Bayesian Shape and Appearance Models

FMRIB Technical Report TR07BP1

Brian Patenaude[1], Stephen Smith[1], David Kennedy[2] and Mark Jenkinson[1]

1:Oxford Centre for Functional Magnetic Resonance Imaging of the Brain (FMRIB),
Department of Clinical Neurology, University of Oxford, John Radcliffe Hospital,
Headley Way, Headington, Oxford, UK
2: Center for Morphometric Analysis, MGH, Boston

Abstract

Active shape and appearance models are widely used in image segmentation. In this paper, 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 which are otherwise problematic due to rank deficient covariance estimates. The appearance model is framed as the conditional distribution for a given shape, which is useful as then the posterior can then be used to fit the model to new data. The conditional distribution may also be used in terms of predicting one shape distribution given the location of another shape. This framework generalizes to other types of data beyond shape and intensity; for example age as a predictor of shape. The analytic form for the conditional distribution scales appropriate covariances in such a way that one does not need an empirical/arbitrary weighting for relating intensity variance to shape variance as is usually required. For this paper the framework is applied to sub-cortical brain models.

1 Introduction

The accuracy and robustness of medical image segmentation algorithms are important for studying normal and pathological cases. The challenge is to achieve accuracy and robustness in the presence of low contrast-to-noise. A trained technologist or clinician draws on prior knowledge such as shape, topology, and texture when performing manual segmentation. The brain, for example, consists of many substructures with generally consistent topology, shape, and inter-shape relationships, such that knowledge of one structure aids in the segmentation of another. Our aim is to formulate a shape and appearance model that can incorporate this intra- and inter-structure variability information. Furthermore, the model should be able to account for the high dimensionality of the data with respect to the size of the training set.

In order to improve robustness and accuracy, higher level information is integrated into segmentation algorithms through shape priors. The active shape model (ASM) is one such example that has become widely used in the field of machine vision and medical image segmentation over the past decade [5]. ASMs model the vertices (control points) of a structure as a multivariate Gaussian distribution. Shape is then parameterized in terms of its mean and eigenvectors. 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.

Intensity priors also provide a rich set of information, the active appearance model (AAM) is an extension of the ASM framework that incorporates intensity priors [6]. As with shape, the intensity distribution is modelled 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 weights are determined using the root-mean-squared differences in intensity for small deviations in the shape parameters. Using this weighting matrix, the separate intensity and shape parameterizations are combined into a single model. The AAM is fit to new data by minimizing the squared difference between the predicted intensities given a shape deformation and the observed image intensities.

A limitation of shape and appearance models are their dependency on training data. When training these models, particularly in 3D, we are are dealing with a largely underdetermined inverse problem (and hence rank-deficient covariance matrices). 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 $3N$. The dimensionality of appearance models are increased by the number of intensity samples. In our application, the number of control points within a single structure ranges from a few hundred to a few thousand. For a single shape model we are, therefore, dealing with a dimensionality ranging from approximately a thousand, upwards to ten thousand.

In practice, particularly in the medical field, the number of subjects used to train from is small compared to the dimensionality of the model. For example our 3D model of the left putamen consists of 2562 vertices, the shape model would thus have a dimensionality of 3×2562, which by far exceeds our 139 training sets. The situation worsens substantially when modelling multiple structures, since the dimensionality increases whilst the number of subjects remains unaltered. Typically the solution to this mixed-determined problem 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 reflects the eigenvectors that span the unseen variation from the unsampled population.

Structural co-variation is a valuable piece of information when fitting multiple objects. Using this information, we hope to improve robustness and accuracy, particularly in structures with a low contrast-to-noise ratio. Canonical correlation analysis (CCA), "SVD method" (this differs from SVD) [2], combined principal component analysis (CPCA) are various methods for examining co-variation of structure [15, 2]. All these methods provide a means for predicting one structure from another. The methods differ primarily in their optimization criteria; CCA is similar to PCA except that it maximizes the normalized correlation rather than variance, "SVD method" optimizes purely for co-variation in the data (by taking the SVD of the cross-covariance matrix) and is directly related to partial least squares (PLS), and CPCA optimizes the total variation (SVD on the concatenated data). Our framework provides a natural means of incorporating structural co-variation through shape priors and conditional distributions.

The various existing methods for subcortical segmentation encompass both surface-based and volumetric-based approaches. Predominantly, volumetric based approaches are based on a non-linear warp of an atlas to new data [14, 8, 3]. One of the most prevalent methods for subcortical segmentation is ASEG [8]. In addition to an average template, ASEG uses voxel-wise intensity and shape priors, the shape prior is an anisotropic Markov random field on the labels; the prior's parameters are learned from the training data.

Surface-based methods attempt, on the other hand to use learned shape variation as a prior in the segmentation [17, 13, 4]. In [13], the learned shape variation and empirically derived distance and texture metrics are used to help constrain the deformation. The relative weights between constraints are arbitrary and vary across structures. In [4] fuzzy spatial relations are incorporated with a deformable model; the parameters of the fuzzy relations are learned from the data. As with other deformable model techniques, they all require an arbitrary weighting between the forces. In [17] the zero-level set of the signed distance function implicitly models the surface and its co-variation by applying PCA to the signed distance functions. Mutual information between labels and intensities is used to fit the model. We provide a probabilistic framework for incorporating learned metrics that eliminates the need for empirical weightings.

In this paper we are proposing a general Bayesian framework for modelling data from a finite training set. The framework explicitly takes into account the inadequacy of the training data in estimating the covariance matrix of the multivariate Gaussian. The conditional distributions provide a probabilistic model for inter- and intra-intensity/shape co-variation without requiring empirical weightings or the application of ill-conditioned matrix inverse. More generally, this framework can model co-variation between any two attributes within the multivariate Gaussian. Section 2 will lay out the statistical framework for our shape and appearance models. Results of their application and discussion will follow in section 3.

2 Methods

2.1 Training Data and Mesh Parameterization

The training data consists of 139 manually labelled T1 weighted magnetic resonance images of the brain. All the training data was linearly registered to MNI152 space using FLIRT [9]. The sample population spans images of normal and pathological brains (including schizophrenia and Alzheimer's disease). We are modelling 19 structures: brainstem and the left/right amygdala, caudate nucleus, cerebellum, hippocampus, lateral ventricles, nucleus accumbens, putamen, pallidum, thalamus.

