used in appendix C was dropped to avoid confusion with the subscript s that is used here to indicate shape. Furthermore,

$$\alpha_{I,s} = \frac{1}{2}\left(k_{I,s} + 1 - \frac{1}{n_s}\right), \tag{3.37}$$

where $k_{I,s}$ is the dimensionality of the joint distribution of intensity and shape, so that

$$k_{I,s} = (3 + n_I)\, n_{vert}, \tag{3.38}$$

where n_{vert} is the number of vertices and n_I is the number of intensity samples taken per vertex. Finally γ_v is given by

$$\gamma_v = \frac{n_s - \frac{1}{n_s}}{n_s - \frac{1}{n_s} - 2}. \tag{3.39}$$

3.3.2 Conditional Shape Priors

In practice, given the amount of training data, limiting the search space to the joint modes results in shape constraints that are too strict and is overly ambitious when generalizing to the unsampled population. By joint modes we are referring to the eigenvectors of the covariance matrix of the concatenated shapes (i.e. the concatenated vertex coordinates of multiple shapes). Instead, structural covariation is incorporated as a prior in our model as given by

$$p(\boldsymbol{x}_{s1} \mid \boldsymbol{x}_{I1}, \boldsymbol{x}_{s2}) = \frac{p(\boldsymbol{x}_{I1}, \boldsymbol{x}_{s2} \mid \boldsymbol{x}_{s1})p(\boldsymbol{x}_{s1})}{p(\boldsymbol{x}_{I}, \boldsymbol{x}_{s2})}, \tag{3.40}$$

where $\boldsymbol{x}_{s1}$ is the shape vector of the structure being fitted, $\boldsymbol{x}_{s2}$ is the fixed shape vector constraining the search space of $\boldsymbol{x}_{s1}$, and $\boldsymbol{x}_{I1}$ is the vector of intensity samples associated with $\boldsymbol{x}_{s1}$.

Using our proposed framework, $p(\boldsymbol{x}_{I1}, \boldsymbol{x}_{s2} \mid \boldsymbol{x}_{s1})$ can be learnt from the data, where $\boldsymbol{x}_{I1}$, $\boldsymbol{x}_{s2}$ are combined into a single partition and $\boldsymbol{x}_{s1}$ into another.

If we make the basic assumption of independence between $\boldsymbol{x}_{I1}$ and $\boldsymbol{x}_{s2}$, (3.40) simplifies to

$$\begin{aligned} p(\boldsymbol{x}_{s1} \mid \boldsymbol{x}_{I1}, \boldsymbol{x}_{s2}) &= \frac{p(\boldsymbol{x}_{I1} \mid \boldsymbol{x}_{s1})p(\boldsymbol{x}_{s2} \mid \boldsymbol{x}_{s1})p(\boldsymbol{x}_{s1})}{p(\boldsymbol{x}_{I1})p(\boldsymbol{x}_{s2})} \\ &= \frac{p(\boldsymbol{x}_{I1} \mid \boldsymbol{x}_{s1})p(\boldsymbol{x}_{s1} \mid \boldsymbol{x}_{s2})}{p(\boldsymbol{x}_{I1})} \\ &\propto p(\boldsymbol{x}_{I1} \mid \boldsymbol{x}_{s1})p(\boldsymbol{x}_{s1} \mid \boldsymbol{x}_{s2}). \end{aligned} \tag{3.41}$$

By making the assumption of independence we are reducing the maximum distri-

bution dimensionality that we are trying to estimate. By making this assumption we are potentially throwing away information about the interaction between a given shape and the intensity profiles of another shape; this would be most pronounced for neighbouring structures. Even though we discard the cross-covariation between shape and neighbouring structures' intensity for computational simplicity (which is also excluded for single structure fits), it does have the added benefit of reducing the number of parameters in the covariance matrix to be estimated. A portion of the discarded information is in fact redundant, if the structures border each other the intensity samples for a single structure would overlap the neighbouring structure.

The negative-log posterior is now given by

$$-\ln p(\boldsymbol{x}_{s1} \mid \boldsymbol{x}_{I1}, \boldsymbol{x}_{s2}) \propto -\ln p(\boldsymbol{x}_{I1} \mid \boldsymbol{x}_{s1}) - \ln p(\boldsymbol{x}_{s1} \mid \boldsymbol{x}_{s2}). \tag{3.42}$$

This differs from (3.32) in that the shape prior $p(\boldsymbol{x}_{s1})$ is replaced by a conditional shape prior $p(\boldsymbol{x}_{s1} \mid \boldsymbol{x}_{s2})$. The evaluation of the conditional distribution of one shape given another can be simplified to a single $n_s \times n_s$ by $n_s \times 1$ matrix multiplication at the beginning of the search, and an $n_s \times n_s$ matrix times an $n_s \times 1$ for each new parameter estimate of $\boldsymbol{b}_1$ that is visited. See appendix D for details.

3.3.3 Optimization

By minimizing the cost function we are maximizing the probability of the shape given the observed image intensities. To minimize the cost function a conjugate gradient descent search scheme is employed. With each iteration a conjugate gradient

search scheme minimizes in a direction orthogonal (conjugate) to the previous. The conjugate gradient is a widely used optimization method that typically has better convergence properties than a steepest descent search. In our application, because we restrict ourselves to the shape space defined by the eigenvectors we take the gradient of the cost function with respect to the mode parameters (search along the eigenvectors). In practice we truncate the number of modes at a fixed number. Later in this chapter, when discussing validation we will investigate the effect of varying the number of modes of variation that are included. The eigenvectors are ordered by descending variance, hence the lower modes contribute less information to the model (or even just noise) than the upper modes. Including more modes of variation increases the number of parameters being estimated; the more parameters there are to optimize, the more difficult the optimization may become. Thus, if the information contributed by the increased number of modes of variation does not make up for the increased optimization difficulty, the inclusion of more modes may decrease performance. In practice, the optimization time increases with the number of modes, thus if the number of modes does not improve performance it is best not to include them. This is confirmed empirically by performing leave-one-out cross-validation over several different truncation levels for the number of modes of variation (see results in appendix E).

3.3.4 Computational Simplifications

Given that shape deformations are constrained to linear combinations of the modes of variation we can make some computational simplifications. Typically, we are dealing

with operations involving large covariance matrices (e.g. 1926×1926 for the left putamen) that are computationally expensive since obtaining the conditional covariance matrix is dependent on the inverse of the covariance matrix. However, we are able to eliminate all operations on $k \times k$ matrices, making the models computationally feasible in practice. The simplification exploits the low rank of the covariance matrices which may thus be expressed by far fewer eigenvalues than the dimensionality.

Conditional Mean as a Function of the Mode Parameters $\boldsymbol{b}_2$

Given training data with two partitions $\mathcal{Z}_1$ and $\mathcal{Z}_2$ ($\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ are the matrices of the demeaned partitions of the training data) and the latter being parameterized in (3.31), the conditional mean can be expressed as a function of predictor model parameter $\boldsymbol{b}_2$. This provides an efficient method for calculating the conditional mean at run time rather than operating on the full covariance matrices. The conditional mean expressed in terms of the shape parameter vector $\boldsymbol{b}_2$ is given by

$$\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \boldsymbol{Z}_1 \left[\boldsymbol{V}_2 \boldsymbol{D}_{2,s} \boldsymbol{D}_{\epsilon_2,s}^{-1} \sqrt{\frac{\gamma_v}{n_s - 1}} \right] \boldsymbol{b}_{2,s}. \tag{3.43}$$

In general the ",s" subscript refers to the upper-left submatrix, such that the maximum dimension is n_s. $\boldsymbol{b}_{2,s}$ is the first n_s rows of $\boldsymbol{b}_2$. All matrices within square brackets in (3.43) are of size $n_s \times n_s$ except $\boldsymbol{b}_{2,s}$ which is $n_s \times 1$. If we truncate the number of modes at L, only the first L columns of $\boldsymbol{Z}_1 \left[V_2 \boldsymbol{D}_{2,s} \boldsymbol{D}_{\epsilon_2,s}^{-1} \sqrt{\frac{\gamma_v}{n_s - 1}} \right]$ are needed. See appendix C.1 for details of the derivation.

Evaluating Conditional Covariance Operations

In order to simplify the calculation of the conditional probability we need to simplify operations involving a covariance matrix (3.25d). Conditional covariance may be used in two ways: 1) to calculate conditional modes of variation, e.g. to model the variation of the thalamus given that we know the location of the putamen; and 2) to explicitly calculate the probability of a predicted measure, e.g. to calculate the probability of intensities in the image given a known shape.

In case 1, we need to calculate the eigenvectors and eigenvalues for $\boldsymbol{\Sigma}_{1|2}$; although conditional modes of variation are not actually used in practice, the eigenvectors and eigenvalues are used in further simplifications. To calculate the eigenvectors directly from $\boldsymbol{\Sigma}_{1|2}$ can be a very expensive operation given that the number of control points in practice is large. In case 2, we need to evaluate $(\boldsymbol{x}_I - \boldsymbol{\mu}_{I|s})^T \boldsymbol{\Sigma}_{I|s}^{-1} (\boldsymbol{x}_I - \boldsymbol{\mu}_{I|s})$. For both cases we will exploit the fact that n_s is typically much less than k to simplify the calculations, though the results are valid for any $n_s < k$. These simplifications are left to appendix C.2.

3.3.5 Validation and Accuracy

The accuracy of the algorithm was evaluated using leave-one-out cross-validation across the 317 training sets. For all evaluations the manual segmentations are regarded as the gold standard. The segmentation performance was measured using the

Dice overlap metric given by

$$D = \frac{2TP}{2TP + FP + FN}, \tag{3.44}$$

where TP is the true positive volume, FP is the false positive volume, and FN is the false negative volume as given by

$$TP = \sum_{k=1}^{N_z} \sum_{j=1}^{N_y} \sum_{i=1}^{N_x} (G(i,j,k)S(i,j,k)) \tag{3.45a}$$

$$FP = \sum_{k=1}^{N_z} \sum_{j=1}^{N_y} \sum_{i=1}^{N_x} (S(i,j,k) - G(i,j,k)S(i,j,k)) \tag{3.45b}$$

$$FN = \sum_{k=1}^{N_z} \sum_{j=1}^{N_y} \sum_{i=1}^{N_x} (G(i,j,k) - G(i,j,k)S(i,j,k)), \tag{3.45c}$$

where G is the binarised gold standard image of the single structure and S is the binarised segmentation for a single structure and N_x, N_y, N_z are the number of voxels in the x, y and z directions respectively.

The volumetric output used to compute the Dice metric results from filling the output mesh. The mesh filling process consists of two steps: 1) drawing the mesh outline and 2) filling the interior. As a consequence of these two steps, we know whether an output voxel belongs to the boundary or the interior of the structure. In practice we then threshold the boundary voxels based on the intensity statistics of the interior voxels. The intensity distribution is assumed to be Gaussian with the mean and variance being estimated by the mode and full-width-half-maximum (FWHM) of the intensity histogram. A z-score is calculated at each boundary voxel and is

thresholded based on a set z-value. The z-score is equal to the number of standard deviations away from the population mean the sample lies, and is given by

$$z = \frac{x - \bar{x}}{\sqrt{s^2}}, \tag{3.46}$$

where x is the sample, $\bar{x}$ is the estimated mean value, and s^2 is the sample variance.

To investigate the effect of inaccuracies inherent in moving between mesh and volumetric representations, we introduce a boundary-corrected Dice (BCD) measurement. Assuming the boundary voxels to be unreliable, though correctable, then the BCD is the maximum achievable Dice overlap out of all boundary correcting schemes. The BCD is given by

$$BCD = \frac{2(TP_{int} + G_{bound})}{2(TP_{int} + G_{bound}) + FP_{int}}, \tag{3.47}$$

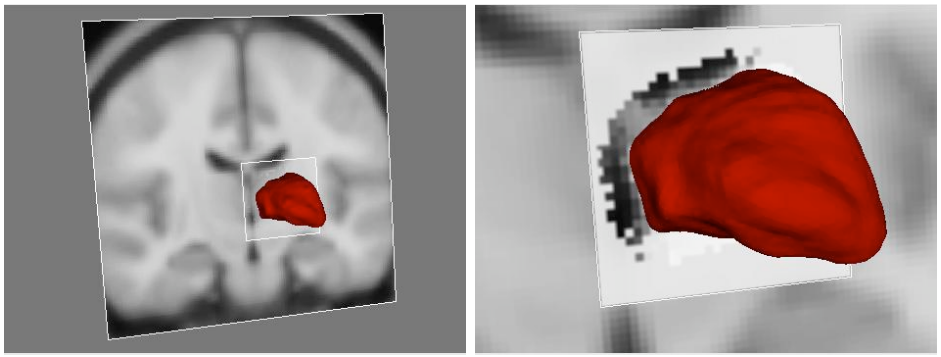
where TP_{int}, FP_{int} respectively are the true positive and false positive volumes that are contained within the interior of the filled mesh. G_{bound} is the ground truth volume contained within the boundary of the filled mesh. Because the boundary voxels were corrected based on the ground truth all the boundary voxels are true positives. The BCD is a method similar to that proposed in Crum et al. [15], where the condition for overlap at the boundary voxels is relaxed based on the assumption that the boundary is wrong.

3.4 Results and Discussion

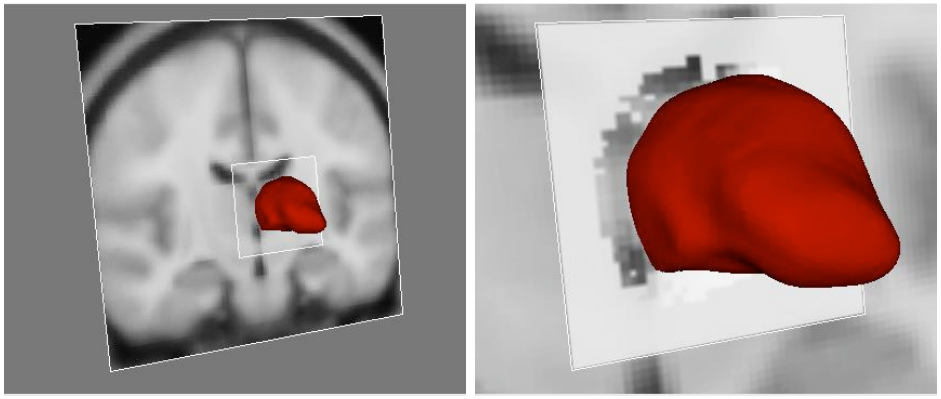
We will first qualitatively demonstrate a Bayesian appearance model, then follow with results from the leave-one-out experiments. By varying the shape parameter corresponding to the individual modes of variation, we can observe the surface deformations as well as the predicted intensity distribution. Figure 3.2 is a graphical depiction of ± 3 standard deviations along the first mode of variation for the left thalamus and the conditional intensity mean associated with it; the model is overlaid on the MNI152 template. For each vertex, 13 intensity samples were taken along the surface normal at a 0.5mm intervals. The first mode is predominantly one of translation; the translation typically correlates with an increased ventricle size as can be seen by the enlarging dark band in the conditional mean intensity.

The 17 subcortical structures that were modelled were all independently fit to each of the 317 training sets using 10, 20, 30, 40, 50, 60, and 70 eigenvectors. ϵ_s and ϵ_I were chosen to be 0.0001% of the total shape and intensity variance respectively. Appendix E includes the boxplots of BCD for each structure as a function of the number of eigenvectors used. Figure 3.4 shows the Dice measure for each structure corresponding to its optimal number of modes of variation (as determined from the LOO cross-validation results in appendix E). The Dice measures are given for the output volume with and without using boundary correction (z threshold of 3).

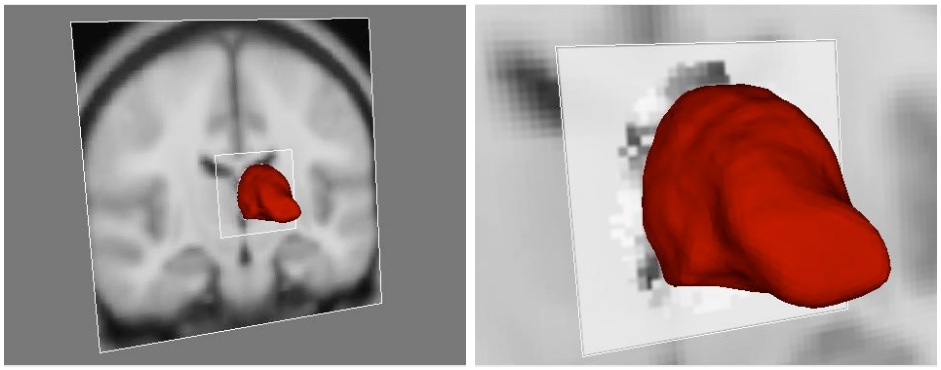
As seen in the results in appendix E, increasing the number of included modes of variation in the model improves accuracy (as measured by the Dice overlap) up to approximately 30 modes, depending on the structure. Figure 3.3 depicts the volumetric results for a single subject from the LOO using 40 modes of variation. The amount



(a) -3σ



(b) mean



(c) $+3\sigma$

Figure 3.2: First mode of variation for the left thalamus. The first column shows the thalamus surface overlaid on the MNI 152 template. The second column is a zoomed-in view, with the conditional mean overlaid in the square patch. The enlarging dark band of intensities at the thalamus border represent the enlarging ventricle that correlates with the translation and shape change seen in the thalamus.

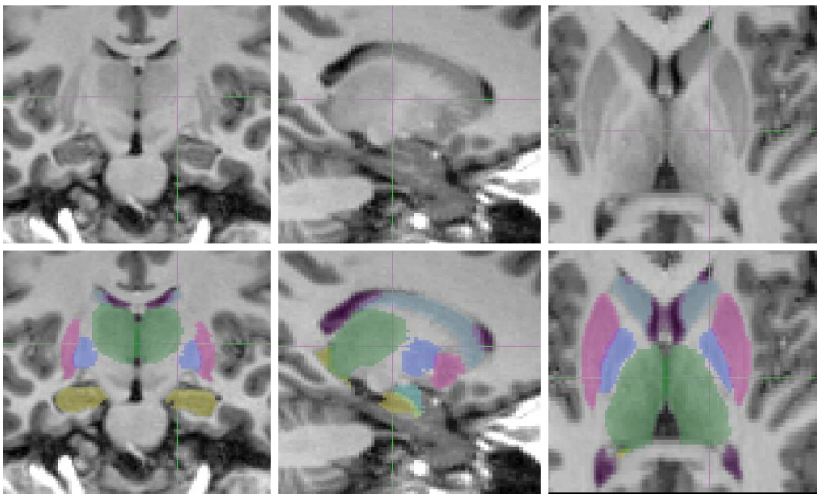


Figure 3.3: Subcortical segmentation produced by fitting the models to a single subject from the LOO.

of improvement due to increasing the number of modes levels off beyond this and in fact at times reduces performance. The reduction in performance is most likely attributed to an overly complex model. For equal performance, the minimal number of eigenvectors is most desirable as it has a large benefit for computation time. The effect of the number of eigenvectors is a reflection of the trade-off between simplicity and complexity.

The results from figure 3.4 show the improvement in overlap due to boundary correction. All structures showed improvement except for the amygdala and accumbens which showed no improvement in overlap with boundary correction. The threshold for boundary correction was arbitrarily chosen, and a higher threshold shows improved results for the amygdala and accumbens, suggesting that the optimal threshold may vary across structure. We have chosen a single threshold (not based on an example image) so as to avoid any bias due to choosing the threshold from the test data.

BCD is useful for comparisons as it decouples the boundary correction algorithm from the fitting, however when performing a practical evaluation for comparison with other software BCD is not appropriate. The mean Dice overlaps from figure 3.4 are comparable with that reported for ASEG. ASEG would appear to perform better on the lateral ventricles and caudate whereas our methodology appears to have a greater mean overlap for the putamen and amygdala. A more thorough comparison of these methods is required to accurately measure relative performance since the methodologies were tested on different data.

Qualitatively the automated method provided smoother boundaries in the posterior/anterior direction than the manual segmentations. The manual segmentation tends to be noisiest in this direction, which is a symptom of performing the segmentation in sequential coronal slices. Therefore the manual segmentation takes into account the 2D geometry for a single coronal slice, the shape model on the other hand relies on the 3D geometry. Since we currently do not have access to the precise inter-rater variability statistics for the manual segmentation, we were unable to make a strict comparison between the manual labelling method and our automated one.

To examine the influence of ϵ_s and ϵ_I on performance we performed a leave-one-out experiment for the left hippocampus for all combinations of ϵ_s and ϵ_I equal to 1, 0.1, 0.01, 0.001, 0.0001%. Figure 3.5 show the boxplots of the BCD for this experiment. By increasing ϵ_s we are decreasing the rigidness of the shape prior (more extreme shapes become more likely under our model). From figure 3.5 the performance appears to be rather insensitive to the choice of ϵ_s. This is encouraging since ϵ_s is empirically estimated. However, for large values of ϵ_I the performance decreased, because it reduces the penalty for intensity deviation away from the conditional mean intensity

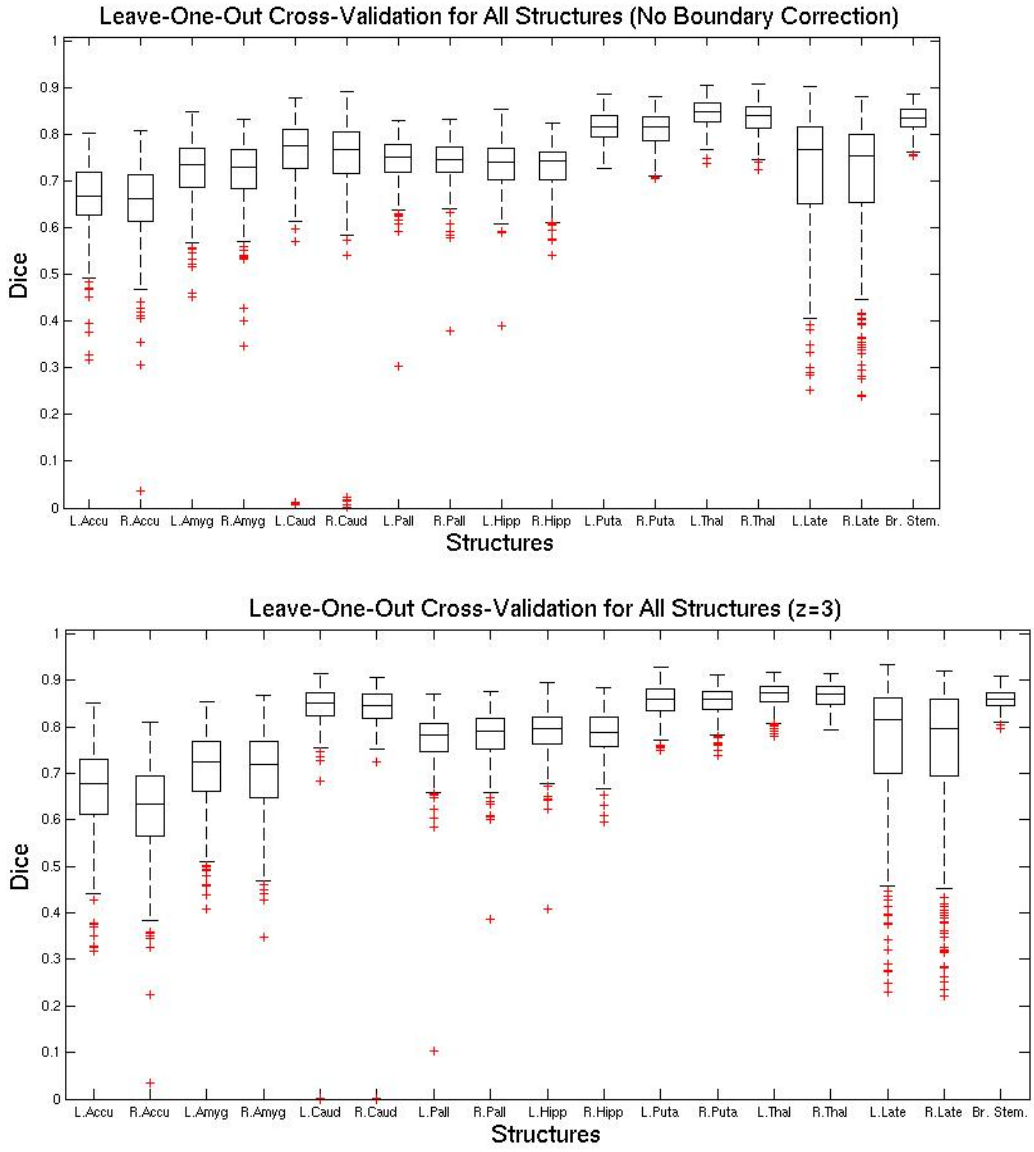


Figure 3.4: Leave-one-out overlap results using optimal number of modes of variation and ϵ_s and ϵ_I equal 0.0001%

and hence relies too strongly on the shape prior. As ϵ_I approaches infinity then the posterior mode approaches the mean shape.

Figure 3.6 shows the BCD results for the left caudate, lateral ventricles and hippocampus when normalizing intensity by its own mode versus that of the thalamus. This concept was discussed in chapter 2 when discussing intensity normalization. The idea being that these structures fail at times due to a confounded estimate of the mode of the intensity distribution, therefore to eliminate these failures we can normalize by a neighbouring structure from which the mode is more robustly estimated. The thalamus was chosen for its robust fitting and proximity to the other structures. The close proximity is desirable to reduce the potentially negative effect of bias field. The improvement is clear for the ventricles. For the caudate and hippocampus, the median and inter-quartile range are similar for both normalization procedures, however, by normalizing by the thalamus intensity mode the extreme outliers were eliminated. Overall, by using a reference distribution that is more robustly estimated from new data we eliminate outliers without decreasing the overall median and inter-quartile range. We achieved the desired result of eliminating outliers due to poor intensity estimation. Figure 3.7 shows the caudate when it failed (as seen in figure 3.6) using its own intensity for normalization and succeeded when normalizing by the thalamus.