The volumetric labels are parameterized by deforming a 3D mesh representation of the most typical (across

subjects) structure to each subject. The necessary cross-subject vertex correspondence is preserved by within-surface motion constraints and minimal smoothing forces within the 3D deformable model [10, 11, 16]. By sampling normalized intensities along the surface normal at each vertex, we are able to generate appearance models; we used 13 samples per vertex at a 05.mm interval. In this paper, we normalize the intensities by subtracting the median intensity across a given structure; however, residuals to a planar fit (instead of median) could be used to model out linear intensity drifts in x,y,z directions.

2.2 Shape and Appearance Models from Finite Training Data

2.2.1 Mathematical Model

Given that we have a finite set of training data $z = \{\tilde{x}_1 \dots \tilde{x}_{n_s}\}$. our model of the underlying distribution is a multivariate Gaussian distribution given by

$$p(x_i \mid \mu, \lambda) = N_k(x_i \mid \mu, \lambda) \tag{1}$$

where k is the dimensionality of $x_i \; \epsilon \; \Re^k$, μ is the mean and λ is a $k \times k$ positive-definite precision matrix.

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

$$p(x_{obs} \mid z) = \int p(x_{obs} \mid \mu, \lambda) p(\mu, \lambda \mid z) d\mu d\lambda \tag{2}$$

where $x_{obs} = \{x_{n_s+1}, ..., x_{n_s+m}\}$ is a set of new observables, such that $m \geq 1$.

Given the sufficient statistics $t(z)$ [1], it can be shown that

$$\begin{aligned} p(x_{obs} \mid z) &= p(x_{obs} \mid t(z)) \\ &= \int p(x_{obs} \mid \mu, \lambda) p(\mu, \lambda, \mid t(z)) d\mu d\lambda \end{aligned} \tag{3}$$

We use the sufficient statistics for the multivariate Gaussian given by

$$t(z) = (n_s, \bar{x}, S) \tag{4}$$

where

$$\bar{x} = n_s^{-1} \Sigma_{i=1}^{n_s} \tilde{x}_i, \tag{5}$$

$$S = \Sigma_{i=1}^{n} (\tilde{x}_i - \bar{x})(\tilde{x}_i - \bar{x})^t. \tag{6}$$

The expression for $p(x_{obs} \mid \mu, \lambda)$ is given by the predictive model in (1). In order to evaluate (3) we now need to derive an expression for $p(\mu, \lambda, \mid t(z))$.

Using Bayes' theorem,

$$p\left(\mu, \lambda \mid t(z)\right) = \frac{p(t(z) \mid \lambda, \mu) p(\lambda, \mu)}{\int p(t(z) \mid \lambda, \mu) p(\lambda, \mu) d\mu d\lambda} \tag{7}$$

where

$$p(t(z) \mid \mu, \lambda) = p(S \mid \bar{x}, \mu, \lambda) p(\bar{x} \mid \mu, \lambda) \tag{8}$$

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

$$p(\bar{x} \mid n_s, \mu, \lambda) = N_k(\bar{x} \mid \mu, n_s \lambda) \tag{9}$$

$$p(S \mid \bar{x}, n_s, \mu, \lambda) = Wi_k \left(S \mid \frac{1}{2}(n_s - 1), \frac{1}{2}\lambda \right) \tag{10}$$

δ is the dirac delta function. Wi_k is a Wishart distribution with $\frac{1}{2}(n-1)$ degrees of freedom, and a precision matrix of $\frac{1}{2}\lambda$. For this case, to satisfy the requirements of the Wishart distribution n must be greater than k. Substituting (9) and (10) back into (8), we arrive at

$$\begin{aligned} &p(t(z) \mid \mu, \lambda) \\ &\quad = N_k(\bar{x} \mid \mu, n_s\lambda) Wi_k\left(S \mid \frac{1}{2}(n_s-1), \frac{1}{2}\lambda\right) \end{aligned} \tag{11}$$

To calculate the posterior $p(x_{obs} \mid t(z))$, we need to specify the prior $p(\mu, \lambda)$. Using the conjugate prior, and introducing the hyperparameters n_0, μ_0, and β, the prior is given by

$$p(\mu, \lambda \mid \mu_0, n_0, \beta) = N(\mu \mid \mu_0, n_0\lambda) Wi_k(\lambda \mid \alpha, \beta) \tag{12}$$

By substituting (11) and (12) into (7), followed by (7) and (1) into (3), and then integrating, we obtain

$$\begin{aligned} &p(x_{obs} \mid z, n_0, \mu_0, \beta, \alpha) = \\ &St_k(x_{obs} \mid \mu_n, (n_0+n_s+1)^{-1}(n_0+n_s)\alpha_n\beta_n^{-1}, 2\alpha_n) \\ where &\mu_n = (n_0+n_s)^{-1}(n_0\mu_0 + n_s\bar{x}) \\ &\beta_n = \beta + \frac{1}{2}S + (n_s+n_0)^{-1} n_s n_0 (\mu_0 - \bar{x})(\mu_0 - \bar{x})^t \\ &\alpha_n = \alpha + \frac{1}{2}n_s - \frac{1}{2}(k-1) \end{aligned} \tag{13}$$

and St_k is a multivariate Student distribution. The posterior and marginal distributions resulting from (13) are given in [1].