An alternative approach to solving the problem of outliers is through the use of a conditional shape prior. This is exemplified by fitting the left caudate conditioned on the left thalamus. Both the left thalamus and left caudate were fit using LOO, figure 3.8 depicts the difference between the Dice overlap for the left caudate fit conditioned on the left thalamus and the Dice overlap for the caudate fit on its own. It is clear that the outliers were removed, performing a two-tailed paired t-

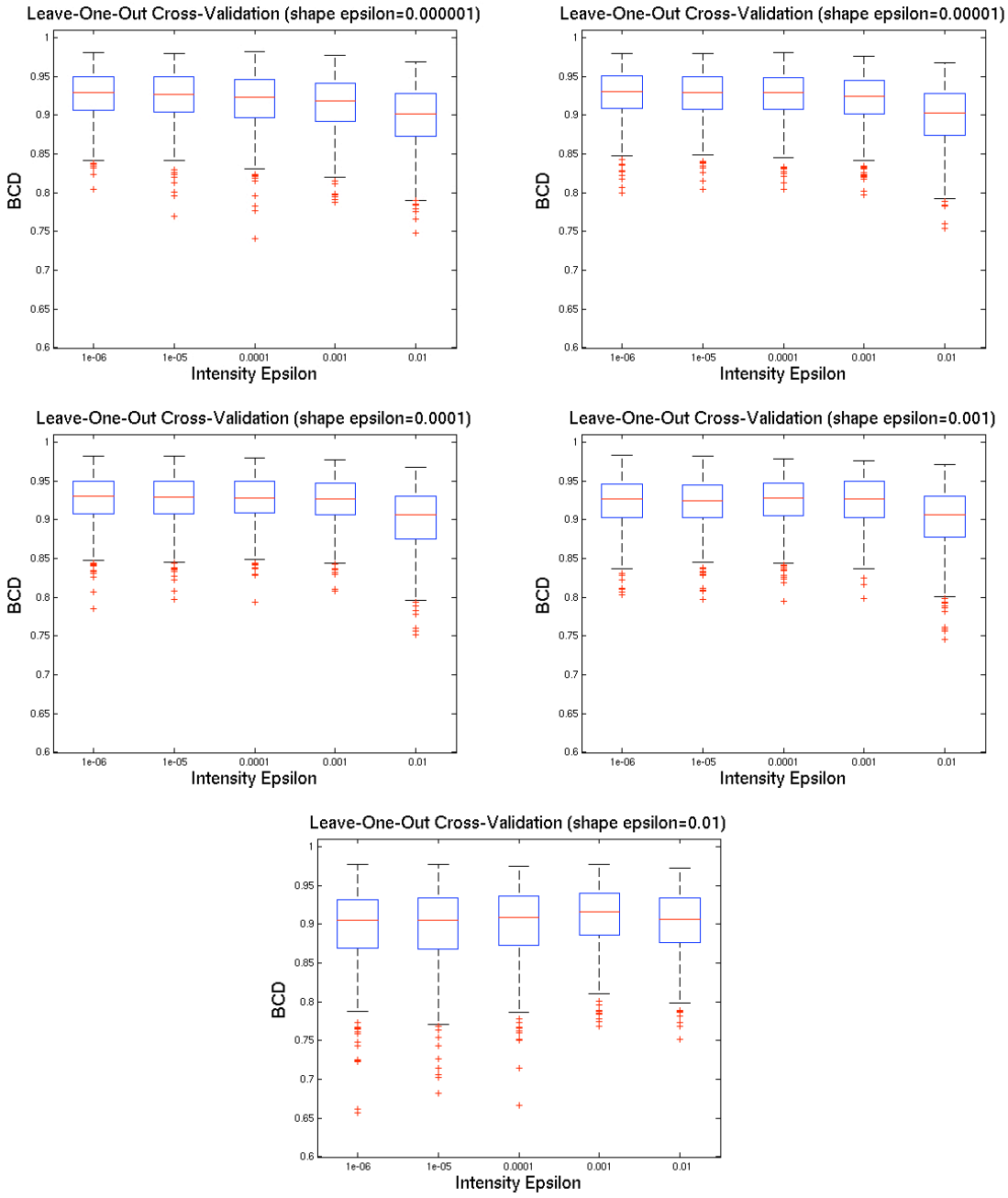


Figure 3.5: Leave-one-out overlap results using 50 modes of variation for all combinations of ϵ_s and ϵ_I equal to 1, 0.1, 0.01, 0.001, 0.0001%

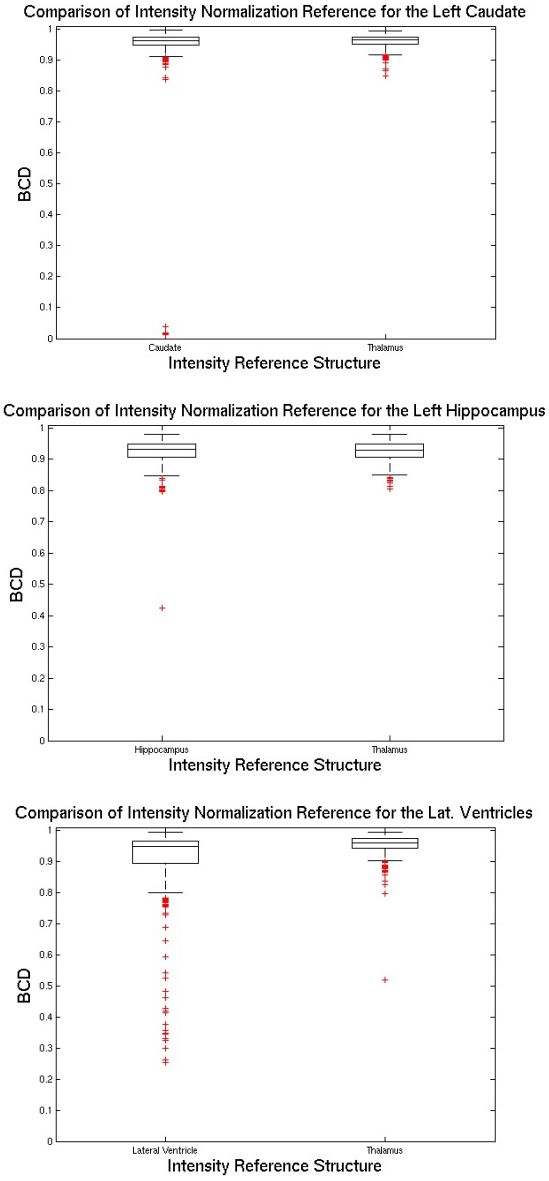


Figure 3.6: Leave-one-out overlap results for the Left Hippocampus using 50 modes of variation. The left boxplot is the BCD when normalizing intensity by its own mode, whereas the right boxplot is the BCD when normalizing intensity by the mode of the thalamus.

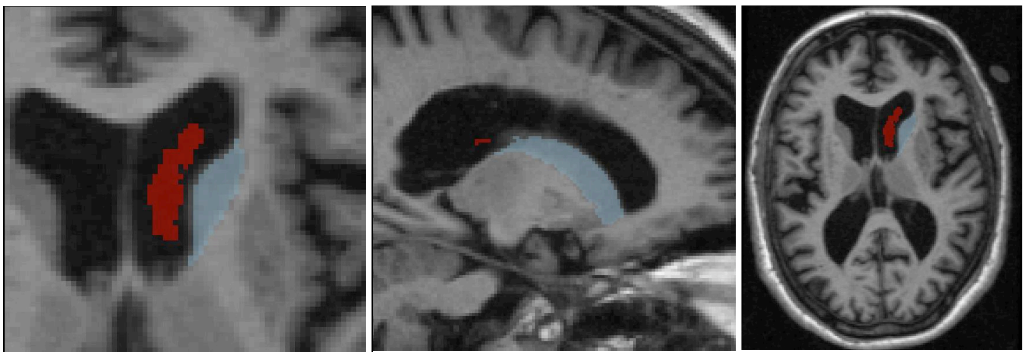


Figure 3.7: The left caudate fit to a T_1-weighted image (using LOO). In red, left caudate fit using its own intensities for normalization. In blue, left caudate fit using the intensities of the left thalamus for normalization.

test between the individual fit and the conditioned fit (omitting outliers), there was no significant change in Dice value (mean difference 6.7×10^{-5}, standard deviation 0.02). Therefore the method was able to remove outliers with no significant effect on the normal fits. When fitting is conditioned on another structure it assumes that the predictive one is correct (in this case the thalamus). The correlation coefficient (omitting outliers) between the Dice overlap for the thalamus and the improvement in Dice for the caudate conditioned on the thalamus (over the individual fit) was found to be 0.026 which converted to a t-statistic is 0.46, suggesting that there is no significant correlation between the two. The lack of correlation in performance may indicate that individually the models are sufficiently well fit, so that the shape conditional is not significantly contributing to the fit, but rather it is dominated by the image intensities (with the exception of the outliers). In general the conditional fit has not proved robust across various structures, though in practice it does provide a useful utility for debugging individual problematic subjects. A more suitable solution

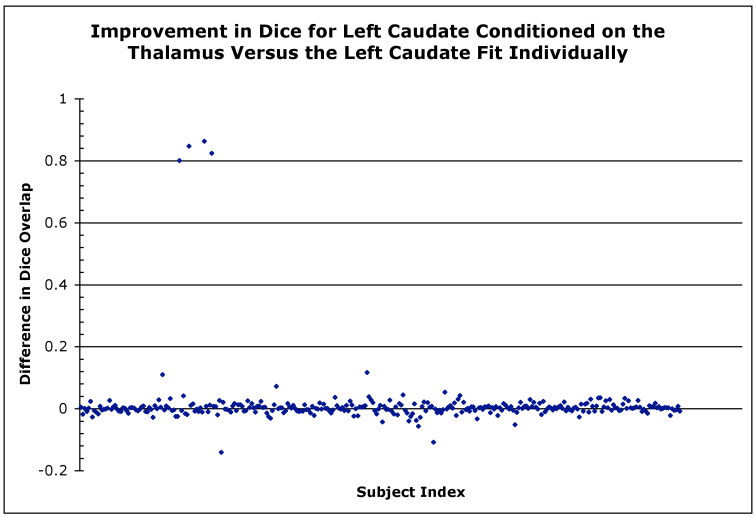


Figure 3.8: Difference in Dice overlap between the left caudate fit conditioned on the left thalamus over that fit individually. The fitting for each used leave-one-out cross-validation with 30 modes of variation for the caudate and 40 modes for the thalamus.

would be to simultaneously fit several individual structures.

The proposed framework for jointly incorporating both shape and intensity information was shown to give accurate and robust results for the segmentation of subcortical structures. The novel aspect of the model was in the modelling of the relationship between shape and intensity through the conditional distribution. The idea is similar to that of the AAM except rather than using an arbitrary empirical estimate of the relationship between shape and intensity, it uses one based on a probabilistic framework. The registration-based methods and level-set methods for segmentation also frame the problem of segmentation in a Bayesian framework although they do not model the relationship between intensity and shape from the training data.

The accuracy of the models using LOO cross-validation is suggestive of the appropriateness of the mesh parameterization method proposed in the previous chapter

as well as its point correspondence. The following chapter will apply our models to clinical data in order to investigate methods for extracting shape information from the resultant surfaces. The results provide further evidence of the appropriateness of our model.

Chapter 4

Size and Shape Analysis from Surfaces with Vertex Correspondence

4.1 Introduction

The end goal of an automated segmentation algorithm is to accurately and robustly extract information from a set of structures. Generally such information may be shape, size, pose, etc... In the medical field, we hope this information will allow us to establish markers of an underlying condition or pathology. These markers may be used to gain a deeper understanding of the mechanisms underlying the pathology and/or to aid in a diagnostic classification in a clinical setting. Discriminants use information to aid in classifications for which it may otherwise be difficult or

even impossible for a human user to make given the same information. The vertex correspondence that is implicit to our shape model and its derived output surfaces facilitates the use of landmark-based statistical shape analysis. The set of corresponding vertices serves as a rich set of information from which to extract shape and/or size metrics. In this chapter we investigate the role of classical multivariate statistics as well as discriminant analysis in interpreting the high-dimensional data (surfaces). In particular, we will focus on methods and metrics that provide easy interpretation in terms of understanding the shape change (although not necessarily the cause of the shape change). The analysis methods are applied to the surfaces, mode parameters and images that are produced from fitting the 17 subcortical shape-and-appearance models to two independent datasets.

A greater understanding of the structural changes in subcortical regions that are associated with a particular pathology may lead to a further understanding of the pathology itself. A measure such as volume (which is commonly used) only provides information about global size of a structure and does not provide any localization of the volume change, nor does it provide any information regarding shape. Localization of differences may isolate the structural variation to particular sub-nuclei of the structure. Amongst other indicators of size and/or shape, we propose a method for assessing local shape variation using independent, vertex-wise analysis across the surface. Differences in Cartesian vertex coordinates reside in an easily visualized and physically meaningful space.

We make use of classical inference techniques to test for differences in group means of volume, surface area, vertex location, and local triangular face area. Each measure contributes some different piece of information regarding shape and/or size. For

univariate tests, the general linear model (GLM) and resulting t-statistic are used to assess statistical difference [31]. In the multivariate case, we use Pillai's trace for assessing differences in means [49]. The statistics will be discussed in section 4.2.

Along with classical inference we use discriminant analysis to investigate group differences in size and/or shape. Discriminant analysis examines the ability of a model to dissociate groups from within a dataset. By model we are referring to the mathematical model of the feature space being used to discriminate, this usually corresponds to a probability distribution function (pdf) of the feature for each class. Generally classification is made based on a decision boundary that is defined by the model and its estimated parameters; the parameters are estimated from a set of subjects with their discriminant features and known classification (training set). In this chapter, the discriminant features consist of size and shape characteristics that are derived from the mesh, such as volume, surface area or vertex location. Good discrimination based on a given feature is indicative of a relationship between that feature and the class membership. The ability of the decision boundary to correctly classify data and its generalization to the larger population is quantified using the prediction error (and its counterpart, prediction accuracy). Methods to assess prediction error in an unbiased manner will be discussed later in this chapter.

We will begin by reviewing the general linear model (GLM) and some univariate and multivariate statistics. This will be followed by a review of discriminant methods, methods for assessing prediction error and recent applications of discriminant analysis in the field of structural neuroimaging. We will then describe the methods used in this chapter for investigating shape differences between two clinical groups. Finally, we will present and discuss the results of applying these statistical tests and discriminant

methods to two clinical datasets. More precisely, they are applied to the output of the fitted models.

4.2 The General Linear Model (GLM), Univariate and Multivariate Tests.

This section is devoted to reviewing some of the standard univariate and multivariate statistics for investigating group differences. In particular, we will review the GLM and ordinary least-squares regression (OLS), along with classical univariate and multivariate parametric statistics such as the t-statistic, Wilks' lambda, Pillai's trace, Hotelling's Trace and Roy's greatest characteristic root. To investigate group differences in size and shape we will apply these methods to volumes, surface area, individual vertex location and the area of the individual triangular faces of the mesh. We will discuss these metrics and their interpretability in further detail in section 4.6.5.

The GLM models the relationship between multiple explanatory variables (EVs) and a response variable. In our case, the explanatory variables would be group membership, age and/or gender, whereas the response variable may be volume, vertex location, etc. The GLM is an additive model that states that an outcome $\boldsymbol{y}$ is a linear combination of input variables $\boldsymbol{x}_i$ with weighting parameters β_i. The model is expressed as

$$\boldsymbol{y} = \sum_{i=1}^{M} \boldsymbol{x}_i \beta_i, \tag{4.1}$$

where M is the number of explanatory variables and $\boldsymbol{y}$ is of dimensionality k.

To estimate the model parameters β_i from the data, we employ the ordinary least-squares (OLS) solution. We will restrict ourselves to the OLS solution since it is a well-established and widely prevalent method for parameter estimation, used throughout most of classical inference. The OLS solution amounts to the minimization, with respect to β_i, of the residual squared-error between the outcome $\boldsymbol{y}$ and the predicted outcome $\hat{\boldsymbol{y}}$. Generalised to the multivariate case, the expression for the residual squared-error is given by

$$E = \sum_{j=1}^{k} \sum_{i=1}^{N} \left[\boldsymbol{y}_{i,j} - \sum_{l=1}^{M} \boldsymbol{x}_{i,l} \beta_l \right] \left[\boldsymbol{y}_{i,j} - \sum_{m=1}^{M} \boldsymbol{x}_{i,m} \beta_m \right], \tag{4.2}$$

where k is the dimensionality of $\boldsymbol{y}_i$, N is the number of samples, $\boldsymbol{y}_{i,j}$ is the j^{th} component of the i^{th} outcome sample, $\boldsymbol{x}_{i,l}$ is the l^{th} component of the i^{th} sample of the explanatory variable and β_l is the l^{th} model parameter of the GLM. The closed form, OLS solution for β_i is derived by taking the derivative with respect to β_i and setting it equal to zero. The solution for the estimated model parameters, expressed in matrix form [45], is given by

$$\hat{\boldsymbol{\beta}} = \left[\boldsymbol{X}^T \boldsymbol{X} \right]^{-1} \boldsymbol{X}^T \boldsymbol{Y}, \tag{4.3}$$

where $\boldsymbol{X}$ is the matrix of explanatory variable samples (design matrix), $\boldsymbol{Y}$ is the matrix of sampled outcomes, and $\hat{\boldsymbol{\beta}}$ are estimated model parameters of the GLM.

The residual error (an approximation to the noise) is assumed to be normally dis-

tributed, $N(0, \mathbf{\Sigma}_e)$, with zero mean. The noise variance is estimated by

$$\mathbf{\Sigma}_e = \frac{1}{N-M} \left(\boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}\right)^T \left(\boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}\right). \tag{4.4}$$

To be consistent with the standard univariate notation for variance, we define the noise variance for the univariate case as $\sigma_e^2 = \Sigma_e$, where σ_e^2 is a scalar.

The OLS solution for $\hat{\boldsymbol{\beta}}$ corresponds to the maximum-likelihood solution under the assumption of Gaussian random variables. Under this assumption, the GLM parameters will be normally distributed with a Wishart-distributed noise variance (Chi-square for the univariate case). The ratio of the two is a Student's distribution and hence the t-statistic is employed to assess the statistical difference between groups in the univariate case. In the multivariate case, Wilks' lambda, Pillai's trace, Hotelling's Trace or Roy's largest characteristic root may be used to test for significance. In this chapter, we use Pillai's trace to assess multivariate differences in group mean since it is typically regarded as the most robust of the four statistics [49].

In the univariate case, either a t-test or F-test may be used to test for the significance of an EV. For a single effect the F-statistic is equal to the square of the t-statistic. Given a hypothesis contrast vector $\boldsymbol{c}$ of dimension $N \times 1$ and a design matrix $\boldsymbol{X}$, the standard expression for the t-statistic is given by

$$t = \frac{\boldsymbol{c}^T\hat{\boldsymbol{\beta}}}{\sigma_e\sqrt{\boldsymbol{c}^T\left(\boldsymbol{X}^T\boldsymbol{X}\right)\boldsymbol{c}}}. \tag{4.5}$$

To investigate group differences in vertex location we need to use a multivariate statistic (three dimensions for a single vertex). Four multivariate statistics may typically

be used to assess the significance of an effect in the multivariate case. The statistics are: 1) Wilks' lambda, 2) Pillai's trace, 3) Hotelling's Trace and 4) Roy's greatest characteristic root. Each of the four statistics is a function of the error covariance matrix E and the effect covariance matrix H. E is the covariance of the residual error (Σ_e) to the full model (X), and H is the (co)variance explained by the explanatory variable(s) of interest. Wilk's lambda, Pillai's trace, Hotelling's Trace and Roy's greatest characteristic root are expressed in equations (4.6a), (4.6b), (4.6c), (4.6d) respectively.

$$\Lambda = \frac{|\boldsymbol{E}|}{|\boldsymbol{E}+\boldsymbol{H}|}, \tag{4.6a}$$

$$\mathcal{V} = \mathrm{Tr}\left((\boldsymbol{H}+\boldsymbol{E})^{-1}\boldsymbol{H}\right), \tag{4.6b}$$

$$\mathrm{HT} = \mathrm{Tr}\left(\boldsymbol{H}\left(\boldsymbol{E}\right)^{-1}\right), \tag{4.6c}$$

$$\mathrm{GCR} = \lambda_0, \tag{4.6d}$$

where $|...|$ is the determinant of the matrix, Tr is the trace of the matrix, and λ_0 is the largest eigenvalue of the matrix given by $\boldsymbol{H}\left(\boldsymbol{H}+\boldsymbol{E}\right)^{-1}$.

The multivariate statistics are a function of the amount of variance explained by the dependent variable of interest (the effect) versus the total error variance. For large sample sizes the choice of statistic is not crucial, although in most situations Pillai's trace seems to be the preferred statistic [49]. Roy's greatest characteristic root is typically the least sensitive of the four, except in situations where the change is in a dominant single direction.

Wilks' lambda varies between 0 and 1 such that the closer the statistic is to zero, the more significant is the effect. It is the case for the other three multivariate

statistics that the larger the statistic the more significant the effect. The F-statistic is more commonly used in statistics and is thus more readily interpretable for testing significance and visualizing local shape differences. It is for this reason that we approximate Pillai's trace with an F-statistic using equations (4.7a), (4.7b), (4.7c), (4.7d).

$$s = \min\left(k, g-1\right), \tag{4.7a}$$

$$t = \frac{|k-g-1|-1}{2}, \tag{4.7b}$$

$$u = \frac{N-g-k-1}{2}, \tag{4.7c}$$

$$F_{(s(2t+s+1),s(2u+s+1))} = \frac{2u+s+1}{2t+s+1}\left(\frac{\mathcal{V}}{s-\mathcal{V}}\right), \tag{4.7d}$$

where N is the total sample size, g is the number of dependent variables, and k is the dimensionality of the response variable, and $\mathcal{V}$ is Pillai's trace. This is a standard approximation for Pillai's trace [49].

4.3 Discriminant Analysis

Discriminant analysis refers to the use of the classification ability of a model to assess group differences, the parameters of which are typically estimated from a set of training data. Training data is made up of a group of subjects, their discriminant features, and their known classification. A discriminant feature is typically a quantity measured from the data, which in our case would be a measure derived from the T_1-weighted image. The segmentation derived from our shape-and-appearance model provides a means of extracting size-and-shape information for subcortical structures

from the image, which in turn may be used to discriminate between groups. In addition to measurements derived from the image, features such as age and gender may also be used. Discriminant analysis does not aim to infer whether two means are equal, but rather asks whether a specified boundary may distinguish between the different groups. The idea is that if a feature may be used to accurately identify class membership then one may conclude that there is a systematic difference between groups. Performance is measured in terms of prediction error. In order to make comparisons between classifiers it is important to assess the variability in the prediction error estimate.

Many discriminant methods are formulated probabilistically and aim to classify based on the maximum posterior likelihood of class membership given the data, $P(G = c \mid X)$, where G is group membership (class is indicated by $c = 0, 1, ..., N-1$) and X is the input feature vector of k dimensions. The decision boundary between class i and j is the function that satisfies $P(G = i \mid X) = P(G = j \mid X)$. The form of the densities may vary across methods, although some of the most common rely on Gaussian models (e.g. linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) [30]). An alternative approach is to define a hyperplane that provides maximum margin between classes; this will be discussed in further detail when reviewing support vector machines (SVM). We will now briefly discuss the role of dimensionality reduction and then review some of the most prevalent discriminant methods that are used in practice.

In general, dimensionality plays a large role in the performance of discriminants. Better discrimination can be achieved by reducing the dimensionality of the space prior to applying a discriminant. The need for dimensionality reduction is a result of

a low sampling density in many applications. The sampling density is the number of samples relative to the dimensionality of the space, and is proportional to $N^{\frac{1}{p}}$ [30]; N is the sample size and p is the dimensionality. As an example of the relationship between sample size, dimensionality and sampling density, we consider a set of evenly spaced samples taken from a space defined by a square (a 2D space). To evenly sample a 3D space defined by a cube (same edge length as the square) using the same number of samples, a sparser sampling is required (larger intervals between samples). To achieve the same sampling density for the cube as for the square, it would require a factor of 1.58 more samples. To achieve the same sampling density for a higher dimensional space (dimensionality of p_2) as for a lower dimensional space (dimensionality of p_1), the sample size required is equal to $N_1 \log_{p_1}(p_2)$ (N_1 is the number of samples for the lower dimensional space). Generally, the higher the sampling density, the more accurate are the parameter estimates that define the space (and hence decision boundary). Frequently in medical imaging we are dealing with dimensions in the order of tens, hundreds or even thousands, and consequently the sample size is frequently less than the number of dimensions; this is analogous to estimating the shape of a 3D-ellipsoid using two samples from that space. The problem is ill-posed from the start, however, since the sampling density changes exponentially with the number of dimensions, by reducing the number of dimensions, we increase our sampling density exponentially.

The idea behind feature selection (dimensionality reduction) is that not all of the dimensions (features) contribute to the discrimination between groups. If we eliminate the non-informative dimensions, we will exponentially increase our sampling density whilst not discarding any useful information. In practice, to determine whether a

dimension contributes to discrimination is another problem for which there is no clear, optimal solution; one common approach is to perform a univariate t-test on the individual dimensions. Despite the fact that informative dimensions may be discarded in the feature selection process, the increase in sampling density may outweigh the loss of information. The key is to find the balance between sampling density and the number of informative dimensions. Even though a well-defined discriminant boundary may exist in the "true" feature space, if we are unable to estimate the boundary accurately, it may be inconsequential.

4.3.1 Linear and Quadratic Discriminant Analysis

Linear discriminant analysis (LDA) is a powerful and well established tool for discrimination [30]. Assuming an underlying Gaussian distribution with equal group variances, a linear decision boundary is constructed such that it minimizes the within-group variance whilst maximizing the between-group variance. Quadratic discriminant analysis is based on the same assumptions except for that of equal variances. We will first derive the general expression for QDA and will then impose the equal variance constraints of LDA.

The class density, $f_c(x)$, takes the form of a multivariate Gaussian as given by,

$$f_c(\boldsymbol{x}) = P(X \mid G = c) = \frac{1}{(2\pi)^{\frac{k}{2}} \left|\boldsymbol{\Sigma}_c\right|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_c)\boldsymbol{\Sigma}_c^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_c)}, \tag{4.8}$$

where c is the class, C is the number of classes, $\boldsymbol{x}$ is the k dimensional feature vector and $\boldsymbol{\Sigma}_c$ is the feature's within-class covariance matrix of size $k \times k$. If the

prior class probability, $\pi_c = P(G)$, is constrained such that $\sum_{c=1}^{C} \pi_c = 1$, then by a straightforward application of Bayes' rule we arrive at,

$$\mathrm{P}\left(G \mid X = \boldsymbol{x}\right) = \frac{f_c(\boldsymbol{x})\pi_c}{\sum_{c=1}^{C} f_c(\boldsymbol{x})\pi_c}. \tag{4.9}$$

The decision about class membership is made based on the ratio of class posteriors. It is convenient to express the ratio as the log-likelihood,

$$\begin{aligned} log\left(\frac{P(G = c_1 \mid X = \boldsymbol{x})}{P(G = c_2 \mid X = \boldsymbol{x})}\right) =& log\left(\frac{|\boldsymbol{\Sigma}_{c_2}|^{\frac{1}{2}}}{|\boldsymbol{\Sigma}_{c_1}|^{\frac{1}{2}}}\right) - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_{c_1})^T \boldsymbol{\Sigma}_{c_1}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{c_1}) \\ &+ \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_{c_2})^T \boldsymbol{\Sigma}_{c_2}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{c_2}) + \log\frac{\pi_{c_1}}{\pi_{c_2}}. \end{aligned} \tag{4.10}$$

This is the decision function used by QDA to classify data; it contains both linear and quadratic terms. Assuming that for all classes, $|\boldsymbol{\Sigma}_c| = |\boldsymbol{\Sigma}|$, then equation (4.10) may be simplified to obtain the following expression for the decision function that is used by LDA (equation 4.11).

$$\begin{aligned} log\left(\frac{P(G = c_1 \mid X = \boldsymbol{x})}{P(G = c_2 \mid X = \boldsymbol{x})}\right) =& \log\frac{\pi_{c_1}}{\pi_{c_2}} - \frac{1}{2}(\boldsymbol{\mu}_{c_1} + \boldsymbol{\mu}_{c_2})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_{c_1} - \boldsymbol{\mu}_{c_2}) \\ &+ \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}_{c_1} - \boldsymbol{\mu}_{c_2}\right). \end{aligned} \tag{4.11}$$

The assumption of equal variance has simplified the discriminant function by eliminating the quadratic terms. A practical implication is that the decision function for QDA requires more parameters that need estimating than for LDA. The increased number of parameters for QDA is due to the estimation of multiple covariance matrices; for the two class case LDA has $k\left(\frac{k}{2} + 1\right)$ fewer parameters than QDA. We are able to pool all subjects across classes when estimating the covariance matrix for LDA. Despite the fact that the equal variance assumption may not always be valid,

LDA may still outperform QDA due to the reduction in the number of parameters. The relative performance between LDA and QDA is dependent on the sample size and nature of the discriminant boundary. LDA and QDA have the advantage of being easily interpretable because of the simplicity of the model.

4.3.2 Logistic Regression

Logistic regression is used to model the posterior probabilities via linear functions of the input variables, the model takes the form of $C-1$ log-odds functions as given by

$$\log\left(\frac{P\left(G=c \mid X=\boldsymbol{x}\right)}{P\left(G=C \mid X=\boldsymbol{x}\right)}\right)=\beta_{c0}+\boldsymbol{\beta_C}^T\boldsymbol{x}, \tag{4.12}$$

for classes $c=1,2,...,C-1$, where β_{c0} is the parameter estimate corresponding to the mean offset, and $\boldsymbol{\beta_C}^T$ is the weighting vector (parameter estimates) corresponding to each discriminant variable. The final class is used for normalization, however the solutions are equivalent, regardless of the normalization class. By constraining the sum of class probabilities to 1 and using some simple calculations, the class probability is expressed as

$$P\left(G=c \mid X=\boldsymbol{x}\right)=\frac{\exp\left(\beta_{c0}+\boldsymbol{\beta_C}^T\boldsymbol{x}\right)}{1+\sum_{c=1}^{C-1}\exp\left(\beta_{c0}+\boldsymbol{\beta_C}^T\boldsymbol{x}\right)}, \tag{4.13}$$

for classes $c=1,2,...,C-1$. Unlike LDA and QDA, logistic regression does not make Gaussian assumptions about the input variables. An iterative reweighted least-squares algorithm may be used to estimate the maximum likelihood solution for the model parameters.

4.3.3 Support Vector Machine

Support Vector Machines (SVM) are a class of discriminant that is based on the idea of finding a maximally separating hyperplane [13]. A maximally separating hyperplane is defined as a hyperplane that maximizes the minimal distance between a point on the hyperplane and the nearest training point. Traditionally, this definition does not allow for mis-classification of subjects (hard margin). It is also possible to have an SVM that uses the concept of a soft margin such that there is an error-tolerance to allow for label mis-classification. The ability to allow for mis-classification is essential for preventing over-fitting of the hyperplane to the data, furthermore, it may also be the case that a separating hyperplane does not exist.

The separating hyper-plane may be defined by a subset of the input features (support vectors) that lie on the boundary. SVMs define the discriminant boundary using a subset of the input vectors; the vectors contained within this subset are the "support vectors". The SVM solves a constrained optimization problem using Lagrange multipliers, where the support vectors correspond to the the input vectors with non-zero Lagrangian multipliers. The SVM allows the use of kernels to map the data into a non-linear space, producing a non-linear discriminant boundary.

4.3.4 Relevance Vector Machines

The relevance vector machine (RVM) [58] is a discriminant method using a sparse representation of the data within a Bayesian framework. Similar to the SVM, the RVM builds the discriminant based on a subset of the input vectors, these are the so-

called relevance vectors. The manner in which the relevance vectors are calculated is fundamentally different from that which is used by the SVM to calculate the support vectors. The RVM models the conditional probability of the target (for our case, the target is a 0-1 group classification) given an input vector as a multivariate Gaussian distribution of the form

$$P\left(\boldsymbol{t} \mid \boldsymbol{x}, \boldsymbol{\omega}, \tau\right) = \mathcal{N}\left(\boldsymbol{t} \mid y(\boldsymbol{x}, \boldsymbol{\omega}), \tau^{-1}\right), \tag{4.14}$$

where

$$y(\boldsymbol{x}, \boldsymbol{\omega}) = \sum_{m=0}^{M} \omega_m \phi_m(\boldsymbol{x}), \tag{4.15}$$

such that $\mathcal{N}$ represents the multivariate Gaussian (Normal) distribution, $\boldsymbol{\omega}$ are the parameters of the model or input weightings, $\boldsymbol{x}$ is the input data, τ is the noise variance and $y(\boldsymbol{x}, \boldsymbol{\omega})$ is the weighted sum of $\phi(\boldsymbol{x})$. For a linear RVM $\phi(\boldsymbol{x})$ represents the input data, alternatively, it may represent a kernel that maps the input vectors into a non-linear space. The parameters $\boldsymbol{\omega}$ are given Gaussian priors of the form

$$\boldsymbol{\omega} = \mathcal{N}\left(\omega_m \mid 0, \alpha_m^{-1}\right), \tag{4.16}$$

where one parameter α_m is assigned to each weight, as α_m approaches 0 the contribution of the feature associated with ω_m approaches 0. It is in this manner that the RVM achieves its sparse representation. If the target was a continuous variable rather than a binary classification the RVM may be used to perform regression between the input variables and the target data.

The RVM obtains the good predictive performance seen with the SVM as well as

preserving the sparse representations of the data. In fact the RVM at times produces even sparser representations than for the SVM. A particular advantage of the RVM, as opposed to the SVM, is that the RVM produces posterior probabilities rather than solely hard classifications of the data. The class probabilities are converted to hard classifications by thresholding the probabilities at 0.5.

4.4 Assessment of Performance

For discriminant analysis it is important to assess the generalizability of the model to independent test data. The quality of performance is assessed by the prediction error when applied to the test data. Prediction error is defined as the mean prediction loss that is given by

$$e\bar{r}r = \frac{1}{N}\sum_{i=1}^{N} L(\hat{y}_i, y_i), \tag{4.17}$$

where $L(\hat{y}_i, y_i)$ is the loss function between the true and predicted outcome. We restrict our discussion to binary classification and are thus only concerned with the $0-1$ loss function as given by:

$$L(\hat{y}_i, y_i) = \begin{cases} 0 & \text{if } \hat{y}_i = y_i \ , \\ 1 & \text{if } \hat{y}_i \neq y_i. \end{cases} \tag{4.18}$$

where $\hat{y}_i$ is the predicted outcome, y_i is the actual outcome and L is the loss function. If the test data is not independent of the training data then the prediction error would be overly optimistic due to overfitting. This is an important point when discussing discrimination as a diagnostic tool or for understanding differences between groups.

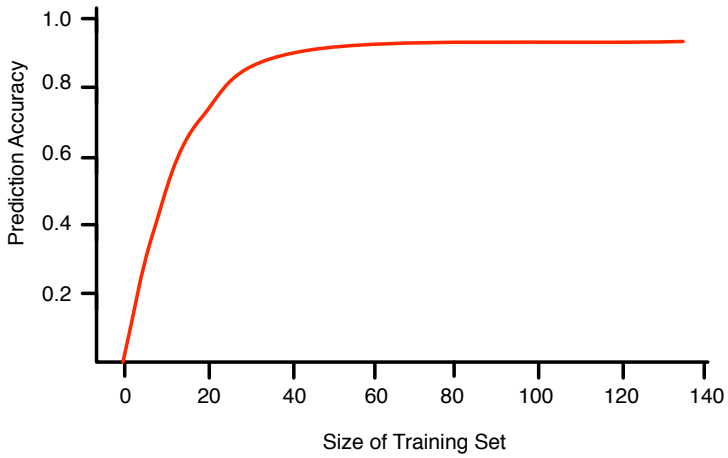


Figure 4.1: Hypothetical learning curve.

If the discriminant method does not generalize well to the population then, based on the discriminant analysis, we are unable to draw any valid conclusions about the population at large.

Ideally, provided with a large amount of training data, the data would be split into three partitions: 1) a partition for training, 2) a partition to optimize any tuning parameters, and 3) a partition on which to test. The third group should be vaulted until after all training has been performed to ensure independence between the training and test data. Unfortunately, particularly in the medical field, the number of available training/test subjects is limited. As such we do not have an adequate sampling density to randomly separate our training set into two or three independent groups whilst retaining an adequate training size to obtain an unbiased estimate of prediction error. The problem is that as we reduce the number of training subjects we may reduce the accuracy of the decision boundary estimation. An example of a typical learning curve is shown in figure 4.1. Bias occurs when the difference in performance for the number of subjects in the original set and the number of samples in

the subsampled set is large (i.e., areas of high slope on the training curve). Hence the further out on the plateau we reside, the more data we may omit from the training set without adding bias. The caveat here is that we do not know the true shape of the learning curve for a given model and application. Care must be taken to not bias the estimated prediction error. Though not ideal, other methods such as boot strapping may be used for assessing the mean and variability in an un-biased manner [30].

Intuitively, estimating prediction error will be overly optimistic when estimating from test data that overlaps the training set. The degree of optimism of the estimate is related to the model complexity and the size of the training set. As the complexity increases, it becomes more likely that the model will overfit the training data. With overlap between test and training data, overfitting to the training data will result in overfitting to the test data and hence an overly-optimistic estimate of prediction error. We would also expect the degree of optimism to go down with increased sample size because of a reduction in overfitting due to a more representative sample set of the population. Therefore, in order to obtain an unbiased estimate of prediction error there must be no overlap between the test and training data.

4.4.1 N-fold Cross-Validation

Cross-validation is a widely used and accepted method for assessing prediction error [57, 39, 21, 24, 61, 30]. The training set is separated into N partitions of equal size, each partition of data is used as a test set for a discriminant that is trained from the remaining partitions. In this manner each sample is used as test data, without ever having overlap between a test sample and the training set. The prediction error

is then accumulated across partitions, and for an N fold cross-validation is given by

$$\bar{\text{err}} = \frac{1}{N}\sum_{i=1}^{N}\bar{\text{err}}_i, \tag{4.19}$$

where $\bar{\text{err}}_i$ is the prediction error associated with using the i^{th} partition as the test set and N is the number of partitions.

There is a trade-off between bias and variance relating to the choice of N; the higher the N (smaller partitions), the lower the bias but the higher the variance. The special case of N equalling the number of training datasets has been given the name Leave-One-Out (LOO) cross-validation. LOO is prevalent in medical imaging because of its low bias. The low bias is particularly desirable because of the low sample sizes, where it is likely that the classifier does not reside on the plateau of the learning curve. LOO gives an approximately unbiased estimate of the true prediction error, however, has higher variance since the training sets are very similar to each other. Unfortunately, the use of LOO cross-validation does not provide an estimate of the variance in the prediction error estimate (i.e., we can estimate classification accuracy, but don't know what the uncertainty of this estimate is).

4.4.2 Boot Strapping

Boot strapping is an alternative method for estimating prediction error. A set of subjects is randomly sampled from the entire training set with replacement (one subject may be drawn multiple times) such that the bootstrap sample set is the same size as the original training set. The discriminant is trained from the bootstrap sample

set and is applied to original training data to assess prediction error. This approach allows for the calculation of statistics such as the expectation and variance of the prediction error. The bootstrap procedure is biased in that the bootstrap sample set partially overlaps the test data. Given B bootstrap samples of size N, the prediction error is given by

$$\hat{Err}_{boot} = \frac{1}{B}\frac{1}{N}\sum_{b=1}^{B}\sum_{i=1}^{N} \mathrm{L}\left(y_i, \hat{f}^b\left(x_i\right)\right), \tag{4.20}$$

where $\hat{f}^b$ is the discriminant function trained from the b^{th} bootstrap sample set, y_i is the true outcome and L is the binary loss function.

By combining cross-validation and bootstrapping we may achieve an estimate of the prediction error that is nearly unbiased by overlap between the test data and the bootstrap sample set. For each bootstrap sample set the discriminant is tested on the subjects from the training set that were not included in that sample set. We then accumulate the loss for each subject across bootstrap sample sets, which is then used to calculate the prediction error. Provided that the number of bootstrap sample sets is sufficiently large, each subject in the original training set should be used for testing at least once. Using the cross-validation/bootstrap, the prediction error is given by

$$\hat{Err}_{boot,cv} = \frac{1}{B}\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|C_i|}\sum_{b\epsilon C_i} \mathrm{L}\left(y_i, \hat{f}^b\left(x_i\right)\right), \tag{4.21}$$

where C_i is the set of bootstrap sample sets for which the i^{th} subject is not included. That is, a given bootstrap sample set (i.e. set of randomly sampled subjects) is included as a member of C_i if the i^{th} subject is not included in that bootstrap sample set. For each element of C_i, the predicted outcome is the result of a new model estimate trained from a random sampling of the training set.

The expected number of unique subjects from the training set that are included in each bootstrap sample set is $0.632N$ [30]. The reduction in the number of unique subjects will affect where we reside on the learning curve, thus bootstrapping will approach the sample-size bias of approximately a 2-fold cross-validation. Given the sample sizes in medical imaging, this could mean an overly conservative, biased estimate of prediction error.

To be able to statistically compare estimates of $\hat{Err}_{boot/cv}$, we also need to estimate the variance in the prediction error estimation. To estimate variance, we estimated $\hat{Err}_{boot,cv}$ M times, with independent sampling, and evaluate the expected prediction error and variance as given by

$$E[\hat{Err}_{boot/cv}] = \frac{1}{M}\sum_{k=1}^{M} \hat{Err}_{boot/cv,k}, \tag{4.22a}$$

$$Var[\hat{Err}_{boot/cv}] = \frac{1}{M-1}\sum_{k=1}^{M}\left(\hat{Err}_{boot/cv,k} - \frac{1}{M}\sum_{k'=1}^{M}\hat{Err}_{boot/cv,k'}\right)^2. \tag{4.22b}$$

In estimating the variance of the prediction error estimate, since the bootstrap sample sets are independently sampled, there is the possibility that there is a common bootstrap sample set within the M estimations of prediction error. A common sample set would imply correlation across the prediction error estimates; if so, the variance estimate will be biased. A numerical simulation was run for a bootstrap sample set size of 100 (this is the size that is chosen for our analysis in this chapter), the probability of observing a repeated bootstrap sample set was estimated to be less than 10^{-6}. Given the probability of obtaining a duplicate bootstrap sample set, we assume there is no correlation between prediction error estimates and that equation

(4.22b) is unbiased for $N = 100$.

With estimates of the expected value and variance for $\hat{Err}_{boot/cv}$ we are able to test for significant differences in discriminant performance using a t-test. Rather than prediction error, prediction accuracy is often reported (particularly in neuroimaging). All the same theory applies to accuracy except that the prediction accuracy is defined as $\text{Pred}_{acc} = 1 - \bar{\text{err}}$. This applies for the cross-validation and bootstrap estimates alike. We will report the prediction accuracy in this chapter.

4.5 Literature Review of Discriminant Analysis in Neuroimaging.

Lao et al. [39] use a non-linear SVM to classify gender and age via morphometric measures. Rather than utilizing each voxel individually, they focus on incorporating the inter-voxel correlations to discriminate. Images are aligned to a template using non-linear warps, these transformations are mass-preserving such that in areas of contraction and expansion the tissue density is adjusted to ensure that the total tissue mass remains constant. The result of the shape transformation is a white matter, grey matter, and CSF density map with voxel correspondence across subjects, the values of which are combined across all voxels and all maps to form a high dimensional feature vector used for discrimination. The features are fed into a non-linear SVM for discrimination, but first dimensionality reduction is performed by downsampling followed by a wavelet decomposition and feature selection. For validation, the method is applied to simulated atrophy data where the underlying truth is known.

In addition, permutation testing is used to confirm that an expected prediction error of 50% is attained in the absence of any effect. LOO cross-validation is used to assess prediction accuracy between age groups as well as gender for real data. A prediction accuracy between males and females of 96.7% was reported. For age discrimination, the subjects were divided into four groups with a mean age of 58.14, 64.39, 72.78 and 82.35 respectively. A prediction accuracy between group 1 and 4, 1 and 2, 2 and 3, and 3 and 4 were 97.5%, 81.5%, 71.6%, and 74% respectively. The high prediction accuracy between groups 1 and 4 suggests that there is larger morphological difference between the larger age gap. The main contribution of this work is the use of all voxels simultaneously versus the use of individual voxels separately. In order to cope with the large dimensionality of the problem down-sampling, wavelet decomposition and feature selection is performed. Our goal in using discriminant analysis is not purely in finding the maximum discrimination but rather in using discriminability to understand shape differences. For this reason we discriminate based on several features independently, since the different features provide different pieces of shape information. By separating the discriminants, we may separate out the shape changes that are attributed to the group difference.

Davatzikos et al. [16] investigates the application of discriminants to groups of normals and schizophrenia patients. The methodology used for discrimination is the same as that previously described for [39]. The discriminant was able to separate the training data with 100% accuracy, which is not atypical for non-linear classifiers given their flexibility. A prediction accuracy of 81.1% was reported for women and men combined; when considering discrimination within males and females in isolation, accuracy rates of 85.0% and 82.0% were reported respectively. A permutation

test is used to assess the null distribution and test for significance. Since variability in the discriminant accuracy is not assessed, conclusions regarding the significant differences between the reported prediction accuracies cannot be made. For example, it is unclear whether the discriminants perform significantly better for a gender in isolation over the two combined.

In Fan et al. [24] an adaptive feature extraction is proposed where ROIs may adapt to the particular pathology. The original set of features are the tissue density maps that were generated using the same mass-preserving transformation scheme that was used for [39, 16]. Pearson's correlation coefficient between the tissue density and group for each voxel is used to assess to the relevance of each voxel for discrimination. In addition, the spatial consistency of features is measured using a two-way random effects model using a feature matrix based on immediate neighbours. The product of the relevance and spatial consistency measure is used as a single measure on which a watershed algorithm is applied to define discriminant regions. This approach groups regions based on discriminability and not on pre-defined anatomical regions. Sub-volumes of these adaptively selected regions are then defined using a forward selection algorithm based on SVM discriminability, such that beginning with a single voxel, voxels are iteratively added to the regions as they improve the discriminability. The classification was performed on the two datasets, both comprised of normals and schizophrenics, the first group were females and the second males. The best average classification rate for males and females was 91.8% and 90.8% respectively. The method provides an adaptive framework for grouping and isolating voxel-based discriminative regions; this results in an efficient dimensionality reduction and good discriminative performance. In this paper, they do not discuss the meaning

of the spatial patterns, in particular whether the grouped regions corresponded to anatomical/functional regions. In our work we focus on the use of subcortical regions where we are interested in a hypotheses about the particular structure. For example, to pose the question does the hippocampus discriminate well between controls and AD patients. We also aim to extend the question to ask whether a particular region or shape characteristics of a particular structure discriminates well.

Yoon et al. [61] apply discriminant analysis to the cortical surface in order to discriminate between controls and schizophrenics. The discrimination was performed using surface thickness that was defined by the Euclidean distance from vertices on the inner cortical grey matter surface to the nearest vertex on the outer cortical grey matter surface. SVM is applied to the feature vectors after a PCA dimensionality reduction is performed on each lobe (as defined anatomically). Discrimination accuracy was reported have been improved when only using a subset of the feature vectors. The optimal features did not correspond to the principal component of highest variance, rather, a univariate (between-subject) t-test was used for feature selection. The maximum prediction accuracy (between normals and schizophrenics) when truncating based on variance was 77%, when using the p-value for feature selection, a maximum prediction accuracy of 100% was achieved. There was no variance in the metrics reported, but more importantly the feature selection appeared to be performed outside the LOO process and hence introduces bias. By "outside the LOO process", we are meaning that the t-test was performed using the entire dataset; hence including the test data in the feature selection criteria (which is part of the discriminant method). The feature selection ought to be re-evaluated for each LOO training set.

Duchesnay et al. [22] describes the use of automatically identified sulci for discrim-

inating between gender from a database of 143 subjects. 116 sulci with 27 shape features are marked for each brain. They propose a means of feature selection to balance the trade off between high and low dimensional analysis. An initial univariate feature selection (t-test) is used to reduce the feature space to around 100-200 features. A suboptimal greedy search algorithm is used to further reduce the feature space. As an objective function it averages across five cross-validation runs, where the model was constructed from a random sampling of 80% of the data and is tested on the remaining. Finally, features that are common to a sulci are combined into a single feature. The new feature is a linear combination of the actual features and is calculated using LDA so that it weights the most discriminant features. Discrimination is then performed using LDA, a one layer hidden perceptron, and an SVM. Using feature selection and the linear projections, the SVM achieved a prediction accuracy of 95.8% (93.7% using LDA). When the linear projection step was omitted, SVM and LDA gave a prediction accuracy of 75.5% and 91.6% respectively. The linear projection scheme provides a dramatic increase in performance for the SVM but only a slight increase for LDA; this is a prime example of LDA's robustness to low sampling densities. Although in the optimal paradigm the SVM outperforms LDA, LDA was more robust and still provided good prediction accuracy; this is our motivation for employing LDA and QDA in this chapter. The method provides a means of utilizing the multivariate patterns from an initial high dimensional feature space. The largest drawback with performing discrimination on a large number of features is the problem that interpretation of the discriminant patterns may be difficult; although the large number of features was required to achieve the high prediction accuracy.

4.6 Experimental Methods

We are investigating the use of vertex-wise statistics and discriminant analysis to discern group differences in subcortical structures. Volume and total surface area will be included in our analysis; however, of more interest is the use of the shape information that is inherently contained in the mesh modelling. In this section, we will outline the to which the methods are applied, the choice of discriminant features (metrics) and the discriminant methods to be used in this chapter. Along with the choice of metric, the registration of the surfaces to a standard space (spatial normalization) is discussed.

4.6.1 Test Data

The shape-and-appearance models are fit to two independent clinical datasets; we then apply the size and shape analysis to the fit models. The first group is composed of 19 elderly controls and 39 AD patients. The subjects were scanned at three time-points, where the second and third time-points were performed 6 and 12 months following the first. The T_1-weighted image resolution was $0.9375mm \times 1.5mm \times 0.9375mm$. A single time-point is used for the analysis in this chapter, which was the third since the disease should have progressed the most and the shape effects be the most pronounced for this later time point. AD is structurally characterized by the loss of grey matter, particularly atrophy in the hippocampus at the subcortical level [48, 27, 34].

The second dataset is a group of adult schizophrenia patients and controls; there are 78 subjects in total, 47 controls and 31 patients. All the acquisitions are $1mm$ isotropic resolution T_1-weighted images. Enlarged ventricles and reduced volume in the hippocampus are typically attributed to the pathology [44]. We expect that the structural differences in the subcortical structures in a cross-sectional comparison of schizophrenia versus normals ought to be more subtle than for elderly AD patients versus controls.

4.6.2 Structural Segmentation

The discriminant features used are metrics derived from the surfaces that result from the fitting of the models proposed in the previous chapter to new image data. The models were applied to all subjects from both datasets. The two-stage linear registration was performed to align all the images to the MNI152 template. The brain stem model as well as the left and right model for the nucleus accumbens, amygdala, caudate, hippocampus, pallidum, putamen, thalamus, and lateral ventricles were all independently fit to each subject. Table 4.1 includes the number of modes of variation used for each structure; the number of modes was chosen based on the LOO results contained in appendix E.

Three outputs will be used for analysis: 1) the mode parameters (the parameters that were optimized), 2) the surface mesh (vertex coordinates), and 3) the image representation (of the segmentation). Given the model and linear transformation between the native and MNI152 space, the surface mesh is constructed from the mode parameters which in turn is filled to generate the image representation.

Structure	Number of Modes
L/R Accumbens	50
L/R Amygdala	50
L/R Putamen	40
L/R Pallidum	40
L/R Thalamus	40
L/R Hippocampus	30
L/R Caudate	30
L/R Lat. Ventricles	40
Brain Stem	40

Table 4.1: Number of modes of variation retained for each model when fit to image data.

4.6.3 Statistical Analysis

A multivariate multiple regression model is applied to each metric. In the univariate case the t-statistic is used; in the multivariate case the F-statistic approximation to Pillai's trace is used. For summary measures, as well as the maxima for the local metrics, the statistics may be easily visualized graphically. However, in order to be able to interpret the local difference in discrimination, the local metrics such as vertex coordinates are visualized on a 3D surface. In addition to the statistic, we display the scaled mean vertex-displacement using vectors plotted on the surface.

4.6.4 Choice of Discriminants and Assessment of Prediction Error

Given that several measures, over several structures, over two datasets are being used, we restrict the number of discriminants that are used to two. LDA and QDA were chosen for their robustness and ease of interpretability. Interpretability is important

since we are interested in understanding how shape changes with a disease. LDA and QDA both model the feature space of each class as a Gaussian distribution. LDA and QDA are usually robust and are typically not as susceptible to overfitting as are non-linear methods such as SVMs or RVMs. For subcortical volumes, our experience is that RVMs do not typically outperform LDA and/or QDA, and if so, only slightly.

To estimate and compare prediction errors, the expectation and variance of the prediction was computed using bootstrap cross-validation. The bootstrap sample size was 100 and was repeated 20 times so as to estimate the variance. For the schizophrenia versus controls comparisons, the bootstrap/cross-validation will be commented on, however the LOO results will be primarily given. LOO results were presented because the bootstrap/cross-validation results were believed to have a greater sample-size bias.

4.6.5 Spatial Normalization and Metrics for Size and Shape

Metrics may either be local or global in nature. A local metric measures a feature within a certain region of interest such that multiple measurements are made per structure (e.g. vertex coordinates). A global metric is a single summary measure describing the overall size and/or shape of the structure (e.g. volume). Summary measures provide insight into the gross structural variation, however, do not provide localized information such as variation in particular sub-nuclei. The global metrics are typically pose invariant and require correspondence at a larger scale such as on a per structure basis, for example, the volume of the left putamen of a subject corresponds to the volume of the left putamen of another subject. Comparisons

using local metrics typically require dense-correspondence which tends to be more difficult to establish. When considering metrics such as spatial location, pose becomes important. In addition to the choice of metric, we will discuss the choice of spatial normalization.

The Metrics

A practical implication to be considered when considering a metric is that it dictates the dimensionality of the feature space in question. The dimensionality of the measure(s) will impact the accuracy in the estimation of the model parameters and hence the sensitivity of the shape statistics and performance/generalizability of the discriminant.

Five metrics are used herein to investigate group differences in size-and-shape of structures. The metrics used are: 1) the mode parameters, 2) total volume, 3) total surface area, 4) vertex location and 5) local surface area (of the triangular faces). The metrics chosen are merely a subset of the many that exist - other potential metrics include the RMS distance of the vertices to the mesh centroid (a measure of scale), local curvature, and maximum curvature.

We will now describe the metrics in more detail. The mode parameters of the model are a dimensionality reduction measure; a single parameter reflects shape change described by the corresponding eigenvector. The shape change may in fact reflect global and/or local shape change depending on the eigenvector. Typically the upper few eigenvectors include the global variation. Total volume is a global measure of size

that is derived from the filled and boundary-corrected mesh. Total surface area is measured directly from the mesh and is the summation of the area of each triangular face on the surface. Vertex location is measured as a three-dimensional millimeter Cartesian coordinate and reflects the spatial location in the space of the surface, which is dependent on the spatial normalization. Local surface area is simply the area of single triangular face.