The full expression for a multivariate Student distribution is given by

$$St_k(x \mid \mu, \lambda, \alpha) = c\left[1 + \frac{1}{\alpha}(x-\mu)^t\lambda(x-\mu)\right]^{-\frac{\alpha+k}{2}} \tag{14}$$

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

$$V[x] = \lambda^{-1}\frac{\alpha}{\alpha-2}. \tag{15}$$

2.3 Choice of priors

The first prior chosen is $n_0 = 0$, as this is a flat, non-informative prior on $p(\mu)$ and results in $p(x_{obs} \mid z, n_0, \mu_0, \beta, \alpha)$ being centered at $\bar{x}$ with no dependence on μ_0. By substituting $n_0 = 0$ back into (13) we obtain

$$\begin{aligned} &p(x_{obs} \mid z, n_0 = 0, \beta, \alpha) \\ &\quad = St_k(x_{obs} \mid \mu_n, \frac{n_s}{n_s+1}\alpha_n\beta_n^{-1}, 2\alpha_n) \\ &\mu_n = \bar{x} \\ &\beta_n = \beta + \frac{1}{2}S \\ &\alpha_n = \alpha + \frac{1}{2}n_s - \frac{1}{2}(k-1) \end{aligned} \tag{16}$$

Now, expanding (16) into the general form for the multivariate Student distribution, and rearranging we obtain

$$\begin{aligned} &p(x_{obs} \mid z, n_0 = 0, \beta, \alpha) \\ &= c\left[1 + \frac{1}{n_s - \frac{1}{n_s}}(x-\bar{x})^t\left(\frac{S+2\beta}{n_s-1}\right)^{-1}(x-\bar{x})\right]^{\frac{-(k+n_s-\frac{1}{n_s})}{2}} \end{aligned} \tag{17}$$

The prior α is chosen such that the sample covariance is normalized by $n-1$ (which corresponds to the standard unbiased estimate of a covariance matrix). It follows that

$$2\alpha_{n_s} = n_s - \frac{1}{n_s} \tag{18}$$

$$\alpha = \frac{1}{2}\left(k + 1 - \frac{1}{n_s}\right). \tag{19}$$

This meets the minimum criteria for degrees of freedom, $2\alpha > k-1$, as given by (12). For rotational invariance, β is chosen to be a scaled identity matrix ϵI. ϵ is typically chosen as a percentage of the total variance.

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

Our model takes the final form

$$\begin{aligned} &p(x_{obs} \mid z, \epsilon) \\ &\quad = St_k\left(x_{obs} \mid \bar{x}, \frac{S + 2\epsilon^2}{n_s - 1}, n_s - \frac{1}{n_s}\right) \end{aligned} \tag{20}$$

The variance given by

$$V[x_{obs}] = \left(\frac{S + 2\epsilon^2}{n_s - 1}\right)\gamma_v \tag{21}$$

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

2.4 Conditional distributions

We are interested in conditional distributions across partitions of the joint multivariate Gaussian model. A partition is a subset, x_{ij}, of x_i 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 shape and intensity, or into different shapes. Shape/intensity partitions are used to predict intensity distributions given a particular shape, where as the shape/shape partitions predict shape distributions given another shape.

If x can be partitioned such that,

$$x = (x_1, x_2) \tag{22a}$$

$$\lambda = \begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{bmatrix} \tag{22b}$$

$$k = k_1 + k_2 \tag{22c}$$

where k_j is the dimensionality of the jth partition, then z can be partitioned in the same manner, such that $z = (z_1, z_2)$, where $z_j = \{\tilde{x}_{ij} \dots \tilde{x}_{n_s j}\}$. It follows that

$$p(x_1 \mid x_2, z) = St_{k_1}(x_1 \mid x_2, \mu_{1|2}, \lambda_{1|2}, \alpha_{1|2}) \tag{23a}$$

where

$$\begin{aligned} \mu_{1|2} &= \mu_1 - \lambda_{11}^{-1}\lambda_{12}(x_2 - \mu_2) \\ &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \end{aligned} \tag{23b}$$

$$\begin{aligned} \lambda_{1|2} &= \lambda_{11}\left[\frac{\alpha + k_2}{\alpha + (x_2 - \bar{x_2})^T\Sigma_{22}^{-1}(x_2 - \bar{x_2})}\right] \\ \Sigma_{1|2} &= \left(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)\left[\frac{\alpha + (x_2 - \bar{x_2})^T\Sigma_{22}^{-1}(x_2 - \bar{x_2})}{\alpha + k_2}\right] \end{aligned} \tag{23c}$$

$$\alpha_{1|2} = \alpha_{1,2} + k_2. \tag{23d}$$

For a partitioned covariance matrix the prior β will be defined as a piece-scaled identity matrix, such that for two partitions β is given by

$$\beta = \begin{bmatrix} \epsilon_1^2 I & 0 \\ 0 & \epsilon_2^2 I \end{bmatrix}. \tag{24}$$

2.5 Parameterization of Bayesian Models from a Finite Training Data

Defining $\tilde{z}$ as the demeaned training set z, we express $\tilde{z}$ in terms of its SVD,

$$\tilde{z} = UDV^T \tag{25}$$

Where U are the eigenvectors of the covariance matrix, D are the singular values, and V is the parameter matrix needed to reconstruct the original data.

Adding a scaled identity matrix to a covariance matrix is equivalent to adding a scalar to each eigenvalue, D_i^2, including the zero eigenvalues which correspond to the null space [12]. The covariance matrix, $2\epsilon^2 I + S$, in terms (25) is given by

$$\begin{aligned} \Sigma\gamma_v &= (2\epsilon^2 I + S)(n_s - 1)^{-1}\gamma_v \\ &= U(D^2 + 2\epsilon^2 I)U^T(n_s - 1)^{-1}\gamma_v \\ &= UD_\epsilon^2 U^T(n_s - 1)^{-1}\gamma_v \end{aligned} \tag{26}$$

where D_ϵ^2 is a diagonal matrix consisting of the eigenvalues of $2\epsilon^2 I + S$.

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 savings when n_s is much less than k_j.

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

$$x = \bar{x} + U\frac{D_\epsilon}{\sqrt{(n_s - 1)\gamma_v}}b \tag{27}$$

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

2.6 Bayesian Appearance Models

Our mathematical framework is now applied to appearance models. The joint distribution of shape and intensity are being modelled as a multivariate Gaussian distribution. From our training set, using the model given by (20), we learn the joint intensity/shape distribution, $p(x_i, x_s)$. Given that $p(x_i, x_s \mid z)$ is partitionable, we can calculate the conditional intensity distribution, $p(x_i \mid x_s, z)$, given a particular shape and a finite training set. $p(x_i \mid x_s, z)$ takes the form of (23) with x_i and x_s corresponding to partitions x_1 and x_2 respectively. The shape partition is modelled using (27), so for any b_s vector (new shape instance) we can predict the intensity distribution.