The total volume gives an indication of the size of the structure whereas total surface area is related to size although it also gives an indication of exposed surface. Volumetric and surface area effects may indicate grey matter loss or growth in a patient versus normal population. Vertex location is perhaps the most easily interpretable metric as it reflects local displacement of the vertices in a real world space, interpretation of which may be easily visualized using 3D rendering. Comparisons between the triangular face areas gives an indication of local expansion and contraction of the surface, and should be related to the vertex displacements. For example, if the mean vertex displacement of neighbouring vertices are consistently inwards, it would be expected that there would be a mean reduction in surface area for the triangular faces located in that region.

Spatial Normalization of the Surfaces

In the statistical shape analysis literature a distinction is drawn between shape and size-and-shape. It is the case for both that translation and rotation (pose) are removed prior to analysis. The difference is that scale is removed when considering shape whereas scale is retained for size-and-shape. In neuroanatomy, variation in

global scale of a structure may be of interest as it may correspond to variations in grey matter volume due to pathology. Thus strictly speaking we are usually interested in investigating the variation in size-and-shape of the deep grey structures, although for completeness we will also consider shape. The choice of structural normalization is important and may alter the interpretation of the results. The purpose of normalization is to remove uninteresting variation such as pose and perhaps scale. The removal of pose may also be necessary to satisfy the requirement of the statistical tests (e.g. testing for differences in spatial location of a vertex). Note that even the global (whole-head) normalization scaling may or may not be interesting to keep for inclusion in (e.g.) discrimination testing; there is no general consensus on this question, even in common analysis areas such as VBM and cortical thickness [43]

The models used to generate the output surfaces are native to the MNI152 space. The linear transformation matrix used to align the model into an image's native space provides a means of transforming our surfaces to and from the MNI152 space. The transformation will remove global affine differences between the MNI and native space (with emphasis on subcortical regions). Therefore, the surfaces used by the statistical tests and discriminant methods may be reconstructed in either the MNI or the native space of the image. When reconstructed in MNI space the surfaces have a common reference frame, although local pose differences may still exist. As to whether local residual pose differences are of interest or not is dependent on the question being asked. When reconstructed in the native space, assuming no pre-alignment of the images, the surfaces have no reference frame and therefore must all be aligned to a reference prior to analysis. Different reconstruction methods may or may not be appropriate for all metrics. Furthermore, additional surface alignment may be

necessary, depending on the reconstruction method and metric used. The effect of various combinations of reconstruction and alignment methods will be investigated in this chapter. In particular, for the mesh-based (shape) metrics we use MNI152 space surfaces with no spatial normalization (investigating size, shape, and local pose differences), as well as the native space surfaces with either 6 dof (size-and-shape differences) or 7 dof (shape differences) normalization. For volume and surface area the native space surfaces with no normalization will be used.

After reconstruction of the surfaces in the appropriate space, alignment may be necessary or desired. We restricted the alignment procedure used to either rigid body (six dof) or rigid body with global scaling (seven dof). Given the existence of correspondence between the meshes, the pose (and scale) parameters may be derived directly from the vertices. The target of the alignment is the mean structure as defined by the original shape model. Since the target is in MNI space, all the meshes are aligned to MNI space. The alignment will remove all local pose information (and scale if seven dof is used). Pose parameters are estimated based on a vertex-wise least-squares minimization as outlined by Horn et al. [32]. The transformation of the initial structure, x_{init}, to the target structure, x_{targ}, that minimizes the squared-deviation between the two structures is given by

$$x_{targ} = sR(x_{init} + \Delta) \tag{4.23}$$

where Δ is the translation component (difference in centroid location) as given by,

$$\Delta = \begin{bmatrix} \Delta_x \\ \Delta_y \\ \Delta_z \end{bmatrix} = \begin{bmatrix} \bar{x}_{targ,x} - \bar{x}_{init,x} \\ \bar{x}_{targ,y} - \bar{x}_{init,y} \\ \bar{x}_{targ,z} - \bar{x}_{init,z} \end{bmatrix} \tag{4.24}$$

where $[\bar{x}_{targ,x}\ \bar{x}_{targ,y}\ \bar{x}_{targ,z}]$ and $[\bar{x}_{init,x}\ \bar{x}_{init,y}\ \bar{x}_{init,z}]$ are the centroids of the target and initial structures respectively.

The rotation matrix, R, is given by,

$$R = M\left(M^T M\right)^{-\frac{1}{2}} \tag{4.25}$$

where M is defined as,

$$M = \sum_{i=1}^{N} x'_{init,i}\left(x'_{targ,i}\right)^T \tag{4.26}$$

and $x'_{init,i}$ and $x'_{targ,i}$ are the i^{th} demeaned vertex (represented as a column vector) for the initial and target structures respectively.

Finally the scale parameter, s, may be calculated by the ratio of the RMS distance from the vertices to the centroid for each structure and is given by

$$s = \sqrt{\frac{\sum_{i=1}^{N}\left|x'_{init,i}\right|^2}{\sum_{i=1}^{N}\left|x'_{targ,i}\right|^2}} \tag{4.27}$$

An alternative method of normalization would be to reconstruct the meshes in the MNI152 space. This would provide a shape normalization corresponding to the linear transformation that was determined in the pre-processing stage (two-stage affine

alignment) prior to fitting the models. The transformation was such that its inverse registered the model into the native image space prior to optimization of the model parameters. In fact, the meshes may be reconstructed directly into the MNI152 space using the model and the estimated mode parameters. Applying the original transformation to the mesh in the native image space is equivalent to reconstructing the mesh from the MNI space model. Given a linear, invertible transformation matrix A, it can be easily shown that

$$A\left((A^{-1}\mu) + (A^{-1}U)Db^T\right) = \mu + UDb^T \tag{4.28}$$

where the left hand side of the equation is the transformation of the native space mesh (constructed from the native space model, $A^{-1}\left(\mu + UDb^T\right)$) and the right hand side is the mesh constructed from the MNI space model.

We will now outline the normalization that was used in this chapter, prior to computing each metric. For volume and surface area the native space surfaces with no spatial normalization/alignment was used. The mode parameters are invariant to spatial alignment, and so no alignment is used. For vertex-coordinate and local surface area analysis, the data was analysed using three different normalization techniques. The normalization procedures were: 1) reconstruction in native space with 6 dof alignment to the mean shape in MNI space, 2) reconstruction in native space with 7 dof alignment to the mean shape in MNI space, and 3) reconstruction in MNI152 space.

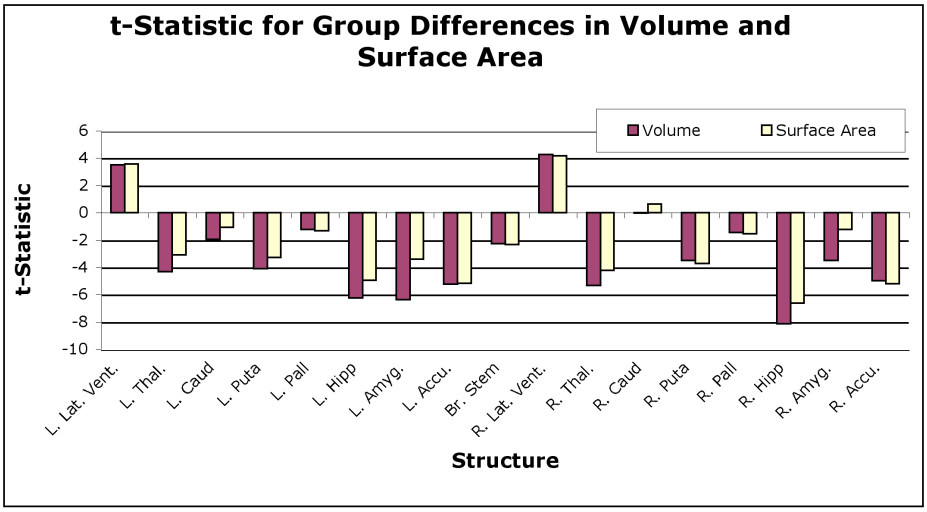
4.7 Results and Discussion

In order to investigate group differences in the overall size of subcortical structures, total volume and surface area are used. Volume and surface area are treated independently and used as a response variable in the GLM. In addition, mode parameters, vertex coordinates, and local triangular face areas are used to investigate shape difference by using the GLM as well as discriminant analysis. The results for AD and schizophrenia will be discussed separately, after which comparisons will be drawn.

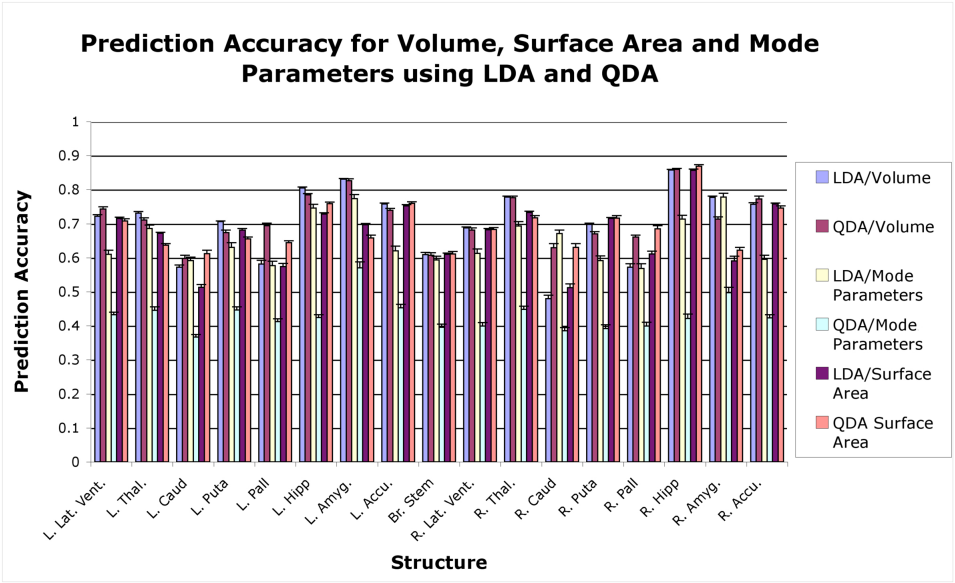
4.7.1 Controls versus AD patients

Since we did not have access to age and gender information for the AD/controls group, the only EV used is group membership. Figure 4.2(a) shows the t-statistics associated with the group difference in volume and surface area for each subcortical structure for the AD dataset.

For the AD/controls group, a two-tail t-test with 56 degrees of freedom and $\alpha = 0.05$ results in a critical t-value of 1.99 (if one is not correcting for multiple comparisons across structures). At $\alpha = 0.05$, a significant reduction in volume for the brain stem as well as a bilateral reduction in the amygdala, hippocampus, and accumbens was found. There was also a bilateral increase in ventricular volume. The same differences were observed for surface area in the same structures (though typically less significant) except for the right amygdala for which there was no significant difference in surface area. This reduction in grey matter, and in particular the hippocampal atrophy was expected for AD [48, 27, 34]. This corresponds well with our finding that the most



(a) t-statistic for volume and surface area between AD and controls. A negative t-value indicates a reduction in volume for AD versus Controls. The critical t-value is 1.99 at $\alpha = 0.05$ (without multiple comparison correction across structures).



(b) Prediction accuracy between AD and controls (using boostrap/cross-validation). Each colour corresponds to a different combination of discriminant and discriminant feature. LDA and QDA indicate the discriminant used.

Figure 4.2: t-statistic and prediction accuracy for volume and surface area.

significant volumetric effect was in the right hippocampus.

Figure 4.2(b) shows the prediction accuracy for the AD/control group for the cases where volume, surface area, and mode parameters are independently used as the discriminant feature. The mode parameters incorporate both size and shape information, however, the dimensionality of the feature space is larger than for volume or surface area. The dimensionality of the feature space is equal to the number of mode parameters that were included in the discriminant, which for our case was the number of modes included in the fitting (see table 4.1). The prediction accuracy of the discriminant appears to correlate with the t-statistic for that structure. This is not surprising since the t-statistic and the discriminant are both related to the separation of the means relative to the group variances.

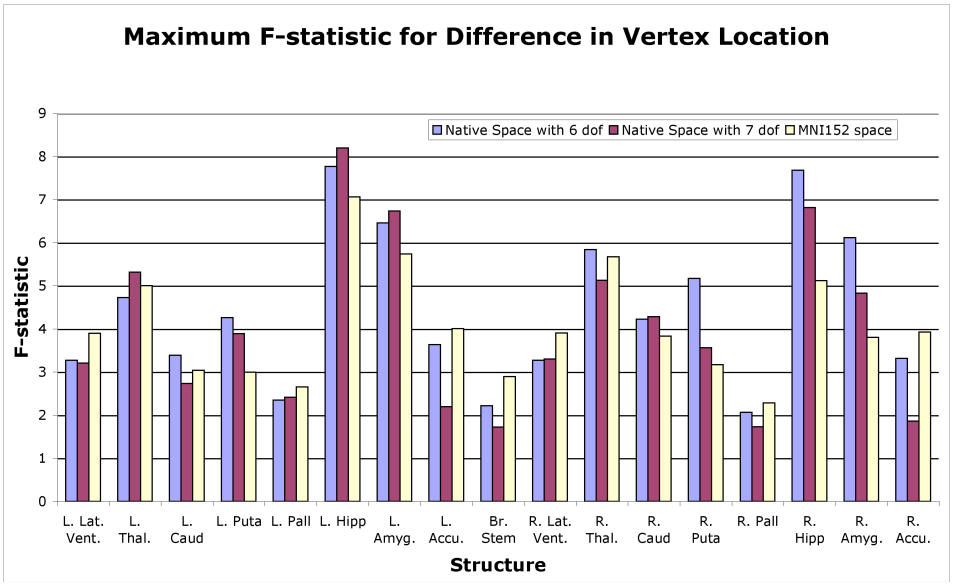
The mode parameters generally do not discriminate well for this data, and this may be related to the higher dimensionality of the space in combination with the reduced effective sample size, which is due to the bootstrap method. It is feasible to make use of a feature selection method to reduce the dimensionality and obtain improved discrimination. A structure's volume typically provided better discrimination than did the surface area, which is consistent with the observation that differences in surface area were less significant than those for volume (figure 4.2(a)). It is worth bearing in mind that the prediction error may, generally, be conservative as a result of sample size bias due to the bootstrap method.

In addition to the size metrics reported, and perhaps of more interest, are localized shape differences. Figure 4.3(a) shows the maximum F-approximation to Pillai's trace that was obtained from applying the GLM to each vertex independently, for each type

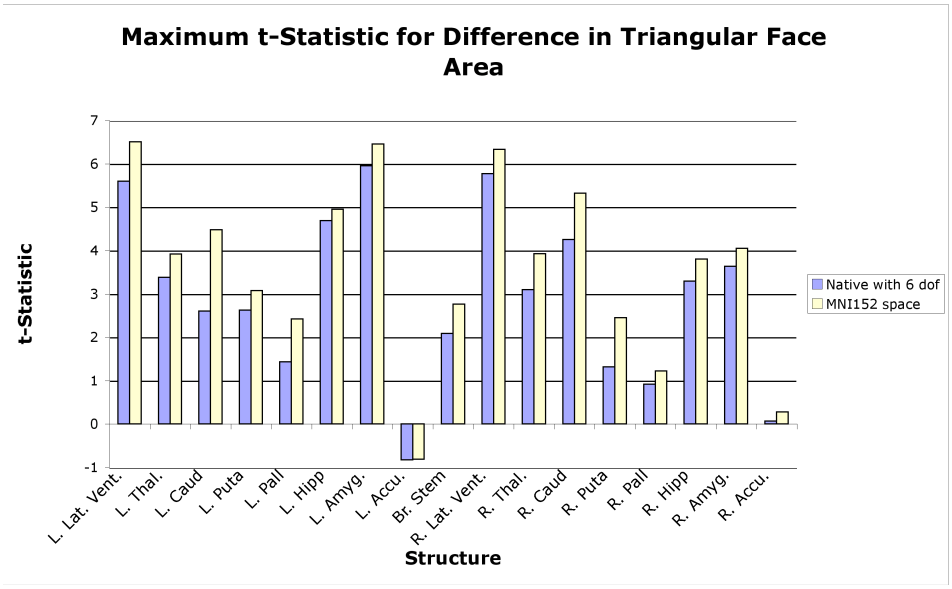
of spatial normalization. In addition, figure 4.3(b) contains the maximum t-statistic obtained from applying the GLM to the area of each triangular face. In addition to the maximum F and t-statistic, figures 4.4(a) and 4.4(b) show the maximum prediction accuracy based on vertex location and triangle area respectively.

As with other measures, the maximum F-approximation correlates with the maximum discriminant performance. The maximum prediction accuracy that was achieved was 0.956 (using the left hippocampus). The maximum discriminant was achieved by applying LDA to the set of surfaces that were reconstructed in native space and aligned to the mean surface (from the model) using 6 degrees of freedom. Given that scale was removed, the high prediction accuracy is indicative that it was in fact a shape difference in the hippocampus that discriminated well between AD and controls. Shape change may correspond to local atrophy as well as geometrical changes. The local discriminants provide significantly better discrimination than for volume and surface area. The difference is suggestive that local regions consistently change in AD, while variation in other regions of the structure may add noise which makes discrimination using global metrics such as volume more difficult.

The maximum statistics and prediction accuracies provided in figures 4.3 and 4.4 give an indication of whether there exists a localized region that differs between groups, however, they do not provide any information about the actual shape differences. The maximum statistic/prediction accuracy is the maximum of the set of statistics/prediction accuracies that were independently calculated for each vertex. To illustrate the shape differences observed in the data, the mean shape for groups are rendered in 3D. In addition, vectors indicating the displacement of the mean vertex location between groups are rendered on the surface. The statistic or prediction

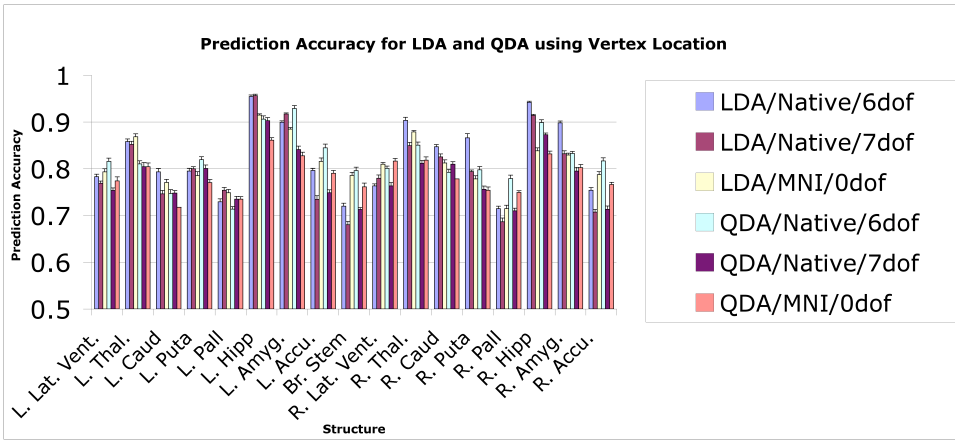


(a) Maximum F-approximation to Pillai's trace for difference in mean vertex location. At $\alpha = 0.05$, with 8 and 110 degrees of freedom, $F_{crit} = 2.03$ (without multiple comparison correction across vertices).

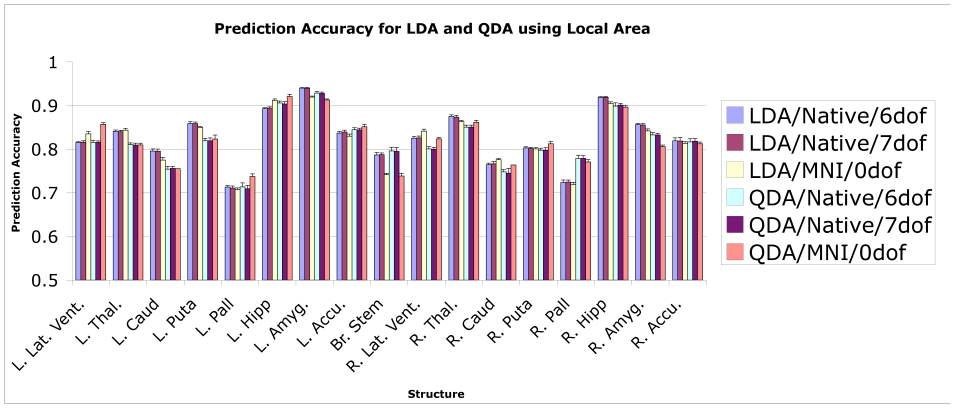


(b) Maximum t-statistic for differences in mean area for an individual triangular patch. At $\alpha = 0.05$, using a two-tailed test, $t_{crit} = 1.99$ (without multiple comparison correction across faces).

Figure 4.3: Maximum of local surface-based statistics. Each colour corresponds to a different type of spatial normalization. Native and MNI152 refer to the meshes reconstructed in the native and MNI152 space respectively. 6 dof indicates that the surfaces were aligned with a rigid body rotation. 7 dof indicates that the surfaces were aligned with a global scale, rotation and translation. No alignment was used for the MNI152 space surfaces.



(a) Maximum prediction accuracy based on a single vertex spatial coordinates.

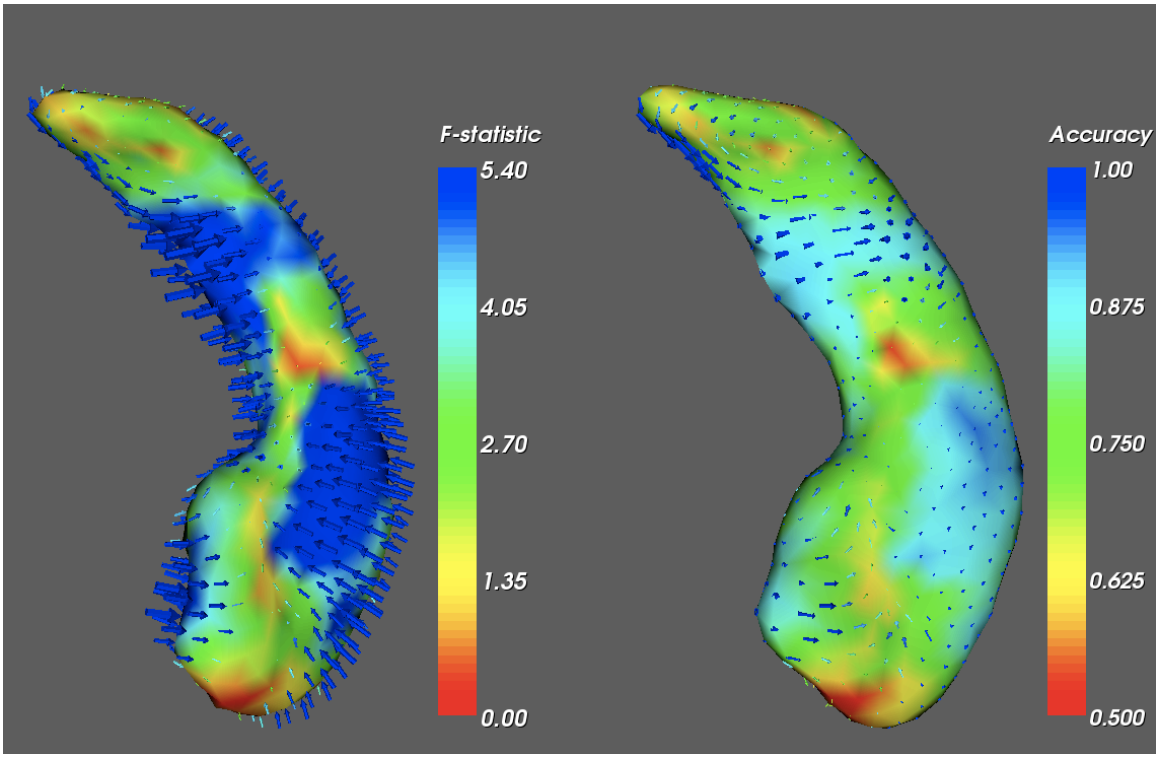


(b) Maximum prediction accuracy based on the area for an individual triangular patch.

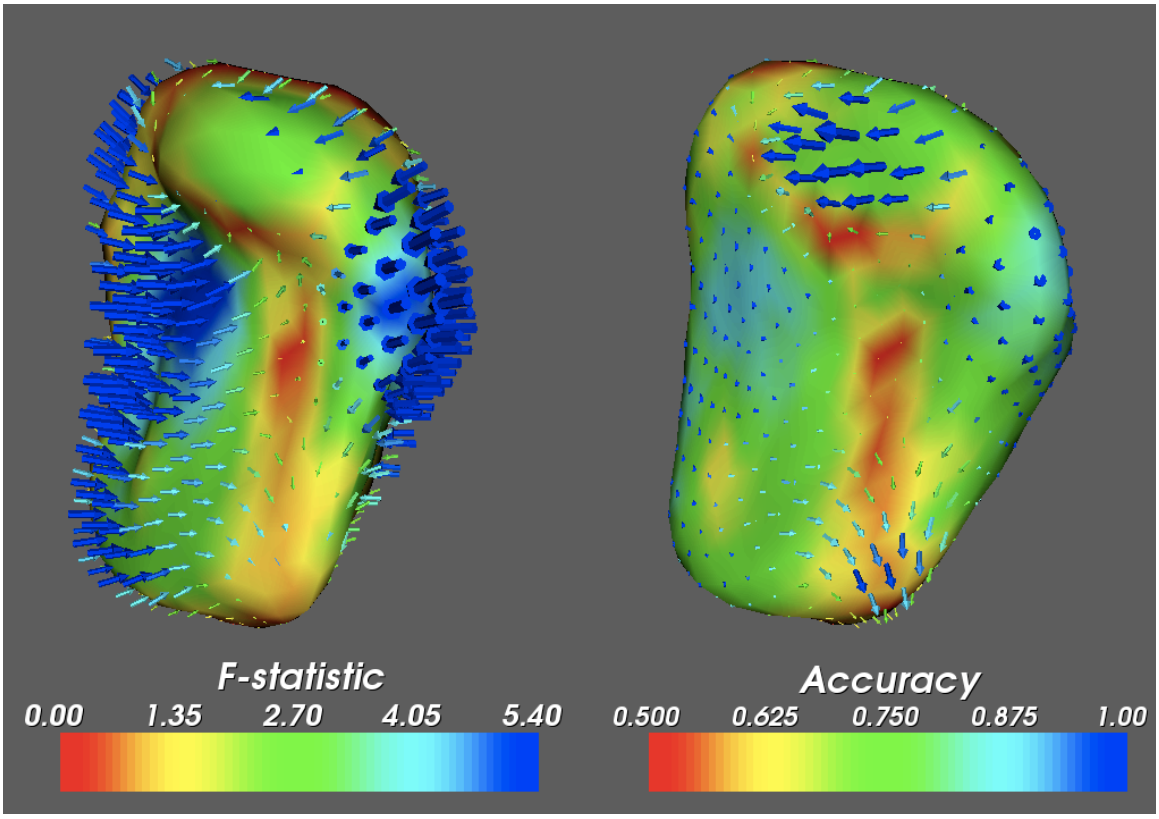
Figure 4.4: Maximum of local surface-based discriminants for AD versus controls (using boostrap/cross-validation). Each colour corresponds to a different combination of discriminant, space of surface reconstruction, and type of spatial normalization. LDA and QDA indicate the discriminant used. Native and MNI refer to the surfaces reconstructed in the native and MNI152 space respectively. 6 dof indicates that the surfaces were aligned with a rigid body rotation. 7 dof indicates that the surfaces were aligned with a global scale, rotation and translation. 0 dof indicates no alignment was used.

accuracy at each vertex is used to colour the surface. The left side of figures 4.5(a), 4.5(b) and 4.5(c) shows the F-approximation map (based on vertex location) for the right hippocampus, thalamus, and amygdala respectively. The right side of the figures show the prediction accuracy map (based on vertex location) for the right hippocampus, thalamus, and amygdala respectively. The surface F-approximations and prediction accuracies depicted were all calculated after surface reconstruction in the native space followed by a six degrees of freedom alignment. The reconstruction and spatial normalization was chosen such that it corresponds to the maximum prediction accuracy for the structure (see figure 4.4). In the right hemisphere, the hippocampus, amygdala and thalamus showed a significant improvement in prediction accuracy between the case where six degrees of freedom is used versus seven degrees of freedom. The seven degrees of freedom normalization removes the size component and isolates shape variation. Therefore, for these structures, the improved discrimination due to the inclusion of size is suggestive that a global scaling is associated with AD as well as shape metrics. In contrast, in the left hippocampus there is no significant difference in discrimination with and without the inclusion of size, which would suggest that AD is associated with shape change in the left hippocampus and not global scaling.