2.7 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 very large covariance matrices, which are very computationally expensive as well as using large amounts of memory. Furthermore, the calculation of the conditional is dependent on calculating the inverse of the covariance matrix. We are able to eliminate all operations on $k \times k$ matrices, making the models computationally feasible in practice.

2.7.1 Conditional Mean as a Function of the Predictor Parameters b_2

Given training data with two partitions z_1 and z_2 and the latter being parameterized in (27), the conditional mean can be expressed as a function of predictors model parameter 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 b_2 is given by

$$\mu_{1|2} = \mu_1 + z_{1dm}[V_2 D_{2,s} D_{\epsilon_2,s}^{-1} \sqrt{\frac{\gamma_v}{n_s - 1}} b_{2,s}] \tag{28}$$

In general the ",s" subscript refers to the upper-left submatrix, such that the maximum dimension is n_s. $b_{2,s}$ is the first n_s rows of b_2. All matrices within square brackets of (28) are of size $n_s \times n_s$ except $b_{2,s}$ which is $n_s \times 1$. If we truncate modes at L, only the first L columns of $\tilde{z}_1[V_2 D_{2,s} D_{\epsilon_2,s}^{-1} \sqrt{\frac{\gamma_v}{n_s - 1}}]$ are needed. See appendix A.1 for details of the derivation.

2.7.2 Evaluating Conditional Covariance Operations

In order to simplify the calculation of the conditional probability we need to simplify operations involving a covariance matrix (23c). Conditional covariances 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. 2) To explicitly calculate the probability of a predicted measure, e.g. to calculate the probability of certain intensity profile given a known shape.

In case 1, we need to calculate the eigenvectors and eigenvalues for $\Sigma_{1|2}$; though 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 $\Sigma_{1|2}$ can be a very expensive operation given that the number of control points in practice is very large. In case 2, we need to evaluate $(x_I - \mu_I)^T \Sigma_{1|2}^{-1} (x_I - \mu_I)$. For both cases we will exploit the fact that n_s is typically much larger than k to simplify the calculations, though the results are valid for any $n_s < k$. These simplification are left to appendix A.2.

2.8 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. Hence, when fitting the Bayesian appearance model $p(x_I \mid p_s)$, we aim to minimize -$\ln p(x_s \mid x_I)$, as given by

$$p(x_s \mid x_I) = \frac{p(x_I \mid x_s) p(x_s)}{p(x_I)} \tag{29}$$

$$-\ln p(x_s \mid x_I) = -\ln p(x_I \mid x_s) - \ln p(x_s) + \ln p(x_I) \tag{30}$$

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 minimizing (30) with respect to b_s, the shape model parameters. Given the shape model, and a search with respect to b_s, the posterior simplifies down to

$$\begin{aligned}
&p(x_s \mid x_I) \\
&= C + \frac{k_I}{2} \ln\left(\frac{\alpha_{I,s} + k_s}{\alpha_{I,s} + b_s^T b_s \gamma_v}\right) \\
&- \frac{(\alpha_{I,s} + k_s + k_I)}{2} \\
&\ln\left(1 + \frac{1}{\alpha_{Is} + b_s^T b_s \gamma_v} (x_I - \mu_{I|s})^t \lambda_{cII} (x_I - \mu_{I|s})\right) \\
&+ \frac{(\alpha_s + k_s)}{2} \ln\left(1 + \frac{1}{\alpha_s} b_s^T b_s \gamma_v\right) \\
&- \frac{(\alpha_I + k_I)}{2} \ln\left(1 + \frac{1}{\alpha_I} (x_I - \mu_I)^t \lambda_I (x_I - \mu_I)\right)
\end{aligned} \tag{31}$$

where α_{Is} is the degrees of freedom of $p(x_I \mid p_s)$, λ_{cII} is the unscaled conditional precision matrix. k_I and k_s are the intensity and shape partition dimensionality respectively. $\mu_{I|s}$ is the conditional mean given shape (28)(which is a function of b_s). x_I are the observed intensities. See appendix B for the derivation.

2.9 Conditional Shape Priors

In practice, limiting the search space to joint modes of variation is difficult and the shape constraints are too strict. Given the amount of training data, limiting the search space to the joint modes is overly ambitious when generalizing to the unsampled population. Instead, structural co-variation is incorporated as a prior in our model as given by

$$p(x_{s1} \mid x_{I1}, x_{s2}) = \frac{p(x_{I1}, x_{s2} \mid x_{s1})p(x_{s1})}{p(x_I, x_{s2})} \tag{32}$$

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

If we make the naive assumption of independence between x_{I1} and x_{s2}, (32) simplifies to

$$\begin{aligned} p(x_{s1} \mid x_{I1}, x_{s2}) &= \frac{p(x_{I1} \mid x_{s1})p(x_{s2} \mid x_{s1})p(x_{s1})}{p(x_{I1})p(x_{s2})} \\ &= \frac{p(x_{I1} \mid x_{s1})p(x_{s1} \mid x_{s2})}{p(x_{I1})}. \end{aligned} \tag{33}$$

By making the assumption of independence we are reducing the maximum distribution dimensionality we are trying to estimate. Though, 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 prenounced for neighbouring structures.

The negative-log posterior is now given by

$$\begin{aligned} -\ln p(x_{s1} \mid x_{I1}, x_{s2}) = &-\ln p(x_{I1} \mid x_{s1}) - \ln p(x_{s2} \mid x_{s1}) \\ &+ \ln p(x_I). \end{aligned} \tag{34}$$

This differs from (30) in that the shape prior $p(x_{s1})$ is replaced by a conditional shape prior $p(x_{s1} \mid x_{s2})$. The evaluation of the conditional 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 a $n_s \times n_s$ matrix times a $n_s \times 1$ for each new parameter estimate of b_1 that is visited. See appendix D for details.

2.10 Model Fitting and Evaluation

The quality of fitting was evaluated using a leave-ten-out procedure. The training set was randomly divided into 13 groups of 10 and one group of 9. From these groups 12 training datasets were created, each of size 120, by excluding one of the groups of 10 and the group of 9. A separate model was fit to each of the training sets. When fitting the model to new data, the model is registered into the native space using a global affine transformation. To register the model, the linear transformation matrix need only be applied to the average shape and eigenvectors (see appendix C).