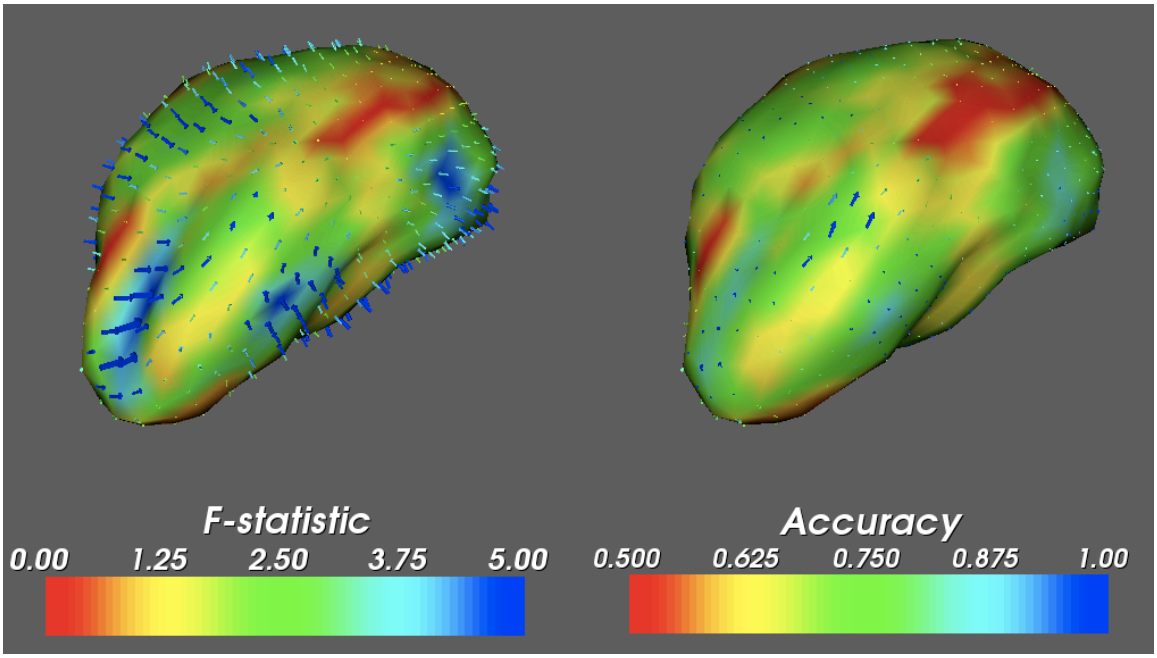
Using either volume or surface area, only the hippocampus and amygdala were able to discriminate with an accuracy above 80%. Using vertex location or local surface area (local contraction/expansion), lateral ventricles, thalamus, putamen, hippocampus and amygdala were all able to achieve a prediction accuracy above 80%. Using individual vertex location, we were also able to significantly improve the prediction accuracy of the amygdala and hippocampus. This suggests that by using our proposed model and fitting method, we may apply vertex-wise analysis to investigate localized



(a) Right Hippocampus



(b) Right Amygdala



(c) Right Thalamus

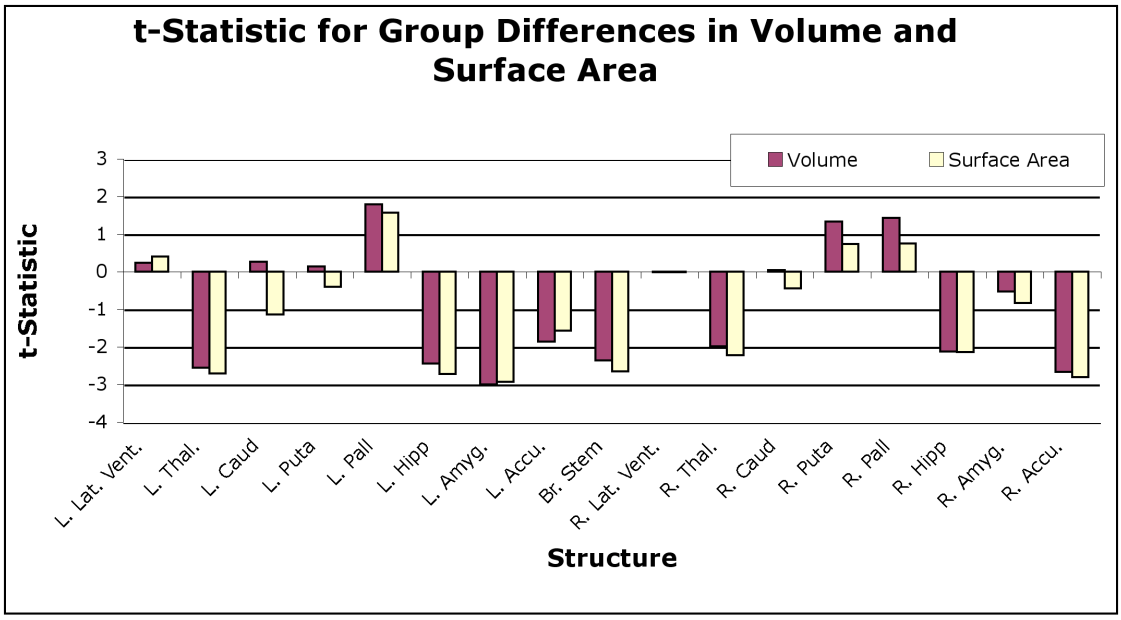
Figure 4.5: F-statistic (left) and prediction accuracy (right) for each vertex based on vertex coordinates (AD versus controls). The left is the mean AD surface, the right is the mean control surface. Arrows are vectors from each mean vertex for the controls to each corresponding mean vertex for AD. Surface colour reflects the F-statistic or prediction accuracy (see colour bar). Arrows are coloured by magnitude (red=0mm, dark blue=2mm).

changes across various regions of the brain that are associated with pathology.

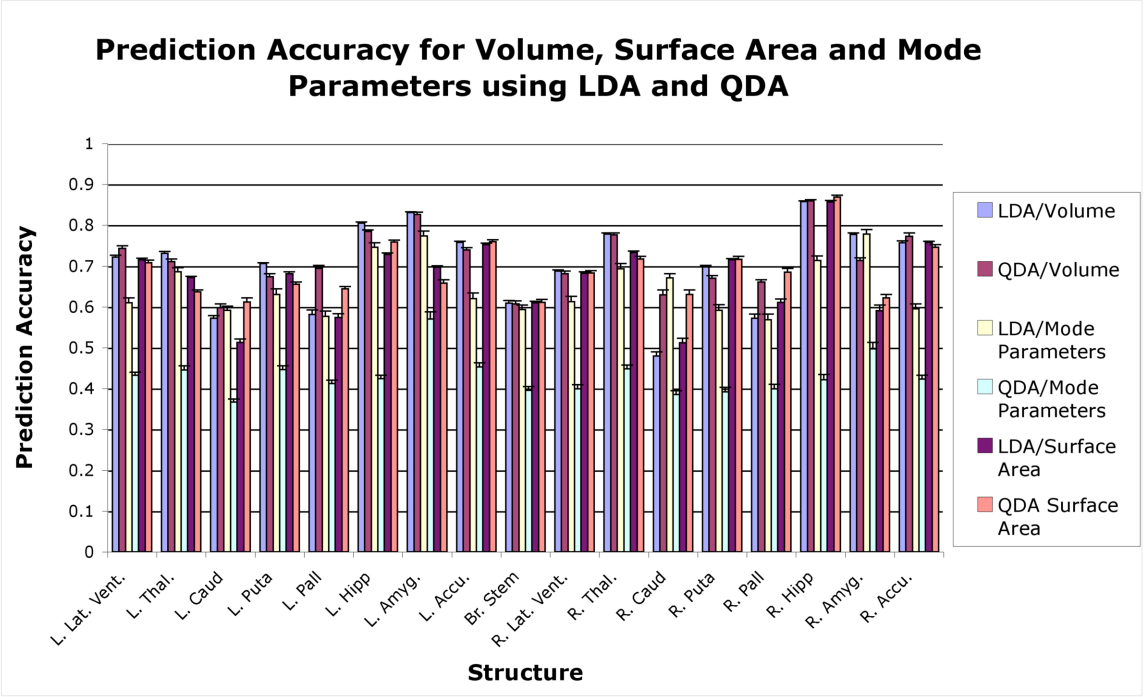
4.7.2 Controls versus Schizophrenia

For the schizophrenia/controls group, in addition to group membership, age and gender were included as confound variables in the GLM. The same set of analysis was used for the schizophrenia dataset as was just presented for the AD dataset. Figure 4.6(a) show the t-statistics associated with the group difference in volume and surface area for each subcortical structure. For the schizophrenia/controls group, a two-tail t-test, with 76 degrees of freedom and $\alpha = 0.05$, results in a critical t-value of 2.00. In the schizophrenia group there was a significant reduction in the left thalamus, hippocampus, and amygdala when compared to controls. In the right hemisphere there is a significant reduction in the volume of the hippocampus and accumbens (the right thalamus is just under threshold). The brain stem also has significant reduction in volume. For all the significant changes in volume there is a corresponding significant change in surface area (in addition, surface area is significant for the right thalamus). Volumetric effects in these structures have been previously reported for schizophrenia [44, 6, 40, 23]. This serves as a confirmation that our automated method has sufficient sensitivity to pick up expected volumetric effects. Perhaps the most intriguing aspect of this finding, from a methodological standpoint, is that the method showed sufficient sensitivity to detect changes in accumbens volume. In T_1-weighted images, even when performed manually, the segmentation of the nucleus accumbens is difficult due to its small size and poor contrast at the putamen and caudate borders.

The bootstrap/cross-validation discriminant analysis did not reach a prediction ac-



(a) t-statistic for volume and surface area between schizophrenia and controls. At $\alpha = 0.05$, using a two-tailed test, $t_{crit} = 2.00$ (without multiple comparison correction across structures).



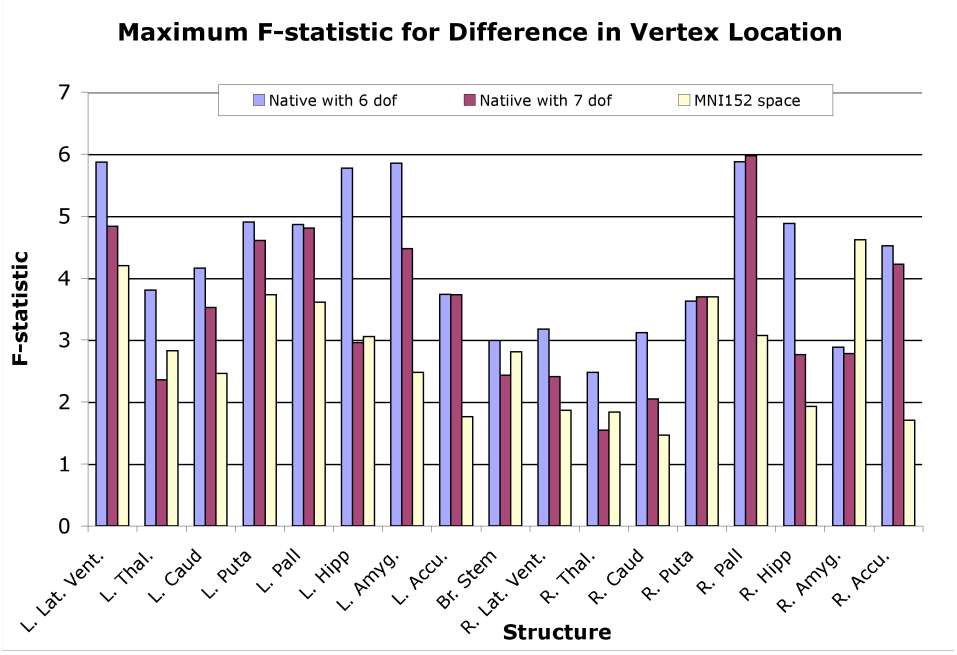
(b) Prediction accuracy between schizophrenia and controls (using leave-one-out cross-validation). Each colour corresponds to a different combination of discriminant, space of surface reconstruction, and type of spatial normalization. LDA and QDA indicate the discriminant used.

Figure 4.6: *t*-statistic and prediction accuracy for volume and surface area.

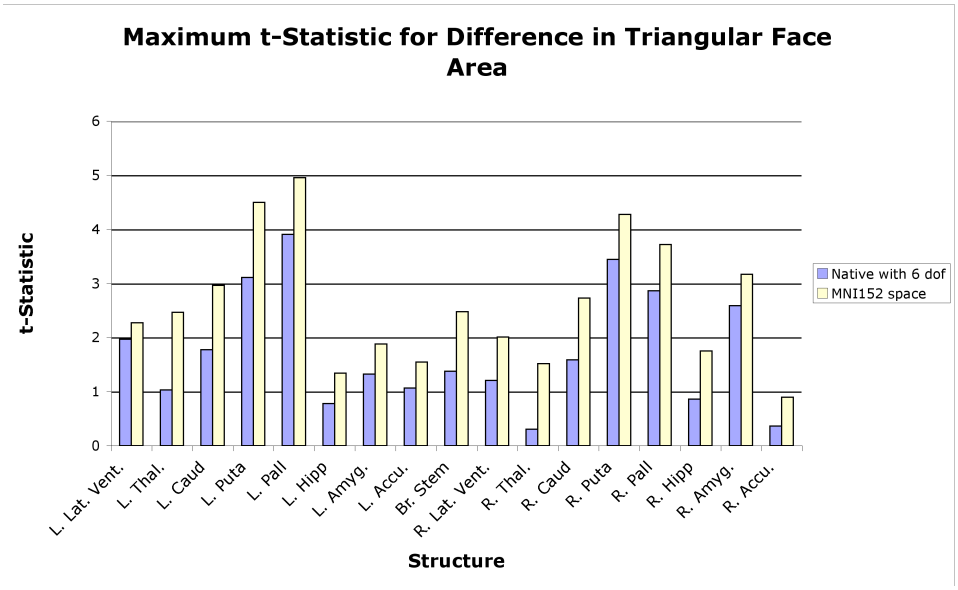
curacy above approximately 67%. Therefore we report LOO results instead of bootstrap results, the poor prediction accuracy from the bootstrap analysis is believed to be attributed to sample size bias. The use of LOO cross-validation has the limitation of not providing an estimate of the variance; this makes it difficult to draw conclusions about differences in prediction accuracy. Figure 4.6(b) shows the prediction accuracy for LDA and QDA when using the volume, surface area and mode parameters using LOO cross-validation. The maximum surface-statistics reflecting the differences between schizophrenia and controls is given in figure 4.7, the maximum prediction accuracy from the discriminant analysis is shown in figure 4.8.

The spatial maps for the surface-statistics of the left accumbens and left amygdala are shown in figure 4.9. The best prediction accuracy for both was for QDA, although six degrees of freedom were used for the alignment of the amygdala whereas seven degrees of freedom were used for the accumbens. Assuming that the results are not due to a random occurrence, the difference in accuracy between six and seven degrees of freedom is suggestive that there is a systematic difference in scale for the left amygdala and not for the left accumbens. The fact that removal of scale improves accuracy for the left accumbens suggest that the scale variability is merely noise.

As with AD, there is consistency between the F-statistic and prediction accuracy, as both indicate that there were local regions of the structure that is associated with the pathology. The best discriminants tend to perform better using QDA rather than LDA, which suggests that shape variance in the patient group is indeed different than for the controls. However, it is difficult to ascertain from the LOO results whether QDA truly does discriminate better or whether it is merely a random occurrence.

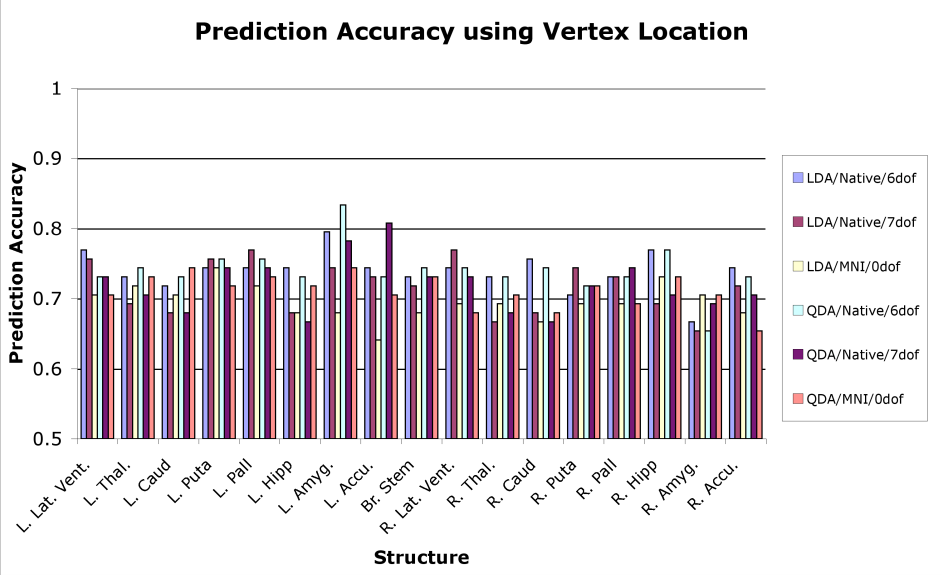


(a) Maximum F-approximation to Pillai's trace for difference in mean vertex location for schizophrenia versus controls. At $\alpha = 0.05$, with 4 and 146 degrees of freedom, $F_{crit} = 2.68$ (without multiple comparison correction across vertices).

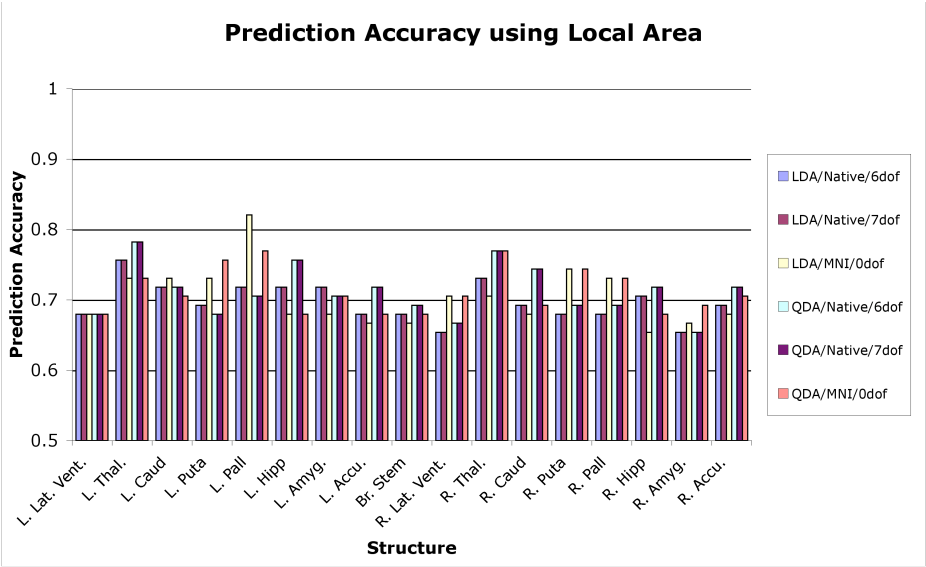


(b) Maximum t-statistic for differences in mean area for an individual triangular patch for schizophrenia versus controls. At $\alpha = 0.05$, using a two-tailed test, $t_{crit} = 2.00$ (without multiple comparison correction across faces).

Figure 4.7: Maximum of local surface-based statistics. Each colour corresponds to a different type of spatial normalization. Native and MNI152 refer to the meshes reconstructed in the native and MNI152 space respectively. 6 dof indicates that the surfaces were aligned with a rigid body rotation. 7 dof indicates that the surfaces were aligned with a global scale, rotation and translation. No alignment was used for the MNI152 space surfaces.

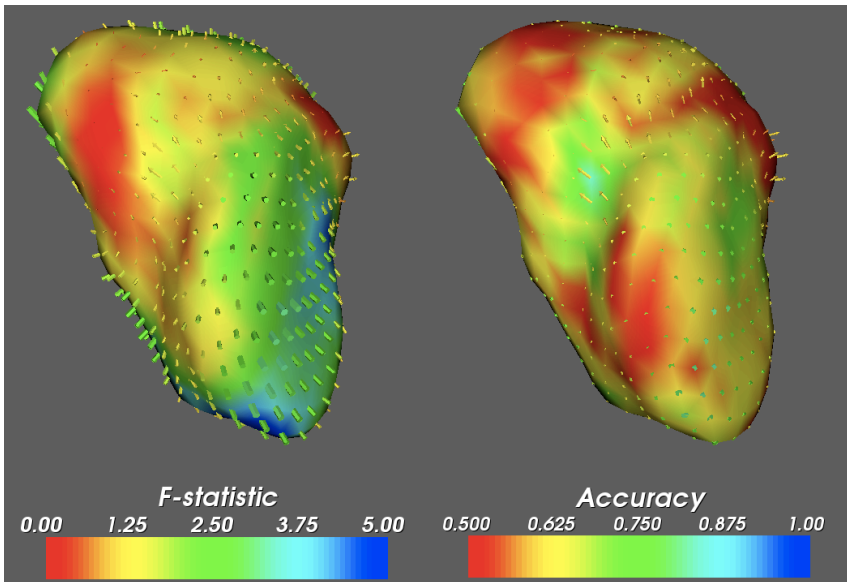


(a) Maximum prediction accuracy based on a single vertex spatial coordinates.

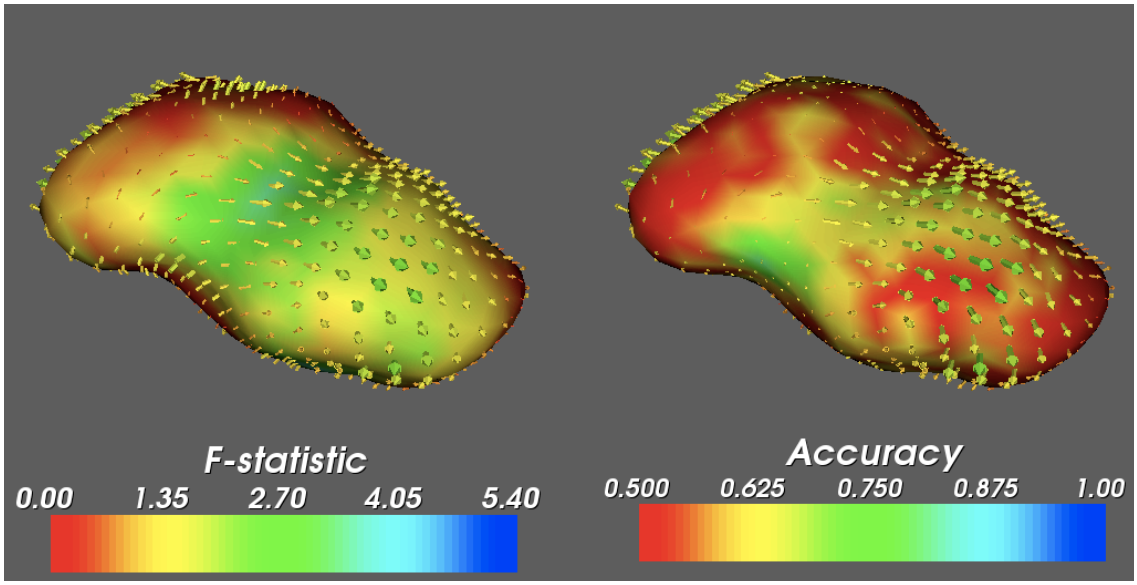


(b) Maximum prediction accuracy based on the area for an individual triangular patch

Figure 4.8: Maximum of local surface-based discriminants for schizophrenia and controls (using leave-one-out cross-validation). Each colour corresponds to a different combination of discriminant, space of surface reconstruction, and type of spatial normalization. LDA and QDA indicate the discriminant used. Native and MNI refer to the surfaces reconstructed in the native and MNI152 space respectively. 6 dof indicates that the surfaces were aligned with a rigid body rotation. 7 dof indicates that the surfaces were aligned with a global scale, rotation and translation. 0 dof indicates no alignment was used.



(a) Right Amygdala



(b) Right Accumbens

Figure 4.9: F-statistic (left) and prediction accuracy (right) for each vertex based on vertex coordinates (schizophrenia versus controls). The left is the mean schizophrenia surface, the right is the mean control surface. Arrows are vectors from each mean vertex for the controls to each corresponding mean vertex for schizophrenia (size is scaled by 3 for visualization purposes). Surface colour reflects the F-statistic or prediction accuracy (see colour bar). Arrows are coloured by magnitude (red=0mm, dark blue=2mm).

4.8 Conclusions

Investigation of differences in volume and surface area provide evidence for global changes in size and shape. However, it does not indicate that the change is an isotropic global scaling. By making use of the rich set of shape information provided by the surfaces we are able to investigate local group differences in shape. Furthermore, by acting directly on vertex coordinates the group differences become easily interpretable.

We examined group differences using classical univariate and multivariate inference techniques (t and F statistics) as well as using discriminant analysis. Because discriminant analysis uses prediction error based on real data it is indicative of a systematic difference between groups. By using bootstrap cross-validation we are able to estimate the expectation of the prediction error and its variability, and thus are able to compare prediction accuracies. For the schizophrenia dataset we use LOO cross-validation because of a believed sample size bias, and as a result it makes it difficult to establish whether any difference in prediction accuracy is a true effect or a random occurrence.

Generally, across both datasets, structures which more significantly differed in a feature corresponded to better discriminants for that feature. This is evident when viewing the common spatial pattern of the F-statistic and prediction accuracy (performed vertex-wise). The differences between AD and controls was more significant than for schizophrenia versus controls, and the discriminants also performed better for the AD dataset.

By performing discrimination on the vertices independently, we reduce the dimen-

sionality of the feature space. More importantly, the discriminant is much more interpretable because the feature space is a three-dimensional Cartesian space that may be easily rendered and visualized. This reduction in dimensionality comes at the price of discarding inter-vertex correlation that may aid in discrimination. Feature selection methods applied to a larger set of vertices may possibly be used to improve prediction accuracy.

Surface area typically gave results which were inferior to those given by volume, nor did it contribute much additional information beyond what the volume already provided. Generally the mode parameters did not discriminate well, because of the larger number of dimensions. Therefore, the use of a feature selection/dimensionality reduction process may have improved their accuracy. The local shape metrics (vertex coordinates and triangle face area) consistently provided improved prediction accuracy. The fact that local metrics improved discrimination is suggestive that for AD and schizophrenia, the changes associated with disease are much more complex than a mere uniform size reduction.

A consequence of using vertex-wise tests, is that we are now performing several hundred independent tests for a single structure. This is problematic when making a decision as to whether or not there is a significant group difference. For a single test, at a significance level α, its repeated application increases the probability that an effect was due purely to chance; α will no longer be a true reflection of the false positive rate. Bonferroni correction [31] is one of the most common methods to account for multiple comparisons, however, it is too conservative since it does not account for the correlation between vertices (which is clearly present). In this chapter, we have only interpreted the spatial maps of the statistic to investigate localized shape differences;

multiple comparisons becomes a problem only when converting the statistic to a p-value.

The surface maps for statistics and prediction accuracy depict localized shape changes, where the regions most associated with the disease are easily visible. The F-statistic and prediction accuracy maps tend to correspond well with each other, and the values appear to be correlated across neighbouring vertices. The patches that have high prediction accuracy or F values are regions that are believed to affected by pathology.

When interpreting shape differences, in particular with regard to discrimination, it is important to avoid any systematic errors in segmentation that are correlated to pathology. A systematic error in segmentation may still prove to be a good diagnostic tool, however, it provides little information about how actual shape changes are related to pathology.

Finally, the fact that vertex-wise discriminants were able to provide better discrimination than for volume, favours the idea that our parameterization method from chapter 3 did indeed preserve point correspondence. The other feature that provides a compelling argument towards the retention of correspondence is that the mean vertex displacement vectors were mostly normal (or close to) the surface, implying that there was little within-surface motion.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

Beginning with a set of manually segmented images, we have provided a method for surface parameterization, for combined surface and intensity modelling, and for the fitting of the models to new data. In addition, we presented methods to investigate shape change given the vertex-correspondence property of the fitted models.

5.1.1 Mesh parameterization using deformable models

The deformable model provided an accurate and robust means for parameterizing labelled images. When applied to the parameterization of manual labels, many of the challenges associated with fitting deformable models to imaging data are alle-

viated because of the presence of ground truth. Our aim was to produce a highly accurate surface parameterization that achieves correspondence across subjects. To ensure stability and robustness, smoothing and tangential forces are typically used. Unfortunately, the smoothing force eliminates the fine detail, and excessive use of the tangential force will violate our point correspondence criterion (due to unwarranted within-surface motion). A novel surface force was introduced that eliminates the need for surface smoothing and that brings stability to the mesh such that the use of the tangential force may be kept to a minimum. The force acts in favour of equal area triangles in a local neighbourhood. The tangential force is needed at times to help aid with large deformations, particularly in areas with large amounts of contraction where the vertices need to be spread out along the surface to accommodate the contraction and prevent mesh self-intersection.

In practice, for the weightings used, the within-surface motion that is introduced by the tangential force is small and localized. In order to ensure minimal within-surface motion, beginning at zero, an iterative process is used to increment the weighting of the tangential force whenever self-intersection of the mesh occurs; the deformation process is reset for each increment. The result is a highly flexible deformable model which is capable of generating surface parameterizations from a set of label images such that correspondence is preserved across subjects.

There are limitations of the deformable model, since, despite the large amount of flexibility, their application to geometrically complex structures may be difficult. For example, the deformable model had difficulty coping with the fourth ventricle passing directly through the brainstem, the effect of which is a hole through the brainstem. This is not easily parameterized using a deformable model, and so the problem was

circumvented by combining the fourth ventricle with the brainstem. Another situation that poses a challenge to the deformable model is long, thin regions with highly variable length. This was the case for the posterior horns of the ventricles. The horns have a large variability in length, mostly due to partial volume effects. The observed change in length is not a simple reflection of true gross anatomical change but rather is a consequence of the manual segmentation protocol as applied to thin horns where the CSF does not occupy the majority of a single voxel.

Another limitation of the deformable model is that it does not accommodate changes in topology. Again, using the posterior horns as an example, at times the partial volume effects cause a disconnect in the manual labels for the horns; the deformable model cannot split into two, to accurately model this effect. With the exception of these special cases mentioned above, the limitations described are not an issue for the subcortical structures being modelled. The subcortical structures generally have a reasonably consistent shape and do not change in topology across the population. Consequently, not allowing changes in topology should generally seen actually be seen as an advantage.

Results were presented in chapter 2 demonstrating the accuracy of the deformable model. The good accuracy is not surprising given that they are deformed based on the underlying truth. Of more interest is the validity of the point correspondence, which is difficult to assess. The issue of correspondence is addressed by examining the performance of the models. It is assumed that a lack of correspondence leads to inaccurate models, and that, in turn, would lead to poor fitting and insensitive statistical tests of group differences (particularly using vertex-wise statistics). Not only would the absence of correspondence confound the shape model but also the es-

timated relationship between shape and intensity. The overlaps presented in chapter 3 for the LOO cross-validation would therefore suggest that the correspondence is indeed preserved. Furthermore, qualitatively, the modes of variation show displacement perpendicular to the surface. If correspondence was not preserved, one would expect the vector components of the eigenvectors to point in a direction along the surface (reflecting the within-surface motion). In addition, we were able to find group differences in the clinical datasets that were presented in chapter 4. These are all suggestive of successful correspondence between vertices.

To apply deformable models, tuning parameters are required to balance the contribution between image forces and shape forces. This is not a large drawback in our application since the parameterization need only be performed once. The surface parameterizations are used to train a shape-and-appearance model which will be fit to new images; we do not use the deformable models presented in chapter 2 to segment intensity data. Most of the existing segmentation methods that incorporate shape models still require tuning parameters to balance shape constraints with the driving image forces. By using our probabilistic approach, the contribution between shape constraints and intensity information is implicit in the derived analytic expression; this eliminates the need for tuning parameters to balance the contribution between shape and intensity (although for the underdetermined case we do require a prior).

5.1.2 Bayesian Shape and Appearance Models

Active Appearance Models (AAMs) do not explicitly account for the lack of training data - they use an empirical estimate to relate shape to intensity, and do not consider

a predicted intensity covariance matrix given a shape deformation when fitting. We solve the highlighted problems of the AAM by posing the appearance model in our novel Bayesian framework. The proposed Bayesian framework models data from a finite training set with an underlying multivariate Gaussian distribution. To cope with small sample sizes relative to dimensionality, a regularization prior is added to the sample covariance matrix. The scaled identity prior models our belief that there exists more variance in the true population beyond that represented in our training set. By so conditioning the matrix, we allow for the calculation of conditional distributions. The appearance model is considered as the conditional distribution of intensity given shape. Its analytic form takes into account the scaling between shape and intensity, hence eliminating any empirical weighting to represent this relationship. An arbitrary weighting does, however, appear in our prior (this is discussed in the next paragraph). Furthermore, by modelling the appearance model as a conditional distribution, the conditional covariance weights the intensity samples by their uncertainty. The framework facilitates the calculation of conditional distributions across different partitions of the data; we have applied the conditionals to shape and intensity, however it can generalize to other categories of data, such as age or gender.

The disadvantage of the framework is the arbitrary choice of the prior ϵ^2; though it has been shown that it has very little impact on the overall fitting (see figure 3.5). Furthermore, ϵ^2 has some real interpretability as it represents shape or intensity variance. The addition of ϵ^2 provides a means by which we can generalize our models to a larger population, rather than limiting the model to the sampled population.

The difference between ϵ^2 and arbitrary weighting parameters used to describe the relationship between shape and intensity is subtle, but is fundamental. The prior

($\epsilon^2\boldsymbol{I}$) incorporates additional information (our prior belief) so as to accurately estimate the shape or intensity space (not to model the relationship between the two), and is only required when the space is undersampled. ϵ^2 does affect the estimated relationship in that the accuracy of the estimated space affects the accuracy of the estimated relationship. In contrast, the AAM parameterizes (estimates) the shape and intensity space separately, and then estimates weighting parameters relating the two. To further illustrate the difference, in the case where the number of samples exceeds the dimensionality of the shape space, our framework no longer requires ϵ^2, whereas the AAM would still need to estimate the relationship between the shape and intensity parameterizations.

From a practical viewpoint, the prior added to the covariance matrix improves the conditioning of the sample covariance, allowing the inverse to be robustly calculated. The inverse is required to evaluate the conditional mean and covariance. By expressing the conditional intensity mean as a function of the shape model's mode parameters, we can evaluate the conditional mean as a linear combination of a set of vectors that represent the relationship between the mean intensity conditioned on shape; otherwise large matrix multiplications are required. As highlighted in the appendices, many of the operations involving the covariance matrices can be simplified so that we work primarily on the scale of $n_s \times n_s$ (n_s is the number of subjects); from a practical point of view this is very important since the dimensionality of our shape models is large.

When fitting, we maximize the posterior of the shape given observed image intensities; this incorporates both shape and intensity priors in addition to the appearance model. Under this formulation, when fitting a structure, it is straightforward to in-

corporate other structures as a prior. In other words, a structure's known shape and location can be used to re-evaluate the shape distribution of another structure. We can, therefore, make use of more robust and accurate structures to inform the less robust. For example, using the known location of the putamen (segmented very reliably), we may re-evaluate the shape distribution of the accumbens (poorly defined image boundaries) based on the putamen; this reduces the search space and could potentially improve accuracy. Furthermore, when maximizing the posterior, there is no arbitrary weightings between the conditional and the learned shape and intensity priors (ϵ^2 is incorporated within the shape and intensity priors). The model utilizes more information from our training data than does the AAM, in that instead of providing a maximum-likelihood estimate of intensity given a shape deformation, the entire conditional distribution is modelled. The benefit of modelling the entire distribution (mean and variance versus just mean) is that the relative contribution of the intensity samples are determined by their uncertainty (which also incorporates covariation). For example, if an intensity sample has large variability in the training data, then when fitting, deviations from the mean will not be given much emphasis, as opposed to intensity samples that have low variability in the training data. The AAM will place equal weighting on all samples. The framework is, also, sufficiently general that data other than shape and intensity can be easily incorporated into the model (e.g., age and gender).

In summary, the advantages of the Bayesian appearance model are that it: 1) explicitly accounts for small datasets, solving the problem of having a rank-deficient covariance matrix; 2) has an analytic form for the conditional distribution, eliminating the need for empirical weightings between intensity and shape variance (ϵ^2 is used

for the estimation of the shape and intensity variance separately); 3) can maximise the posterior probability to fit the model; which incorporates shape and intensity priors within the appearance model; 4) extends well to incorporating other shapes as priors, not only providing a predicted most-likely guess but also the predicted covariation (the benefits of which were elaborated in the previous paragraph); and 5) can extend beyond shape and intensity such that other metrics can be incorporated into the model.

5.1.3 Shape analysis

Several metrics were used to investigate shape differences between two populations. By fitting the models outlined in chapter 3 to two independent datasets, we were able to detect significant differences in volume and surface area. The affected areas were consistent with those expected for the pathology. Volume and surface area only measure size of a structure and do not localize changes. We proposed the application of a multivariate statistical test on individual vertex coordinates to investigate local shape change. In doing so, the statistic was rendered on the shape surface, providing a map of where the structure differed significantly between groups. In terms of localized shape change, testing on all vertices simultaneously is more difficult to interpret; data in general becomes more difficult to physically interpret beyond three dimensions. Also, at the dimensionality we are dealing with, the number of samples will rarely exceed the dimensionality of the entire surface ($3 \times n_{verts}$) and hence the covariance estimates will be less accurate. Therefore, considering vertices independently serves as a dimensionality reduction technique.

The results of the statistical test were confirmed using the same analysis except using discriminant analysis instead of statistics. LDA and QDA were chosen for robustness and their ease of interpretation. The aim of the discriminant analysis was not purely to achieve the maximum discrimination accuracy possible but rather to help isolate shape changes. As with the statistic surface maps, a prediction accuracy map was generated which helps depict the regions of a structure that are systematically different between groups.

The spatial patterns of relative magnitude coincided well between the statistic and prediction accuracy maps. This was a good confirmation of results, and shows that the proposed models are sufficiently sensitive to extract meaningful localized structural changes from T_1-weighted images. The sensitivity to localized structural changes also argues in favour of the fact that vertex correspondence is achieved by the parameterization method proposed in chapter 2.

The localization of individual vertices serves as a dimensionality reduction step. The discriminant analysis is somewhat limited in that it doesn't incorporate any inter-vertex correlation in the discriminant. Feasibly, feature selection methods may be applied to the mesh vertices to determine a subset of vertices to discriminate with, however this may make interpretability more difficult. We will further discuss possibility of feature selection in the following section on future work.

5.2 Future Work

There is still room for future development of parameterization, modelling, and analysis methods, all of which impact each other. The parameterization method is closely tied to the modelling since it defines the data that we are trying to model. Furthermore, changes in the model impact the type of analysis that may be performed on the data.

The current method for parameterization uses a linear transformation to align the data to a common space, the deformable surfaces are then fit to the aligned data. Since we are modelling the residual variation to the normalization procedure, reducing the residual variance allows for easier modelling of the variation and reduces the search space. The most obvious extension of the current method is to use non-linear registration to align the data to a standard space. A potential difficulty is the transformation of the models into the native space. The effect of a global linear transformation on the mutlivariate Student's distribution is well-defined; this is not the case for a non-linear registration. Our framework was derived for an underlying Gaussian model, and if the non-linear transformations violates this assumption, then our framework is no longer valid. An alternative method may be to use a piece-wise linear transformation - following an initial global linear alignment; another linear registration may be performed for each structure based on an ROI surrounding the structure. The piece-wise linear alignment would simplify the transformation of the model to native space, however the robustness and accuracy of such a method remains to be tested.

Currently, structures are modelled independently despite the fact that they may

share boundaries. Thus, for a boundary shared by two structures, the boundary is independently parameterized and modelled twice; it would make more sense to use a single model for the shared boundary. Using a single model for the boundary has the advantage of reducing the total dimensionality of the model for the combined structures. Several possibilities exist for modelling the combination of shared and non-shared boundaries such as this. Perhaps the simplest method would be to perform a PCA on the entire mesh of the combined structures. Although the dimensionality is not quite as large as for the combined structures without shared boundaries, given the challenges that this type of joint model has posed in the past (believed to be because of dimensionality), this method is likely to have similar problems.

In our previous attempts to use joint shape models, they seemed to over-constrain the search space, resulting in poor performance; the search space was restricted to a linear combination of the eigenvectors of the joint distribution. The problem was believed to be one of dimensionality, by combining structures we increased the dimensionality of the distribution, and since the sample size did not increase, our sampling density diminished. For the low sampling density, it is believed to be over-ambitious to attempt to span the joint variation in addition to the individual structural variation by using the eigenvectors of the joint distribution. An alternative approach would be to partition the joint structure into shared boundaries and non-shared boundaries for each structure. This would provide a principled way to partition the surface so as to reduce our dimensionality problem. In fact, depending on the size of the boundary, with 317 training subjects we may have more samples than dimensions for a shape partition. We will further outline three possible methods by which the surfaces may be fit: 1) fit each non-shared boundary independently, then fit the shared boundary

constrained by the previous using the conditional distribution, alternatively, 2) the shared boundary model may be fit followed by the optimization of the non-shared boundary constrained by the shared boundary by using the conditional distribution, and 3) to fit them all simultaneously constrained by the shape prior. The appropriateness of the first two methods is dependent on which boundaries are well-defined by the image. This should be quite predictable based on our knowledge of anatomy and imaging. The third technique is the most appealing since it avoids a hierarchy and hence avoids confounds due to fits conditioned on biassed estimates. As opposed to the third technique, the second uses a hierarchy, which is based on the assumption that previous tiers are correct. Therefore, for hierarchies the results are dependent on the quality of fit of the upper tiers and errors in the fit will feed down erroneous information and bias the results of the lower levels. The third technique, on the other hand, optimizes the two models simultaneously until a joint maximum probability is reached (information feeds back-and-forth).

On the modelling side, improved accuracy and robustness may potentially be achieved through the incorporation of demographic information such as age and gender. It is well known that structural changes occur with age, for example the lateral ventricles expand with age. Given that demographic information is known prior to segmentation, a conditional mean and covariance/eigenvectors may be used given the demographic. For example, instead of initializing the hippocampus using the mean hippocampus for all ages and searching the space representing the variation due to age, gender, pathology as well as variation in the normal population, we may initialize using the mean hippocampus for an eighty year old male and restrict our search space to the variation due to pathology and the normal population that is associated with

eighty year old males.

In chapter 4, we place emphasis on localizing structural change, however the problem of identifying the anatomical significance of that change remains. It may be possible to use an atlas of a structure to identify certain sub-nuclei associated with the affected regions. This is a potentially difficult task that would most likely require expertise in the particular anatomy, which even with such expertise may prove to be difficult to identify. We propose the integration of diffusion tensor imaging (DTI) information with our shape models to create a surface atlas of connectivity. This would effectively be the surface counterpart to the volumetric approach taken by Behrens et al. [3] to segment regions of the thalamus based on connectivity. Given a set of anatomical and DTI scans, by first fitting a structure to the anatomical image and then by using each vertex to seed the tractography, anatomical connectivity may be associated with each vertex. Repeating this for all subjects, we may build up a statistical surface map for the connectivity of a structure, which in turn may be used to help interpret localized shape change. This would also provide a method for assessing anatomical correspondence of the vertices; if the vertices have anatomical correspondence, a vertex should retain the same anatomical connectivity across the population.

In terms of discrimination, the vertex-wise analysis provided information regarding localized shape changes by apply discriminant analysis to vertices independently. We did not make use of any correlation between vertices within or between structures. Exploring the combination of vertices whilst coping with the high dimensionality is an area of future research. A potential method would be to use a feature selection scheme on the entire set of vertices. An example of a feature selection method would be a threshold based on the F-statistic for each vertex. The inclusion of the spatial

correlation of the F-statistic may also enhance the feature selection method.

The methods proposed provide an automated means to extract subcortical, structural information; this allows for the analysis of large datasets. Furthermore, it encourages exploratory investigation of the changes in subcortical structures. Historically, subcortical analysis has relied on manual segmentations, but because this is a time-consuming and labour intensive process, analysis has been usually restricted to a few structures for which the investigators have a hypothesis. There has been a keen interest in utilizing this method for exploring subcortical shape changes in clinical applications, including schizophrenia, bipolar, AD, as well as for exploring developmental changes. Such applications will hopefully yield new insights into the healthy and diseased brain, as well as providing a rich testing ground to further improve this method.

The End.

Bibliography

[1] J. Ashburner and K.J. Friston. MRI sensitivity correction and tissue classification. *NeuroImage*, 7(4):S107, 1998.

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

[3] T.E.J. Behrens, H. Johansen-Berg, M.W. Woolrich, S.M. Smith, C.A.M. Wheeler-Kingshott, P.A. Boulby, G.J. Barker, E.L. Sillery, K. Sheehan, O. Ciccarelli, A.J. Thompson, J.M. Brady, and P.M. Matthews. Non-invasive mapping of connections between human thalamus and cortex using diffusion imaging. *Nature Neuroscience*, 6(7):750–757, 2003.

[4] Rok Bernard, Bostjan Likar, and Franjo Pernus. Segmenting articulated structures by hierarchical statistical modeling of shape, appearance, and topology. In *MICCAI '01: Proceedings of the 4th International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 499–506, London, UK, 2001. Springer-Verlag.

[5] J.M. Bernardo and A.F.M. Smith. *Bayesian Theory*. John Wiley & Sons, Ltd, 2000.

[6] S Berretta, H Pantazopoulos, and N Lange. Neuron numbers and volume of the amygdala in subjects diagnosed with bipolar disorder or schizophrenia. *Biol Psychiatry*, 2007.

[7] Julian Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society*, 48(3):259–302, 1986.

[8] A. D. Brett, A. Hill, and C. J. Taylor. A method of 3d surface correspondence for automated landmark generation. In *In 8th British Machine Vison Conference*, pages 709–718, 1997.

[9] A. D. Brett and C. J. Taylor. A method of automated landmark generation for automated 3d pdm construction, 1998.

[10] D. L. Collins and A. C. Evans. Animal: Validation and applications of nonlinear registration-based segmentation. *Intern. J. Pattern Recognit. Artif. Intell.*, 11(8):1271–1294, 1997.

[11] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models—their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, January 1995.

[12] T.F. Cootes, G.J. Edward, and C.J. Taylor. Active appearance models. In *Proceedings of the 5th European Conference on Computer Vision-Volume II - Volume II*, volume 1407, pages 484–498, 1998.

[13] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[14] D. Cremers, M. Rousson, and R. Deriche. A review of statistical approaches to level set segmentation: integrating color, texture, motion and shape. *International Journal of Computer Vision*, 72(2):195–215, April 2007.

[15] W.R. Crum, O. Camara, and D.L.G. Hill. Generalised overlap measures for evaluation and validation in medical image analysis. *IEEE Transactions on Medical Imaging*, 2006 (In Press).

[16] C. Davatzikos, D. Shen, R.C. Gur, X. Wu, D. Liu, Y. Fan, P. Hughett, B.I. Turetsky, and R.E. Gur. Whole-brain morphometric study of schizophrenia revealing a spatially complex set of focal abnormalities. *Arch Gen Psychiatry*, 62(11):1218–27, November 2005.

[17] Rhodri H. Davies, Carole J. Twining, Tim F. Cootes, and Chris J. Taylor. Automatic construction of optimal statistical shape models.

[18] Rhodri H. Davies, Carole J. Twining, Tim F. Cootes, John C. Waterton, and Chris. J. Taylor. A minimum description length approach to statistical shape modeling. *IEEE Transactions On Medical Imaging*, 21(5):525–537, May 2002.

[19] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[20] I.L. Dryden and K.V. Mardia. *Statistical Shape Analysis*. John Wiley, 1998.

[21] E. Duchesnay, A. Cachia, A. Roche, D. Rivière, Y. Cointepas, D. Papadopoulos-Orfanos, M. Zilbovicius, J.-L. Martinot, and J.-F. Mangin. Classification from cortical folding patterns. *IEEE Trans. Med. Imag.*, 26(4):553–565, 2007.

[22] E. Duchesnay, A. Roche, D. Riviere, D. Papadopoulos-Orfanos, Y. Cointepas, and J. Mangin. Population classification based on structural morphometry of cortical sulci, 2004.

[23] U Ettinger, M Picchioni, S Landau, K Matsumoto, NE van Haren, N Marshall, MH Hall, K Schulze, T Toulopoulou, N Davies, T Ribchester, PK McGuire, and RM Murray. Magnetic resonance imaging of the thalamus and adhesio interthalamica in twins with schizophrenia. *Arch Gen Psychiatry*, 64(4):401–409, April 2007.

[24] Yong Fan, Dinggang Shen, Ruben C. Gur, Raquel E. Gur, and Christos Davatzikos. Compare: Classification of morphological patterns using adaptive regional elements. *IEEE Transactions on Medical Imaging*, 26(1):93–105, January 2007.

[25] Bruce Fischl, David H. Salat, Evelina Busa, Marilyn Albert, Megan Dieterich, Christian Haselgrove, Andre van der Kouwe, Ron Killiany, David Kennedy, Shuna Klaveness, Albert Montillo, Nikos Makris, Bruce Rosen, and Anders M. Dale. Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33:341–355, January 2002.

[26] M Fleute, S Lavallée, and R Julliard. Incorporating a statistically based shape model into a system for computer-assisted anterior cruciate ligament surgery. *Medical Image Analysis*, 3(3):209–22, September 1999.

[27] N. C. Fox, R. I. Scahill, W. R. Crum, and M. N. Rossor. Correlation between rates of brain atrophy and cognitive decline in ad. *Neurology*, 52(8):1533–1534, May 1999.

[28] Alejandro F. Frangi, Daniel Rueckert, Julia A. Schnabel, and Wiro J. Niessen. Automatic construction of multiple-object three-dimensional statistical shape models: Application to cardiac modeling. *IEEE Transactions on Medical Imaging*, 21(9):1151–1166, September 2002.

[29] A. Ghanei, H. Soltanian-Zadeh, A. Ratkewicz, and F.-F. Yin. A three-dimensional deformable model for segmentation of human prostate from ultrasound images. *Medical Physics*, 28:2147–2153, October 2001.

[30] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, 2001.

[31] William L. Hays. *Statistics*. Wadsworth Publishing, fifth edition edition, 1994.

[32] Berthold K.P. Horn. Closed-form solution of absolute orientation using orthonormal matrices. *Journal of the Optical Society of America A: Optics, Image Science, and Vision*, 5(7):1127–1135, July 1988.

[33] M. Jenkinson, P.R. Bannister, J.M. Brady, and S.M. Smith. Improved optimisation for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 17(2):825–841, 2002.

[34] C. R. Jack Jr., R. C. Petersen, Y. Xu, P. C. O'Brien, G. E. Smith, R. J. Ivnik, B. F. Boeve, E. G. Tangalos, , and E. Kokmen. Rates of hippocampal atrophy correlate with change in clinical status in aging and ad. *Neurology*, 55(4):484–490, August 2000.

[35] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, January 1988.

[36] Andras Kelemen, Gabor Szekely, and Guido Gerig. Elastic model-based segmentation of 3-D neuroradiological data sets. *IEEE Transactions on Medical Imaging*, 18(10):828–839, October 1999.

[37] R. Kimmel and J. Sethian. Computing geodesic paths on manifolds, 1998.

[38] Aaron C. W. Kotcheff and Chris J. Taylor. Automatic construction of eigenshape models by direct optimization;. *Medical Image Analysis*, 2(4):303–314, 1998.

[39] Z. Lao, D. Shen, Z. Xue, B. Karacali, S.M. Resnick, and C. Davatzikos. Morphological classification of brains via high-dimensional shape transformations and machine learning methods. *Neuroimage*, 21(1):46–57, January 2004.

[40] MT Ledo-Varela, JM Giménez-Amaya, and A Llamas. The amygdaloid complex and its implication in psychiatric disorders. *An Sist Sanit Navar*, 30(1):61–74, Jan-Apr 2007.

[41] M. Leventon, O. Faugeraus, and W. Grimson. Level set based segmentation with intensity and curvature priors, 2000.

[42] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *SIGGRAPH '87: Proceedings of the 14th annual conference on Computer graphics and interactive techniques*, pages 163–169, New York, NY, USA, 1987. ACM Press.

[43] E Luders, KL Narr, PM Thompson, DE Rex, RP Woods, H Deluca, L Jancke, and AW Toga. Gender effects on cortical thickness and the influence of scaling. *Human Brain Mapping*, 27(4):314–324, April 2006.

[44] C McDonald, N Marshall, PC Sham, ET Bullmore, K Schulze, B Chapple, E Bramon, F Filbey, S Quraishi, M Walshe, and RM Murray. Regional brain morphometry in patients with schizophrenia or bipolar disorder and their unaffected relatives. *Am J Psychiatry*, 163(3):478487, March 2006.

[45] William Menke. *Geophysical Data Analysis: Discrete Inverse Theory*, volume 45 of *International Geophysical Series*. Academic Press, Inc., 1989.