2.11 Overlap Metric

For all evaluation metrics the manual segmentations are regarded as the gold standard. The segmentation performance was measured using the Dice overlap metric given by

$$D = \frac{2TP}{2TP + FP + FN} \tag{35}$$

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 results from filling the output mesh. The mesh filling process consists of two steps: 1) drawing the mesh outline, and 2) filling the interior. We therefore know whether an output voxel belongs to the boundary or the interior. 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 and correctable at the volumetric level, then given a correction scheme, the BCD is the maximum overlap that can be achieved. Methods for boundary correction are beyond the scope of this paper, and will not be discussed here. The BCD is given by

$$BCD = \frac{2(TP_{int} + G_{bound})}{2(TP_{int} + G_{bound}) + FP_{int} + FN_{int}} \tag{36}$$

where TP_{int}, FP_{int}, FN_{int} are, respectively, the true positive, false positive, false negative volume that is contained within the interior of the filled mesh. G_{bound} is the ground truth volume contained within the boundary of the filled mesh. The BCD seems to be a similar method to that proposed in [7], where the condition for overlap at the boundary voxels is relaxed, based on the assumption that the boundary is wrong.

2.12 Shape Conditioned on Age

The framework can be applied to data other than shape and intensity. Age can be an important predictor of shape. The conditional shape/intensity distribution given age can be used to incorporate our prior knowledge of the subjects' age into the fitting algorithm. We will examine the effect of age on shape. In order to incorporate age into our framework it must have an underlying Gaussian distribution. Although we know that this is not strictly accurate, we adopt it here as an approximation. We model $p(x_{age})$ as a uniform distribution. We can then apply a transformation of random variables given by,

$$y_{age|x_{min},x_{max}} = \sigma\sqrt{2}erf^{-1}\left(2(\frac{(x_{age}-x_{min})}{x_{max}-x_{min}}-1)\right) \tag{37}$$

Where $p(y_{age} \mid x_{min}, x_{max})$ with a Gaussian distribution with zero mean and variance equal to σ. x_{min} and x_{max} are hyperparameters that corresponding to the minimum and maximum age that bounds our uniform distribution (the age bounds of the true age group). The conditional distribution is invariant to σ, for simplicity we choose $\sigma = 1$. As $p(x_s, y_{age})$ can now be modelled as a multivariate Gaussian, our framework can be used to calculate the conditional distribution $p(x_s \mid y_{age})$.

3 Results and Discussion

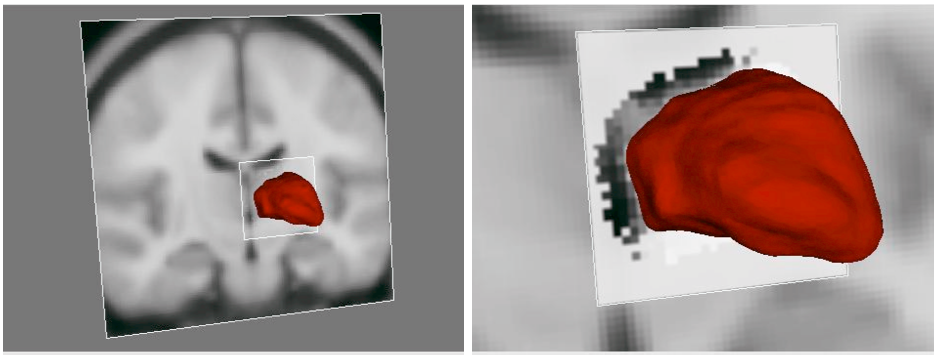
We will first qualitatively demonstrate a Bayesian appearance model, then follow with results on fitting various structures to data. By varying the shape parameter of the individual modes of variation, we can observe the surface deformations; along with the surface deformation the conditional intensity distribution is calculated. Figure 1 is a graphical depiction of ± 3 standard deviations along the the first mode of variation for the left thalamus and the conditional intensity mean associated with it; the model overlays the MNI152 template. For each vertex, 13 intensity samples were taken at a 0.5mm interval. 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.

The left putamen, thalamus, hippocamppus, amygdala, and nucleus accumbens were independently fit to 120 subjects using the leave-ten-out method, using individual appearance models. The left thalamus was also fit conditioned on the left putamen 2(b). The fit was performed across 20 modes of variation; the shape parameters corresponding to the eigenvectors were assumed to be zero. ϵ_s and ϵ_I were chosen to be $10^{-7}\%$ of the total shape and intensity variation.

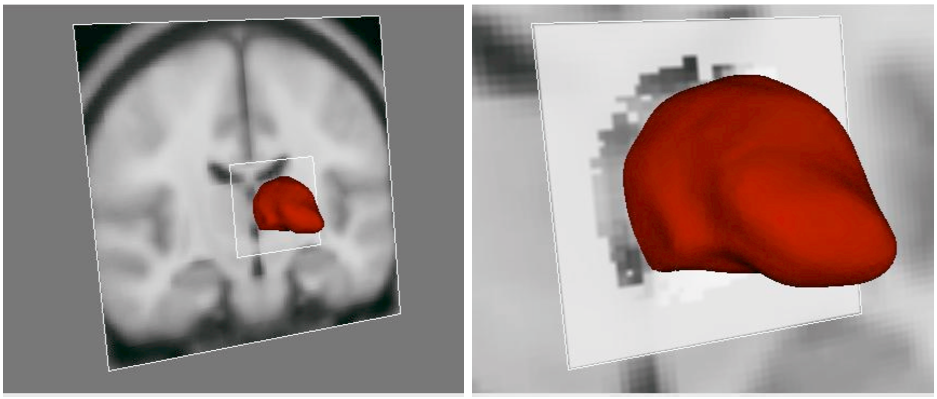
In figure 2 there is a decrease in overlap in the last two-thirds of the data. The data corresponding to the drop in overlap corresponds to lower resolution data. From the standard Dice measurement, it is unclear whether the decrease is due to lower performance at lower resolution or higher sensitivity of the Dice metric to the boundary voxel errors. Using the BCD metric, overlap increases significantly and evens out over resolutions, this leads us to believe that the decrease in Dice is due to an increased sensitivity at the boundary voxels at lower resolution. The overlaps reported on similar data for ASEG [8] lie between the Dice and the BCD overlap that we report in figure 2, being closer to our Dice measurement for the left putamen, amygdala.

To test the benefit of a conditional shape prior and with an increased ϵ_s, we chose the thalamus for subject 40 as a test case (the thalamus fitting performed very poorly without the conditional prior for this subject). Figure 3 shows the fitted thalamus for subject number 40 from 2(b). Figure 3(a) is the manual segmentation. Figure 3(b) is the boundary corrected segmentation when fitting the thalamus, disregarding other structures and with low ϵ_s. Figure 3(c) is the boundary corrected segmentation when fitting the thalamus, disregarding other structures and with higher ϵ_s. Figure 3(d) is the thalamus segmentation when including the conditional shape prior, left thalamus given the left putamen and with low ϵ_s.

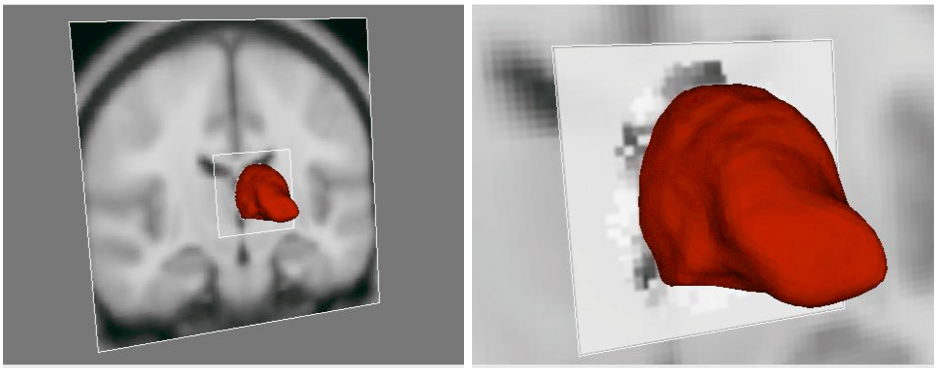
Figures 2(b) and 3(b) show an example of the improved robustness achieved through the incorporation of a conditional shape prior. The poor fit to subject 40 using the single thalamus model is corrected, without significant difference to the rest of the fitting, when using the left putamen as an additional constraint. This would suggest a structural hierarchy across structures using conditional priors would lead to increased robustness. In the case of



(a) -3σ

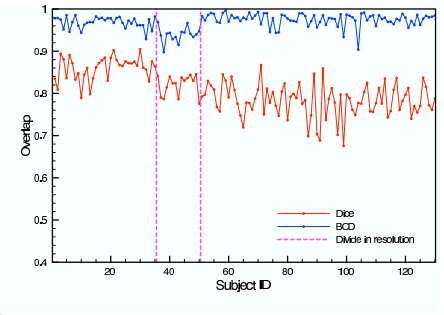


(b) mean

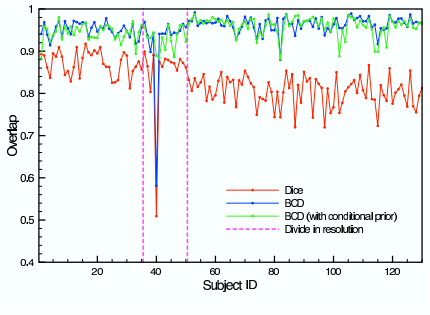


(c) $+3\sigma$

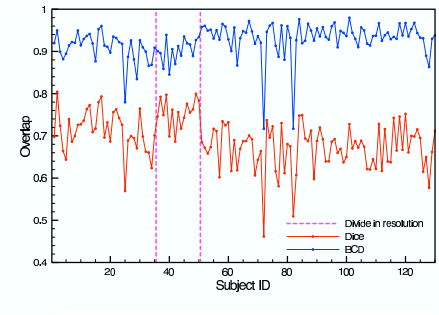
Figure 1: First mode of variation for the left thalamus. The first column shows the thalamus surface overlaid on the MNI 152 template. The second column is a zoomed in view, with the conditional mean overlaid in the square patch. The enlarging dark band of intensities at the thalamus border represent the enlarging ventricle that correlates with the translation and shape change seen in the thalamus.



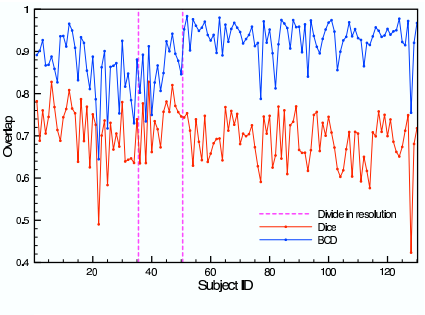
(a) Left Putamen



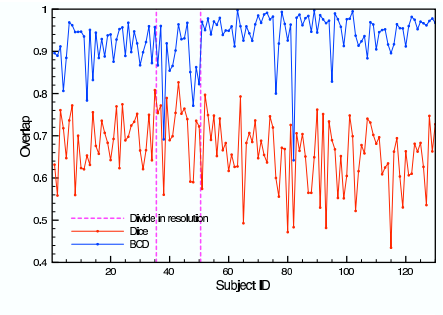
(b) Left Thalamus



(c) Left Hippocampus



(d) Left Amygdala



(e) Left Accumbens

Figure 2: Leave-10-out overlap results using 20 modes of variation and ϵ_I and ϵ_s equal to $1 \times 10^{-7}\%$ of the total shape and intensity variance respectively. The vertical dashed lines are the divisions between different resolution. Subjects 1 to 37, 38 to 50, and 51 to 120 are at $1.5mm^3$, $1mm^3$, and $2.56mm^3$ respectively.

Table 1: Mean ±1 standard deviation of BCD overlap for the left thalamus as a function of shape and intensity prior ϵ_I and ϵ_s. The thalamus was fit to the 120 datasets using the leave-10-out method.