[46] J. Montagnat, H. Delingette, and N. Ayache. A review of deformable surfaces: topology, geometry and deformation;. *Image and Vision Computing*, 19(14):1023–1040, 2001.

[47] John Nolte. *The Human Brain*. Mosby Inc, 1999.

[48] JT O'Brien, S Paling, R Barber, ED Williams, C Ballard, IG McKeith, Gholkar A, WR Crum, MN Rossor, and NC Fox. Progressive brain atrophy on serial mri in dementia with lewy bodies, ad, and vascular dementia. *Neurology*, 56(10):1386–1388, May 2001.

[49] C. L. Olson. On choosing a test statistic in multivariate analysis of variance. *Psychological Bulletin*, 83(4):579–586, July 1976.

[50] Stanley Osher and James A. Sethian. Fronts propagating with curvature-dependent speed: algorithms based on hamilton-jacobi formulations. *J. Comput. Phys.*, 79(1):12–49, 1988.

[51] Alain Pitiot, Herve Delingette, Paul M. Thompson, and Nicholas Ayache. Expert knowledge guided segmentation system for brain mri. *Lecture Notes in Computer Science*, 2879:644–652, 2003.

[52] Kilian M. Pohl, John Fisher, , Ron Kikinis, and William M. Wells. A bayesian model for joint segmentation and registration. *NeuroImage*, 31(1):228–229, May 2006.

[53] Will Schroeder, Ken Martin, and Bill Lorensen. *The Visualization Toolkit An Object-Oriented Approach To 3D Graphics*. Kitware, Inc., 4 edition, 2006.

[54] Dinggang Shen and Christos Davatzikos. Adaptive-focus statistical shape model for segmentation of 3d MR structures. In *MICCAI*, pages 206–215, 2000.

[55] J. G. Sled, A. P. Zijdenbos, and A. C. Evans. A non-parametric method for automatic correction of intensity non-uniformity in mri data. *IEEE Transactions on Medical Imaging*, 17:87–97, February 1998.

[56] S.M. Smith. Fast robust automated brain extraction. *Human Brain Mapping*, 17(3):143–155, November 2002.

[57] Stephen C. Strother, Jon Andersonc, Lars Kai Hansene, Ulrik Kjemse, Rafal Kustra, John Sidtis, Sally Frutiger, Suraj Muley, Stephen LaConte, and David Rottenberg. The quantitative evaluation of functional neuroimaging experiments: The npairs data analysis framework. *Neuroimage*, 15(4):747–771, April 2002.

[58] Michael E. Tipping. The relevance vector machine. In *Advances in Neural Information Processing Systems, San Mateo, CA*, 2000.

[59] A. Tsai, W. Wells, C. Tempany, E. Grimson, and A. Willsky. Mutual information in coupled multi-shape model for medical image segmentation. *Medical Image Analysis*, 2004.

[60] Yongmei Wang, Bradley S. Peterson, and Lawrence H. Staib. Shape-based 3d surface correspondence using geodesics and local geometry. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 644–651, 2000.

[61] Uicheul Yoon, Jong-Min Lee, Kiho Im, Yong-Wook Shin, Baek Hwan Cho, In Young Kim, Jun Soo Kwon, and Sun I. Kim. Pattern classification using principal components of cortical thickness and its discriminative pattern in schizophrenia. *NeuroImage*, 34:1405 – 1415, 2007.

[62] Y. Zhang, M. Brady, and S. Smith. Segmentation of brain mr images through a hidden markov random field model and the expectation maximization algorithm. *IEEE Transactions on Medical Imaging*, 20(1):45–57, 2001.

[63] L. Zhukov, Z. Bao, I. Guskov, J. Wood, and D. Breen. Dynamic deformable models for 3d mri heart segmentation, 2002.

Appendix A

Registration of Statistical Shape Models

In this appendix, we provide the expression for applying transformation matrices to a multivariate Student distribution. In particular, we focus on the case where we are applying a linear transformation to our shape models that are parameterized in terms of their means and eigenvectors.

If $\boldsymbol{x}$ has a multivariate Student distribution, $St_k(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\lambda}, \alpha)$, and $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ such that $\boldsymbol{A}$ is an $m \times k$ matrix of real numbers where $m \leq k$ and $\boldsymbol{A}\boldsymbol{\lambda}^{-1}\boldsymbol{A}^t$ is non-singular, then $\boldsymbol{y}$ has a distribution given by $St_k(\boldsymbol{x}, \boldsymbol{A}\boldsymbol{\mu}, (\boldsymbol{A}\boldsymbol{\lambda}^{-1}\boldsymbol{A}^t)^{-1}, \alpha)$ [5].

Expressing $\boldsymbol{\lambda}$ in terms of its eigenvalues and eigenvectors,

$$\boldsymbol{\lambda} = \boldsymbol{U}\boldsymbol{D}^{-2}\boldsymbol{U}^t \tag{A.1}$$

It follows that the precision matrix for y is given by

$$\begin{aligned}(\boldsymbol{A}\boldsymbol{\lambda}^{-1}\boldsymbol{A}^t)^{-1} &= (\boldsymbol{A}\boldsymbol{U}\boldsymbol{D}^2\boldsymbol{U}^t\boldsymbol{A}^t)^{-1} \\ &= (\boldsymbol{A}\boldsymbol{U})\boldsymbol{D}^{-2}(\boldsymbol{U}\boldsymbol{A})^t\end{aligned} \tag{A.2}$$

$\boldsymbol{AU}$ being the registered eigenvectors; the eigenvalues remain unchanged.

Our shape model has the form of a multivariate Student distribution such that its dimensions correspond to the concatenated x, y, z coordinates of each vertex. Given a 3×3 linear transform matrix, $\boldsymbol{A}_0$, then the transformation matrix $\boldsymbol{A}$ is a $k \times k$ block diagonal matrix of the form

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_0 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{A}_0 \end{bmatrix}, \tag{A.3}$$

such that $\boldsymbol{A}$ is made up of $\frac{k}{3}$ replications of $\boldsymbol{A}_0$ (one for each vertex).

Therefore, applying a linear transformation matrix to a statistical shape model amounts to applying the transformation matrix to each vertex of the mean shape independently as well as to the vectors within each eigenvector independently. When applied to a model of shape and appearance, we assume that the registration only affects the shape distribution, and that the intensity formulation remains unaltered.

Appendix B

Simplification of the Posterior

In this appendix we simplify the posterior probability of shape given intensity to a computationaly convenient form. The expression for the posterior that we wish to maximise is,

$$\ln p(\boldsymbol{x}_s \mid \boldsymbol{x}_I, \mathcal{Z}) = \ln p(\boldsymbol{x}_I \mid \boldsymbol{x}_s, \mathcal{Z}) + \ln p(\boldsymbol{x}_s \mid \mathcal{Z}) \tag{B.1}$$

The full expression for $\ln p(\boldsymbol{x}_I \mid \boldsymbol{x}_s, \mathcal{Z})$ is,

$$\begin{aligned} \ln p(\boldsymbol{x}_I \mid \boldsymbol{x}_s, \mathcal{Z}) = & \ln \left(\frac{\Gamma(\frac{1}{2}(\alpha_{I|s} + k_I))}{\Gamma(\frac{1}{2}\alpha_{I|s})(\alpha_{I|s}\pi)^{k_I/2}} \left|\lambda_{I|s}\right|^{1/2} \right) \\ & + \ln \left(\left[1 + \frac{1}{\alpha_{I|s}} (\boldsymbol{x}_I - \boldsymbol{\mu}_{I|s})^t \lambda_{I|s} (\boldsymbol{x}_I - \boldsymbol{\mu}_{I|s}) \right]^{\frac{-(\alpha_{I|s}+k_I)}{2}} \right) \end{aligned} \tag{B.2}$$

where $\alpha_{I|s}$ is the degree of freedom form the conditional distribution of $\boldsymbol{x}_I$ given $\boldsymbol{x}_s$. Now simplifying the expression,

$$\begin{aligned}
\ln p(\boldsymbol{x}_I \mid \boldsymbol{x}_s, \mathcal{Z}) = &\ln\left(\frac{\Gamma(\frac{1}{2}(\alpha_{I|s}+k_I))}{\Gamma(\frac{1}{2}\alpha_{I|s})(\alpha_{I|s}\pi)^{k_I/2}}\right) + \ln\left(\left|\lambda_{I|s}\right|^{1/2}\right) \\
&- \frac{(\alpha_{I|s}+k_I)}{2}\ln\left(1+\frac{1}{\alpha_{I|s}}(\boldsymbol{x}_I-\boldsymbol{\mu}_{I|s})^t\lambda_{I|s}(\boldsymbol{x}_I-\boldsymbol{\mu}_{I|s})\right) \\
=&C + \frac{1}{2}\ln\left|\lambda_{I|s}\right| - \frac{(\alpha_{I|s}+k_I)}{2}\ln\left(1+\frac{1}{\alpha_{I|s}}(\boldsymbol{x}_I-\boldsymbol{\mu}_{I|s})^t\lambda_{I|s}(\boldsymbol{x}_I-\boldsymbol{\mu}_{I|s})\right) \\
=&C + \frac{1}{2}\ln\left|\lambda_{cII}\left(\frac{\alpha_{I,s}+k_s}{\alpha_{I,s}+\boldsymbol{b}_s^T\boldsymbol{b}_s\gamma_v}\right)\right| \\
&- \frac{(\alpha_{I,s}+k_s+k_I)}{2}\ln\left(1+\frac{1}{\alpha_{I,s}+k_s}(\boldsymbol{x}_I-\boldsymbol{\mu}_{I|s})^t\lambda_{cII}\left(\frac{\alpha_{I,s}+k_s}{\alpha_{I,s}+\boldsymbol{b}_s^T\boldsymbol{b}_s\gamma_v}\right)(\boldsymbol{x}_I-\boldsymbol{\mu}_{I|s})\right) \\
=&C + \frac{k_I}{2}\ln\left(\frac{\alpha_{I,s}+k_s}{\alpha_{I,s}+\boldsymbol{b}_s^T\boldsymbol{b}_s\gamma_v}\right) \\
&- \frac{(\alpha_{I,s}+k_s+k_I)}{2}\ln\left(1+\frac{1}{\alpha_{I,s}+\boldsymbol{b}_s^T\boldsymbol{b}_s\gamma_v}(\boldsymbol{x}_I-\boldsymbol{\mu}_{I|s})^t\lambda_{cII}(\boldsymbol{x}_I-\boldsymbol{\mu}_{I|s})\right)
\end{aligned} \tag{B.3}$$

The full expression for $\ln p(\boldsymbol{x}_s \mid \mathcal{Z})$ is

$$\begin{aligned}
&\ln p(\boldsymbol{x}_s \mid \mathcal{Z}) = \ln\left(\frac{\Gamma(\frac{1}{2}(\alpha_{s+k_s}))}{\Gamma(\frac{1}{2}\alpha_s)(\alpha_s\pi)^{k_s/2}}\left|\lambda_s\right|^{1/2}\right) \\
&+ \ln\left(\left[1+\frac{1}{\alpha_s}(\boldsymbol{x}_I-\boldsymbol{\mu}_s)^t\lambda_s(\boldsymbol{x}_s-\boldsymbol{\mu}_s)\right]^{\frac{-(\alpha_s+k_s)}{2}}\right) \\
&= C - \frac{(\alpha_s+k_s)}{2}\ln\left(1+\frac{1}{\alpha_s}(\boldsymbol{x}_I-\boldsymbol{\mu}_s)^t\lambda_s(\boldsymbol{x}_s-\boldsymbol{\mu}_s)\right) \\
&= C - \frac{(\alpha_s+k_s)}{2}\ln\left(1+\frac{1}{\alpha_s}\boldsymbol{b}_s^T\boldsymbol{b}_s\gamma_v\right)
\end{aligned} \tag{B.4}$$

By substituting equations (B.3) and (B.4) into (B.1), we get an expression for the

full posterior given by

$$
\begin{aligned}
=&C + \frac{k_I}{2} \ln \left(\frac{\alpha_{I,s} + k_s}{\alpha_{I,s} + \boldsymbol{b}_s^T \boldsymbol{b}_s \gamma_v} \right) \\
&- \frac{(\alpha_{I,s} + k_s + k_I)}{2} \ln \left(1 + \frac{1}{\alpha_{I,s} + \boldsymbol{b}_s^T \boldsymbol{b}_s \gamma_v} (\boldsymbol{x}_I - \boldsymbol{\mu}_{I|s})^t \lambda_{cII} (\boldsymbol{x}_I - \boldsymbol{\mu}_{I|s}) \right) \\
&+ \frac{(\alpha_s + k_s)}{2} \ln \left(1 + \frac{1}{\alpha_s} \boldsymbol{b}_s^T \boldsymbol{b}_s \gamma_v \right)
\end{aligned}
\tag{B.5}
$$

Appendix C

Computational Simplifications

Expressing the de-meaned partitions of the training data, $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ in terms of their SVD are given by

$$\boldsymbol{Z}_1 = \boldsymbol{U}_1 \boldsymbol{D}_1 \boldsymbol{V}_1^T \tag{C.1}$$

$$\boldsymbol{Z}_2 = \boldsymbol{U}_2 \boldsymbol{D}_2 \boldsymbol{V}_2^T \tag{C.2}$$

We will now express the partitioned covariance and cross-covariance matrices in terms of (C.1) and (C.2)

$$\begin{aligned} \boldsymbol{\Sigma}_{11} &= \boldsymbol{U}_1 (\boldsymbol{D}_1^2 + 2\epsilon_1^2 \boldsymbol{I})(n_s - 1)^{-1} \boldsymbol{U}_1^T \\ &= \boldsymbol{U}_1 \boldsymbol{D}_{\epsilon_1}^2 \boldsymbol{U}_1^T (n_s - 1)^{-1} \end{aligned} \tag{C.3a}$$

$$\boldsymbol{\Sigma}_{22} = \boldsymbol{U}_2 \boldsymbol{D}_{\epsilon_2}^2 \boldsymbol{U}_2^T (n_s - 1)^{-1} \tag{C.3b}$$

$$\begin{aligned}\boldsymbol{\Sigma}_{12} &= \boldsymbol{\Sigma}_{21}^T \\ &= \boldsymbol{Z}_1\boldsymbol{Z}_2^T(n_s-1)^{-1} \\ &= \boldsymbol{Z}_1\boldsymbol{V}_2\boldsymbol{D}_2^T\boldsymbol{U}_2^T(n_s-1)^{-1}\end{aligned} \tag{C.3c}$$

Rearranging (3.31), such that

$$(\boldsymbol{x}_I - \boldsymbol{\mu}_I) = \boldsymbol{U}_I\frac{\boldsymbol{D}_{\epsilon_I}\sqrt{\gamma_v}}{\sqrt{(n_s-1)}}\boldsymbol{b}_I \tag{C.4}$$

where i is the i^{th} partition.

C.1 Conditional Mean as a Mode of Variation

Substituting (C.3a), (C.3b), (C.3c), and (C.4) into (3.25c),

$$\begin{aligned}\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \boldsymbol{x}_1\boldsymbol{V}_2\big[\,\boldsymbol{D}_{2,s}\;0\,\big]\boldsymbol{U}_2^T\boldsymbol{U}_2\big[\begin{smallmatrix}\boldsymbol{D}_{\epsilon_2,s}^{-2} & 0\\ 0 & \frac{1}{2}\epsilon_2^{-2}\boldsymbol{I}\end{smallmatrix}\big]\boldsymbol{U}_2^T \\ \boldsymbol{U}_2\big[\begin{smallmatrix}\boldsymbol{D}_{\epsilon_2,s} & 0\\ 0 & \sqrt{2}\epsilon_2\boldsymbol{I}\end{smallmatrix}\big]\sqrt{\frac{\gamma_v}{n_s-1}}\boldsymbol{b}_2\end{aligned} \tag{C.5}$$

$$\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \boldsymbol{Z}_1\big[\,(\boldsymbol{V}_2\boldsymbol{D}_{2,s}\boldsymbol{D}_{\epsilon_2,s}^{-1}\sqrt{\tfrac{\gamma_v}{n_s-1}}\boldsymbol{b}_{2,s})+0\,\big]$$

where here $\boldsymbol{b}_{2,s}$ is the upper-left $n_s \times 1$ submatrix of $\boldsymbol{b}_2$.

$$\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \boldsymbol{Z}_1[\boldsymbol{V}_2\boldsymbol{D}_{2,s}\boldsymbol{D}_{\epsilon_2,s}^{-1}\sqrt{\frac{\gamma_v}{n_s-1}}\boldsymbol{b}_{2,s}] \tag{C.6}$$

All matrices within square brackets are of size $n_s \times n_s$ except $\boldsymbol{b}_{2,s}$ which is $n_s \times 1$. If truncating modes at L, only the first L columns of $\boldsymbol{Z}_1[\boldsymbol{V}_2\boldsymbol{D}_{2,s}\boldsymbol{D}_{\epsilon 2,s}^{-1}\sqrt{\frac{\gamma_v}{n_s-1}}]$ are needed.

C.2 Simplifying Conditional Covariance Operations

We will here define

$$\begin{aligned}\boldsymbol{\Sigma}_{1|2} &= \boldsymbol{\Sigma}_{c11}[\frac{\alpha + (\boldsymbol{x}_2 - \bar{\boldsymbol{x}}_2)^T\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \bar{\boldsymbol{x}}_2)}{\alpha + k_2}] \\ &= \boldsymbol{U}_{1|2}\boldsymbol{D}_{1|2}^2\boldsymbol{U}_{1|2}^T\end{aligned} \tag{C.7}$$

where $\boldsymbol{U}_{1|2}$ are the eigenvectors, and $\boldsymbol{D}_{1|2}^2$ is a diagonal matrix of the eigenvalues. For notational convenience we will also define

$$\boldsymbol{\lambda}_{c11}^{-1} = \boldsymbol{\Sigma}_{c11} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \tag{C.8a}$$

such that

$$\boldsymbol{U}_{1|2} = \boldsymbol{U}_{c11} \tag{C.8b}$$

$$\boldsymbol{D}_{1|2}^2 = \boldsymbol{D}_{c11}^2[\frac{\alpha + (\boldsymbol{x}_2 - \bar{\boldsymbol{x}}_2)^T\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \bar{\boldsymbol{x}}_2)}{\alpha + k_2}] \tag{C.8c}$$

C.2.1 Simplifying $(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_{22}^{-1} (\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$

$$\begin{aligned}
&(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_{22}^{-1} (\boldsymbol{x}_2 - \boldsymbol{\mu}_2) = \\
&= \boldsymbol{b}_2^T \sqrt{\frac{\gamma_v}{n_s - 1}} \boldsymbol{D}_{\epsilon_2} \boldsymbol{U}_2^T \boldsymbol{U}_2 \boldsymbol{D}_{\epsilon_2}^{-2} \boldsymbol{U}_2^T (n_s - 1) \boldsymbol{U}_2 \boldsymbol{D}_{\epsilon_2} \sqrt{\frac{\gamma_v}{n_s - 1}} \boldsymbol{b}_2 \\
&= \boldsymbol{b}_2^T \boldsymbol{b}_2 \gamma_v
\end{aligned} \tag{C.9}$$

C.2.2 Simplifying $\boldsymbol{\Sigma}_{c11}$

$$\begin{aligned}
\boldsymbol{\Sigma}_{c11} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{12} \\
&= \boldsymbol{U}_1 \boldsymbol{D}_{\epsilon_1}^2 \boldsymbol{U}_1^T (\frac{1}{n_s - 1}) - (\frac{1}{n_s - 1}) \boldsymbol{U}_1 \boldsymbol{D}_1 \boldsymbol{V}_1^T \\
&\boldsymbol{V}_2 \boldsymbol{D}_2^T \boldsymbol{D}_{\epsilon_2}^{-2} (n_s - 1) \boldsymbol{D}_2 \boldsymbol{V}_2^T \boldsymbol{V}_1 \boldsymbol{D}_1^T \boldsymbol{U}_1^T (\frac{1}{n_s - 1}) \\
&= \boldsymbol{U}_1 (\boldsymbol{D}_{\epsilon_1}^2 - \boldsymbol{D}_1 \boldsymbol{V}_1^T \boldsymbol{V}_2 \boldsymbol{D}_2^T \boldsymbol{D}_{\epsilon_2}^{-2} \boldsymbol{D}_2 \boldsymbol{V}_2^T \boldsymbol{V}_1 \boldsymbol{D}_1^T) \boldsymbol{U}_1^T (\frac{1}{n_s - 1})
\end{aligned} \tag{C.10}$$

We now define

$$\boldsymbol{\Sigma}_{c11} = \boldsymbol{U}_1 \boldsymbol{\Sigma}_{c11V} \boldsymbol{U}_1^T (\frac{1}{n_s - 1}) \tag{C.11}$$

such that,

$$\begin{aligned}
&\boldsymbol{\Sigma}_{c11V} = \boldsymbol{D}_{\epsilon_1}^2 - \boldsymbol{D}_1 \boldsymbol{V}_1^T \boldsymbol{V}_2 \boldsymbol{D}_2^T \boldsymbol{D}_{\epsilon_2}^{-2} \boldsymbol{D}_2 \boldsymbol{V}_2^T \boldsymbol{V}_1 \boldsymbol{D}_1^T \\
&= \left[\begin{smallmatrix} (\boldsymbol{D}_{\epsilon_1,s}^2 + 2\epsilon_1^2 \boldsymbol{I} & 0 \\ 0 & 2\epsilon_1^2 \boldsymbol{I} \end{smallmatrix} \right] \\
&- \left[\begin{smallmatrix} \boldsymbol{D}_{1,s} \\ 0 \end{smallmatrix} \right] \boldsymbol{V}_1^T \boldsymbol{V}_2 \left[\boldsymbol{D}_{2,s} \; 0 \right] \left[\begin{smallmatrix} \boldsymbol{D}_{\epsilon_2,s}^{-2} & 0 \\ 0 & \frac{1}{2}\epsilon_2^{-2} \boldsymbol{I} \end{smallmatrix} \right] \left[\begin{smallmatrix} \boldsymbol{D}_{2,s} \\ 0 \end{smallmatrix} \right] \boldsymbol{V}_2^T \boldsymbol{V}_1 \left[\boldsymbol{D}_{1,s} \; 0 \right] \\
&= \left[\begin{smallmatrix} (\boldsymbol{D}_{1,s}^2 + 2\epsilon_1^2 \boldsymbol{I} - \boldsymbol{D}_{1,s} \boldsymbol{V}_1^T \boldsymbol{V}_2 \boldsymbol{D}_{2,s} \boldsymbol{D}_{\epsilon_2,s}^{-2} \boldsymbol{D}_{2,s} \boldsymbol{V}_2^T \boldsymbol{V}_1 \boldsymbol{D}_{1,s}) & 0 \\ 0 & 2\epsilon_1^2 \boldsymbol{I} \end{smallmatrix} \right]
\end{aligned} \tag{C.12}$$

$$\boldsymbol{\Sigma}_{c11V} = \begin{bmatrix} \boldsymbol{\Sigma}_{c11V,s} & 0 \\ 0 & 2\epsilon_1^2 \boldsymbol{I} \end{bmatrix} \tag{C.13}$$

Express $\boldsymbol{\Sigma}_{c11V,s}$ in terms of its SVD expansion.

$$\boldsymbol{\Sigma}_{c11V,s} = \boldsymbol{U}_{c11V,s} \boldsymbol{D}^2_{c11V,s} \boldsymbol{U}^T_{c11V,s} \tag{C.14}$$

Note that $\boldsymbol{U}_{c11V,s}$ and $\boldsymbol{D}_{c11V,s}$ are $n \times n$ matrices.

It now follows that

$$\begin{aligned} \boldsymbol{\Sigma}_{c11V} &= \boldsymbol{U}_{c11V} \boldsymbol{D}^2_{c11V\epsilon_1} \boldsymbol{U}^T_{c11V} \\ &= \left[\begin{smallmatrix} \boldsymbol{U}_{c11V,s} & 0 \\ 0 & \boldsymbol{I} \end{smallmatrix} \right] \left[\begin{smallmatrix} \boldsymbol{D}^2_{c11V,s} & 0 \\ 0 & 2\epsilon_1^2 \boldsymbol{I} \end{smallmatrix} \right] \left[\begin{smallmatrix} \boldsymbol{U}^T_{c11V,s} & 0 \\ 0 & \boldsymbol{I} \end{smallmatrix} \right] \end{aligned} \tag{C.15}$$

Now substituting the new expression for $\boldsymbol{\Sigma}_{c11V}$ back into $\boldsymbol{\Sigma}_{c11}$, we get

$$\begin{aligned} \boldsymbol{\Sigma}_{c11} &= \boldsymbol{U}_1 \boldsymbol{U}_{c11V} \boldsymbol{D}_{c11V\epsilon_1} \boldsymbol{U}^T_{c11V} \boldsymbol{U}^T_1 \\ &= \boldsymbol{U}_1 \left[\begin{smallmatrix} \boldsymbol{U}_{c11V,s} & 0 \\ 0 & \boldsymbol{I} \end{smallmatrix} \right] \left[\begin{smallmatrix} \boldsymbol{D}^2_{c11V,s} & 0 \\ 0 & 2\epsilon_1^2 \boldsymbol{I} \end{smallmatrix} \right] \left[\begin{smallmatrix} \boldsymbol{U}^T_{c11V,s} & 0 \\ 0 & \boldsymbol{I} \end{smallmatrix} \right] \boldsymbol{U}^T_1 (n-1)^{-1} \end{aligned} \tag{C.16}$$

Define

$$\boldsymbol{U}_1 = \left[\begin{smallmatrix} \boldsymbol{U}_{1,s} & \boldsymbol{U}_{1,s_2} \end{smallmatrix} \right] \tag{C.17}$$

where $\boldsymbol{U}_{1,s}$ and $\boldsymbol{U}_{1,s_2}$ are $k_1 \times n_s$ and $k_1 \times (k_1 - n_s)$ submatrices of $\boldsymbol{U}_1$ respectively.

$$\begin{aligned} \boldsymbol{\Sigma}_{c11} &= \left[\begin{smallmatrix} \boldsymbol{U}_{1,s} & \boldsymbol{U}_{1,s_2} \end{smallmatrix} \right] \left[\begin{smallmatrix} \boldsymbol{U}_{c11V,s} & 0 \\ 0 & \boldsymbol{I} \end{smallmatrix} \right] \left[\begin{smallmatrix} \boldsymbol{D}^2_{c11V,s} & 0 \\ 0 & 2\epsilon_1^2 \boldsymbol{I} \end{smallmatrix} \right] \left[\begin{smallmatrix} \boldsymbol{U}^T_{c11V,s} & 0 \\ 0 & \boldsymbol{I} \end{smallmatrix} \right] \left[\begin{smallmatrix} \boldsymbol{U}_{1,s} & \boldsymbol{U}_{1,s_2} \end{smallmatrix} \right]^T (n-1)^{-1} \\ &= \left[\begin{smallmatrix} \boldsymbol{U}_{1,s} \boldsymbol{U}_{c11V,s} & \boldsymbol{U}_{1,s_2} \end{smallmatrix} \right] \left[\begin{smallmatrix} \boldsymbol{D}^2_{c11V,s} & 0 \\ 0 & 2\epsilon_1^2 \boldsymbol{I} \end{smallmatrix} \right] \left[\begin{smallmatrix} \boldsymbol{U}_{1,s} \boldsymbol{U}_{c11V,s} & \boldsymbol{U}_{1,s_2} \end{smallmatrix} \right]^T (n_s - 1)^{-1} \end{aligned} \tag{C.18}$$

From earlier $\boldsymbol{U}_{1|2} = \boldsymbol{U}_{c11}$

$$\boldsymbol{\Sigma}_{c11} = \left[\, \boldsymbol{U}_{1|2,s} \; \boldsymbol{U}_{1|2,s_2} \,\right] \begin{bmatrix} \boldsymbol{D}^2_{c11V,s} & 0 \\ 0 & 2\epsilon_1^2 \boldsymbol{I} \end{bmatrix} \left[\, \boldsymbol{U}_{1|2,s} \; \boldsymbol{U}_{1|2,s_2} \,\right]^T (n_s - 1)^{-1} \tag{C.19}$$

$$\begin{aligned} \boldsymbol{V}_{1|2} =& \boldsymbol{\Sigma}_{1|2} \frac{(n_s - \frac{1}{n_s} + k_2)}{(n_s - \frac{1}{n_s} + k_2 - 2)} \\ =& \left[\, \boldsymbol{U}_{1|2,s} \; \boldsymbol{U}_{1|2,s_2} \,\right] \begin{bmatrix} \boldsymbol{D}^2_{c11V,s} & 0 \\ 0 & 2\epsilon_1^2 \boldsymbol{I} \end{bmatrix} \left[\, \boldsymbol{U}_{1|2,s} \; \boldsymbol{U}_{1|2,s_2} \,\right]^T \frac{(\alpha + k_2)(\alpha + (\boldsymbol{x}_2 - \bar{x}_2)^T \boldsymbol{\Sigma}_{22}^{-1} (\boldsymbol{x}_2 - \bar{x}_2))}{(\alpha + k_2 - 2)(n_s - 1)(\alpha + k_2)} \\ =& \left[\, \boldsymbol{U}_{1|2,s} \; \boldsymbol{U}_{1|2,s_2} \,\right] \begin{bmatrix} \boldsymbol{D}^2_{c11V,s} & 0 \\ 0 & 2\epsilon_1^2 \boldsymbol{I} \end{bmatrix} \left[\, \boldsymbol{U}_{1|2,s} \; \boldsymbol{U}_{1|2,s_2} \,\right]^T \frac{(\alpha + (\boldsymbol{x}_2 - \bar{\boldsymbol{x}}_2)^T \boldsymbol{\Sigma}_{22}^{-1} (\boldsymbol{x}_2 - \bar{\boldsymbol{x}}_2))}{(\alpha + k_2 - 2)(n_s - 1)} \end{aligned} \tag{C.20}$$

where $\alpha = n_s - \frac{1}{n_s}$

The first n eigenvectors (order by eigenvalues) of the conditional covariance is $\boldsymbol{U}_{1,s}\boldsymbol{U}_{c11V,s}$, which are of dimensions $k_1 \times n_s$ and $n_s \times n_s$ respectively. The eigenvectors greater than n_s are $\boldsymbol{U}_{1,s_2}$. To arrive at the expression for $\boldsymbol{U}_{1|2}$ no operations need be performed on a full $k_I \times k_I$ covariance matrix. Furthermore, you only need to calculate the first n eigenvectors of $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$.