$\epsilon_s(\%)$	$\epsilon_I(\%)$	mean (BCD)	std (BCD)
10^{-7}	10^{-7}	0.956	0.0379
10^{-4}	10^{-4}	0.959	0.0176
10	10	0.942	0.0414
10	10^{-7}	0.947	0.0424
10^{-7}	10	0.947	0.0279

the thalamus given the putamen, we make use of a structure where the segmentation is less sensitive to pathology, to inform one that is more sensitive.

Figure 4 shows the mean overlap ±1 standard deviation for the one group of 9 that was excluded from all models. There is very little variation across models, this is indicative that in practice we have an adequate amount of training data, given that randomly leaving ten out of the model has little impact on the actual fitting.

Table 1 shows the effect of the intensity and shape error prior, ϵ_I and ϵ_s, for the left thalamus. The fitting is fairly insensitive to variation in ϵ_I and ϵ_s (though there is a small peak with reduced variance); hence the conditioning of the matrix does not come at an apparent cost and the choice of value is not critical.

For higher values of the ϵ_s, the majority of subjects tend to have lower overlaps, however, in subject 40 the overlap was significantly increased, as depicted in 3(c). Subject 40 is an extreme pathology, and reflects the type of extra variation that is not included in the sample covariance (from the reduced training set) that we have added through the shape prior. The influence of the error priors are much more complex, as they also effect the conditional distributions, and hence the conditional shape priors as well as the appearance model.

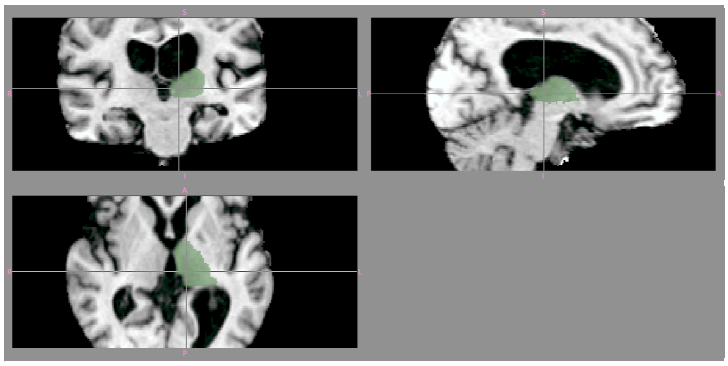
Figure 5 shows the conditional mean left lateral ventricle for ages 22, 53 and 84; as is expect it is increasing with age. Figure 6 shows the predicted mean volume given age; different rates of atrophy can be seen for different structures. Feasibly, by incorporating age into the fitting scheme we could improve robustness by predicting a more accurate mean and covariance for that particular subject's age than the population mean and covariance.

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. 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 prior is used for 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. 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.

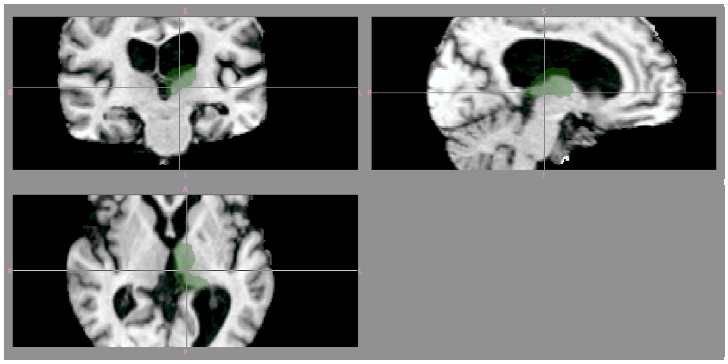
We solve the highlighted problems of the AAM by posing the appearance model our Bayesian framework, where we explicitly account for the lack of training data by the addition of a prior. By conditioning the matrix we allow for the calculation of conditional distributions. The appearance model is considered as the conditional distribution of intensity given shape; the analytic form takes into account the scaling between shape and intensity, hence eliminating any empirical weighting. Furthermore, by modelling the appearance model as a conditional distribution, the conditional covariance weights the intensity samples by the uncertainty.

When fitting we can 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, it becomes straightforward to include other shapes as priors into the fitting. We can, therefore, make use of more robust and accurate structures to inform the less robust. Furthermore, when maximizing the posterior, there is no arbitrary weightings between the conditional and the learned shape and intensity priors. The model utilizes more information from our training data than the AAM, in that instead of providing a maximum-likelihood estimate of intensity given a shape deformation, the entire conditional distribution is modelled. The framework is sufficiently general that data other than shape and intensity can be easily incorporated into the model.

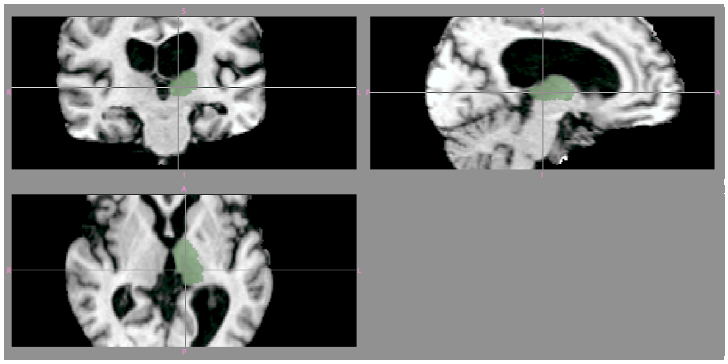
For a single structure shape model, the results are similar to the ASM, the main difference being the addition of a prior that effectively broadens the posterior distribution. The Bayesian appearance model uses the conditionals to predict the intensity distribution from shape. When fitting to new data, the posterior probability of shape given some observed intensities is maximized. The posterior makes use of the prior shape and intensity distribution as well as the conditional. The Bayesian appearance model eliminates the need to retrospectively learn a set of empirical weightings from the training data which relates intensity to shape. Furthermore, the conditional covariance effectively weights the importance of intensity samples by the uncertainty from the distribution; the AAM uses a