C.2.3 Simplifying the calculation of $(\boldsymbol{x}_1 - \boldsymbol{\mu}_{1|2})^T \boldsymbol{\Sigma}_{1|2}^{-1} (\boldsymbol{x}_1 - \boldsymbol{\mu}_{1|2})$

It is worth noting that this calculation would typically be done at run time, so it is important to simplify. $(\boldsymbol{x}_1 - \boldsymbol{\mu}_{1|2})$ is a $k_1 \times 1$ matrix. It is straight forward to show that

$$\boldsymbol{\Sigma}_{c11}^{-1} = \left[\, \boldsymbol{U}_{1|2,s} \; \boldsymbol{U}_{1|2,s_2} \,\right] \begin{bmatrix} \boldsymbol{D}^{-2}_{c11V,s} & 0 \\ 0 & \frac{1}{2}\epsilon_1^{-2} \boldsymbol{I} \end{bmatrix} \left[\, \boldsymbol{U}_{1|2,s} \; \boldsymbol{U}_{1|2,s_2} \,\right]^T (n - 1) \tag{C.21}$$

$$
\begin{aligned}
&(\boldsymbol{x}_1 - \boldsymbol{\mu}_{1|2})^T \boldsymbol{\Sigma}_{1|2}^{-1} (\boldsymbol{x}_1 - \boldsymbol{\mu}_{1|2}) \\
=&(\boldsymbol{x}_1 - \boldsymbol{\mu}_{1|2})^T \left[\boldsymbol{U}_{1|2,s} \; \boldsymbol{U}_{1|2,s_2} \right] \begin{bmatrix} \boldsymbol{D}_{c11V,s}^{-2} & 0 \\ 0 & \frac{1}{2}\epsilon_1^{-2}\boldsymbol{I} \end{bmatrix} \left[\boldsymbol{U}_{1|2,s} \; \boldsymbol{U}_{1|2,s_2} \right]^T (\boldsymbol{x}_1 - \boldsymbol{\mu}_{1|2})(n_s - 1) \\
=&\left[(\boldsymbol{x}_1 - \boldsymbol{\mu}_{1|2}^T)\boldsymbol{U}_{1|2,s} \; (\boldsymbol{x}_1 - \boldsymbol{\mu}_{1|2})^T \boldsymbol{U}_{1|2,s_2} \right] \begin{bmatrix} \boldsymbol{D}_{c11V,s}^{-2} & 0 \\ 0 & \frac{1}{2}\epsilon_1^{-2}\boldsymbol{I} \end{bmatrix} \left[(\boldsymbol{x}_1 - \boldsymbol{\mu}_{1|2})^T \boldsymbol{U}_{1|2,s} \; (\boldsymbol{x}_1 - \boldsymbol{\mu}_{1|2})^T \boldsymbol{U}_{1|2,s_2} \right]^T (n_s - 1) \\
=&(\boldsymbol{x}_1 - \boldsymbol{\mu}_{1|2})^T \boldsymbol{U}_{1|2,s} \boldsymbol{D}_{c11V,s}^{-2} \boldsymbol{U}_{1|2,s}^T (\boldsymbol{x}_1 - \boldsymbol{\mu}_{1|2}) \\
&+ \frac{1}{2}\epsilon_1^{-2} (\boldsymbol{x}_1 - \boldsymbol{\mu}_{1|2})^T \boldsymbol{U}_{1|2,s_2} \boldsymbol{U}_{1|2,s_2}^T (\boldsymbol{x}_1 - \boldsymbol{\mu}_{1|2}))(n_s - 1) \\
=&[(\boldsymbol{x}_1 - \boldsymbol{\mu}_{1|2})^T \boldsymbol{U}_{1|2,s} \boldsymbol{D}_{c11V,s}^{-2} \boldsymbol{U}_{1|2,s}^T (\boldsymbol{x}_1 - \boldsymbol{\mu}_{1|2}) + \frac{1}{2}\epsilon_1^{-2} (\boldsymbol{x}_1 - \boldsymbol{\mu}_{1|2})^T (\boldsymbol{x}_1 - \boldsymbol{\mu}_{1|2})](n_s - 1)
\end{aligned}
\tag{C.22}
$$

Note the dimensions of the matrices. $(\boldsymbol{x}_1 - \boldsymbol{\mu}_{1|2})$ is $k_1 \times n_s$, $\boldsymbol{U}_{1|2,s}$ is $k_1 \times n_s$, and $\boldsymbol{D}_{c11V,s}^{-2}$ is an $n_s \times n_s$ diagonal matrix. Furthermore, the full conditional covariance need not be saved, only the first n_s eigenvectors of the matrix and their respective eigenvalues.

Appendix D

Calculating Conditional Mode Parameters

The probability of one shape given another can be evaluated by the inner product of the mode parameters for a shape model that is based on the conditional distribution. This appendix derives an expression by which the mode parameters of the shape model (no conditional) can be transformed into the mode parameters of the conditional shape model. Provided that the number of modes being optimized in the shape model (no conditional) is less than the number of subjects (n_s), the transformation can be expressed in terms of a two $n_s \times n_s$ by $n_s \times 1$ matrix multiplications.

It can easily be shown that

$$(\boldsymbol{x}_1 - \boldsymbol{\mu}_{1|2})^T \Sigma_{1|2}^{-1} (\boldsymbol{x}_1 - \boldsymbol{\mu}_{1|2}) = \boldsymbol{b}_{1|2}^T \boldsymbol{b}_{1|2} \gamma_v \tag{D.1}$$

where we parameterize the shape $\boldsymbol{x}_1$ in terms of eigenvectors and eigenvalues of the conditional distribution, $p(\boldsymbol{x}_1 \mid \boldsymbol{x}_2)$. $\boldsymbol{x}_1$ for the conditional shape model is given by

$$\boldsymbol{x}_1 = \boldsymbol{\mu}_{1|2} + \boldsymbol{U}_{1|2}\boldsymbol{D}_{\epsilon_{c11}}\sqrt{\frac{\alpha + \boldsymbol{b}_2^T\boldsymbol{b}_2\gamma_v}{\alpha + k_2}}\sqrt{\frac{\gamma_v}{n_s - 1}}\boldsymbol{b}_{1|2} \tag{D.2}$$

Now, parameterizing the shape $\boldsymbol{x}_1$ in terms of $p(\boldsymbol{x}_1)$, we obtain

$$\boldsymbol{x}_1 = \boldsymbol{\mu}_1 + \boldsymbol{U}_1\boldsymbol{D}_{\epsilon_1}\sqrt{\frac{\gamma_v}{n_s - 1}}\boldsymbol{b}_1 \tag{D.3}$$

Equating D.2 and D.3, we arrive at

$$\boldsymbol{\mu}_1 + \boldsymbol{U}_1\boldsymbol{D}_{\epsilon_1}\sqrt{\frac{\gamma_v}{n_s - 1}}\boldsymbol{b}_1 = \boldsymbol{\mu}_{1|2} + \boldsymbol{U}_{1|2}\boldsymbol{D}_{\epsilon_{c11V}}\sqrt{\frac{\alpha + \boldsymbol{b}_2^T\boldsymbol{b}_2\gamma_v}{\alpha + k_2}}\sqrt{\frac{\gamma_v}{n_s - 1}}\boldsymbol{b}_{1|2} \tag{D.4}$$

From appendix C, $\boldsymbol{U}_{1|2} = \boldsymbol{U}_1\boldsymbol{U}_{c11V}$ and $\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \boldsymbol{Z}_1V_2\boldsymbol{D}_2\boldsymbol{D}_{\epsilon_2}^{-1}\sqrt{\frac{\gamma_v}{n_s-1}}\boldsymbol{b}_2$. Substituting these into D.4, we obtain

$$\begin{aligned}
&\boldsymbol{\mu}_1 + \boldsymbol{U}_1\boldsymbol{D}_{\epsilon_1}\sqrt{\frac{\gamma_v}{n_s - 1}}\boldsymbol{b}_1 = \\
&\quad \boldsymbol{\mu}_1 + \boldsymbol{Z}_1V_2\boldsymbol{D}_2\boldsymbol{D}_{\epsilon_2}^{-1}\sqrt{\frac{\gamma_v}{n_s - 1}}\boldsymbol{b}_2 + \boldsymbol{U}_1\boldsymbol{U}_{c11V}\boldsymbol{D}_{\epsilon_{c11V}}\sqrt{\frac{\alpha + \boldsymbol{b}_2^t\boldsymbol{b}_2\gamma_v}{\alpha + k_2}}\sqrt{\frac{\gamma_v}{n_s - 1}}\boldsymbol{b}_{1|2}
\end{aligned} \tag{D.5}$$

Rearranging, it simplifies to

$$\begin{aligned}
\boldsymbol{b}_{1|2} = &\left(\sqrt{\frac{\alpha + k_2}{\alpha + \boldsymbol{b}_2^T\boldsymbol{b}_2\gamma_v}}\boldsymbol{D}_{\epsilon_{c11V}}^{-1}\boldsymbol{U}_{c11V}^t\boldsymbol{D}_{\epsilon_1}\right)\boldsymbol{b}_1 \\
&- \left(\sqrt{\frac{\alpha + k_2}{\alpha + \boldsymbol{b}_2^T\boldsymbol{b}_2\gamma_v}}\boldsymbol{D}_{\epsilon_{c11V}}^{-1}\boldsymbol{U}_{c11V}^t\boldsymbol{D}_1V_1^tV_2\boldsymbol{D}_2\boldsymbol{D}_{\epsilon_2}^{-1}\right)\boldsymbol{b}_2
\end{aligned} \tag{D.6}$$

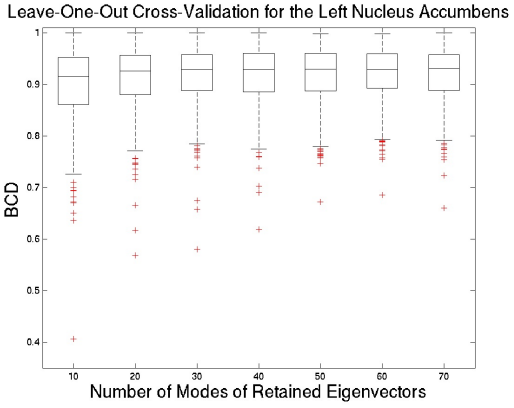
Since all the singular values of $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ above n_s are zero, we can simplify to get

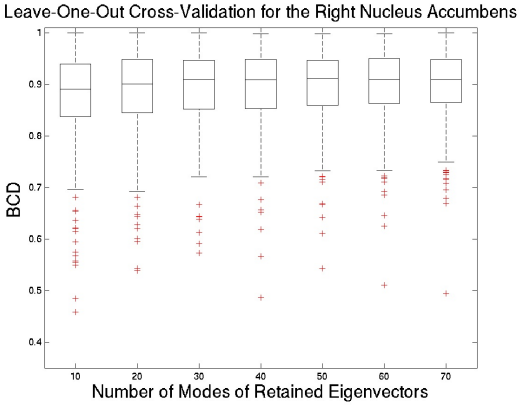
$$\begin{aligned}\boldsymbol{b}_{1|2} = &\left(\sqrt{\frac{\alpha + k_2}{\alpha + \boldsymbol{b}_2^T \boldsymbol{b}_2 \gamma_v}} \boldsymbol{D}^{-1}_{\epsilon_{c11V,s}} \boldsymbol{U}^t_{c11V,s} \boldsymbol{D}_{\epsilon_1}\right) \boldsymbol{b}_1 \\ &- \left(\sqrt{\frac{\alpha + k_2}{\alpha + \boldsymbol{b}_2^T \boldsymbol{b}_2 \gamma_v}} \boldsymbol{D}^{-1}_{\epsilon_{c11V,s}} \boldsymbol{U}^t_{c11V} \boldsymbol{D}_1 V_1^t V_2 \boldsymbol{D}_2 \boldsymbol{D}^{-1}_{\epsilon_{2,s}}\right) \boldsymbol{b}_2\end{aligned} \tag{D.7}$$

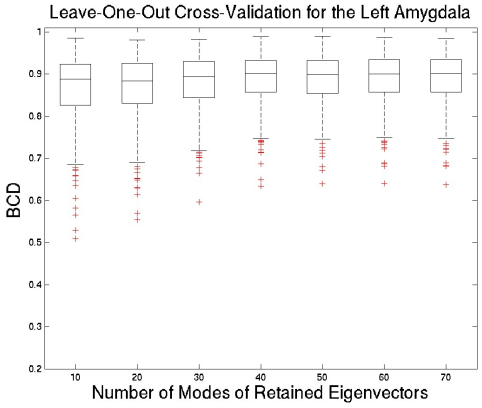
Appendix E

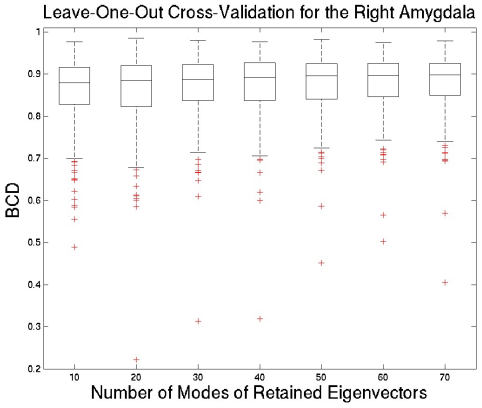
Results from Leave-One-Out Cross Validation

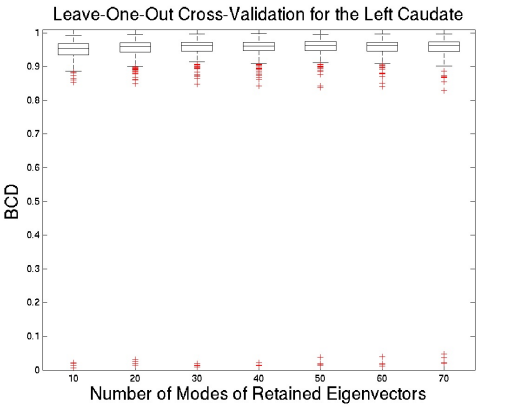
Presented in this appendix are the leave-one-out (LOO) cross-validation results for each modelled structure using 10, 20, 30, 40, 50, 60 and 70 modes of variation. The LOO cross-validation was performed for the entire training set (317 subjects) and are presented in box plots.

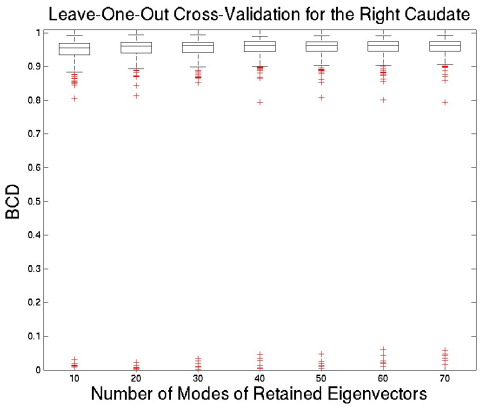

Leave-One-Out Cross-Validation for the Left Nucleus Accumbens
BCD
Number of Modes of Retained Eigenvectors

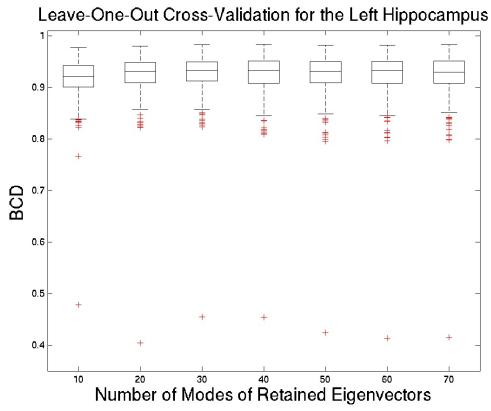

Leave-One-Out Cross-Validation for the Right Nucleus Accumbens
BCD
Number of Modes of Retained Eigenvectors

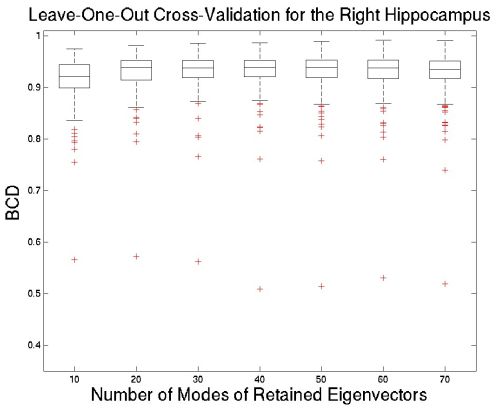

Leave-One-Out Cross-Validation for the Left Amygdala
BCD
Number of Modes of Retained Eigenvectors

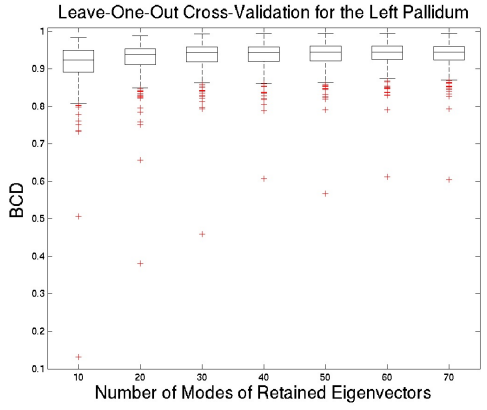

Leave-One-Out Cross-Validation for the Right Amygdala
BCD
Number of Modes of Retained Eigenvectors

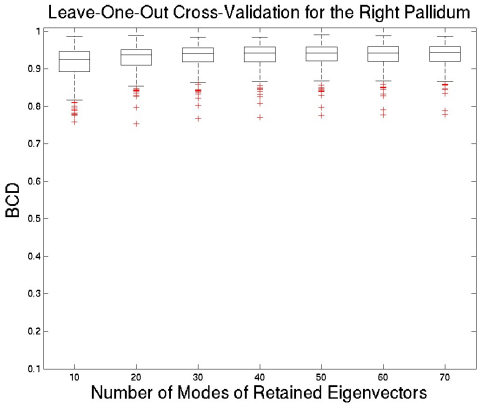

Leave-One-Out Cross-Validation for the Left Caudate
BCD
Number of Modes of Retained Eigenvectors

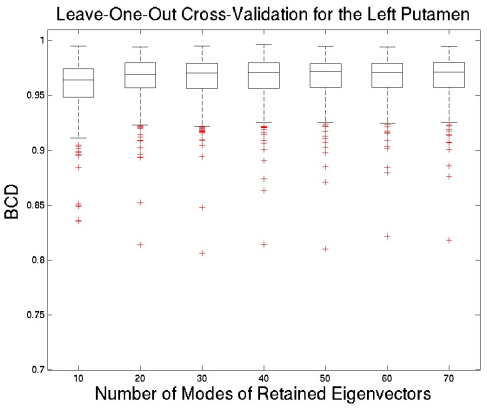

Leave-One-Out Cross-Validation for the Right Caudate
BCD
Number of Modes of Retained Eigenvectors

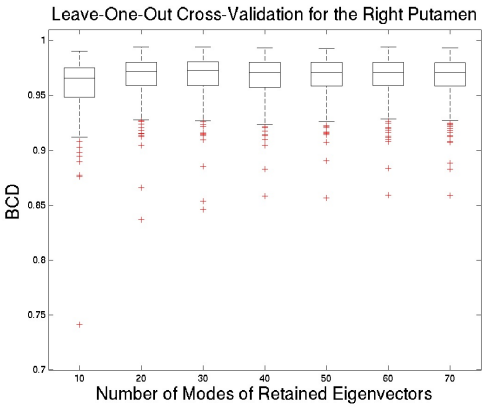

Leave-One-Out Cross-Validation for the Left Hippocampus
BCD
Number of Modes of Retained Eigenvectors


Leave-One-Out Cross-Validation for the Right Hippocampus
BCD
Number of Modes of Retained Eigenvectors


Leave-One-Out Cross-Validation for the Left Pallidum
BCD
Number of Modes of Retained Eigenvectors


Leave-One-Out Cross-Validation for the Right Pallidum
BCD
Number of Modes of Retained Eigenvectors


Leave-One-Out Cross-Validation for the Left Putamen
BCD
Number of Modes of Retained Eigenvectors


Leave-One-Out Cross-Validation for the Right Putamen
BCD
Number of Modes of Retained Eigenvectors

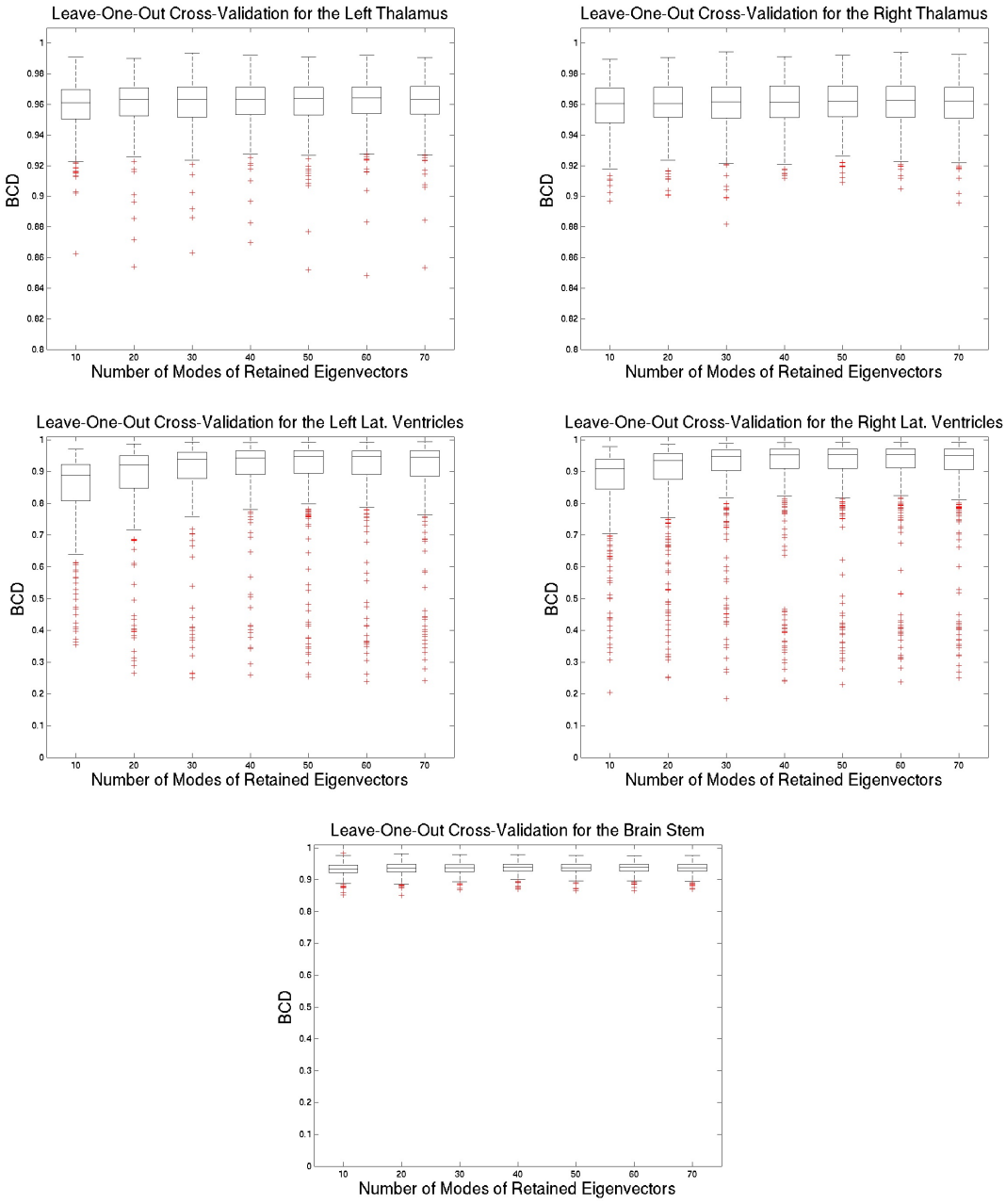


Figure E.1: Leave-one-out overlap results using 10, 20, 30, 40 ,50 ,60 and 70 modes of variation with ϵ_I and ϵ_s equal to 0.0001% of the total shape and intensity variance respectively.