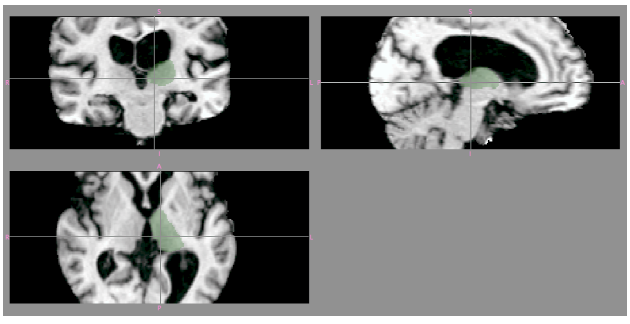
(a) Manual segmentation



(b) Left Thalamus with $\epsilon_s = 1 \times 10^{-7}\%$

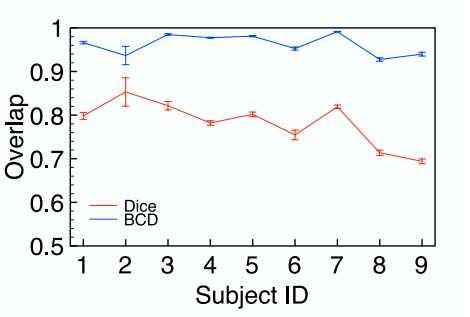


(c) Left Thalamus with $\epsilon_s = 0.001\%$

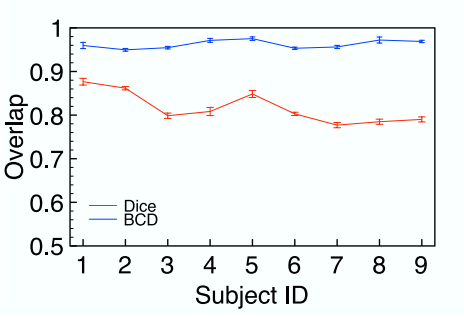


(d) Left Thalamus Given Left Putamen with $\epsilon_s = 1 \times 10^{-7}\%$

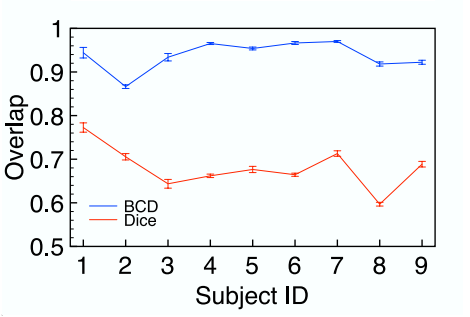
Figure 3: Single subject (40) example of the left thalamus boundary corrected segmentation with low and high ϵ_s and with and without the conditional shape prior.



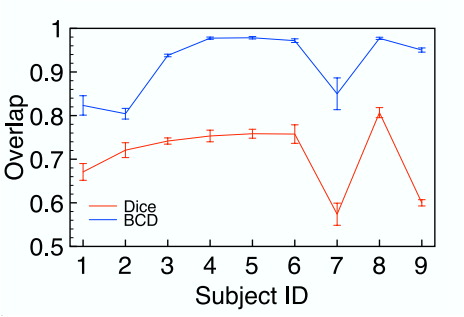
(a) Left Putamen



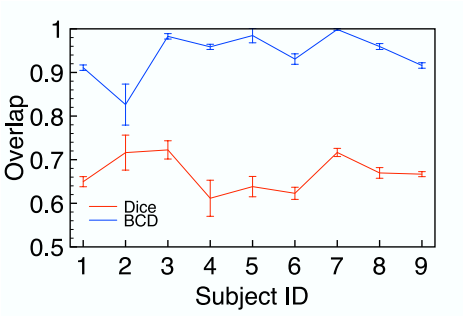
(b) Left Thalamus



(c) Left Hippocampus

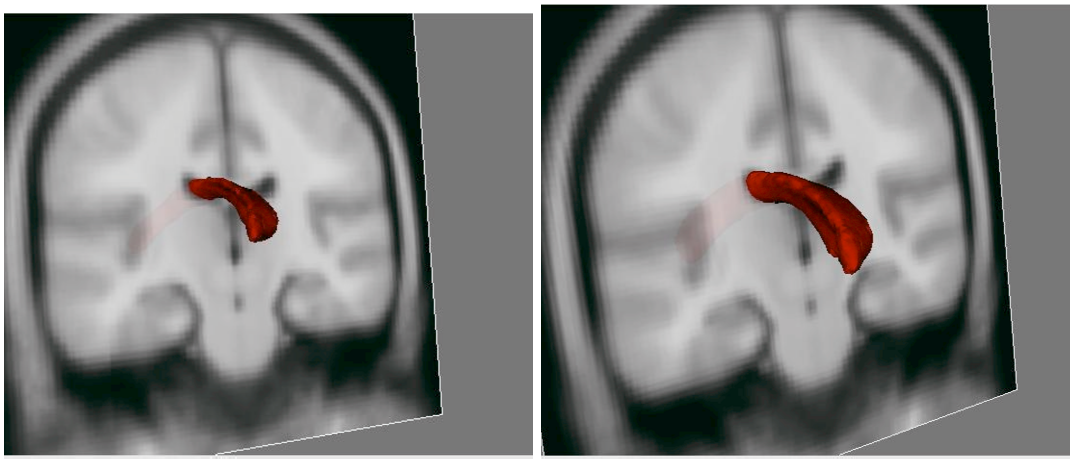


(d) Left Amygdala



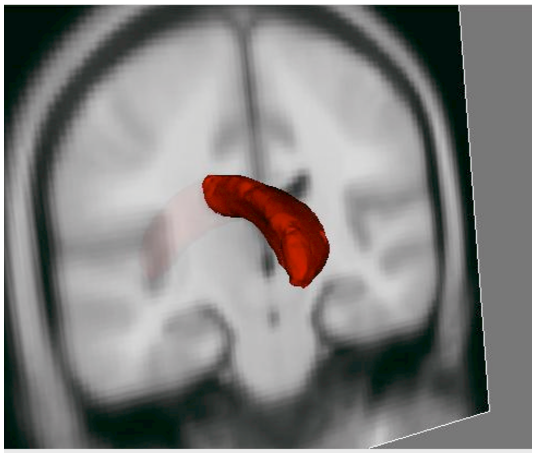
(e) Left Accumbens

Figure 4: Mean overlap ±1 standard deviation for the excluded group of 9.



(a) Age=22

(b) Age=53



(c) Age=84

Figure 5: Conditional mean given age for the left lateral ventricle

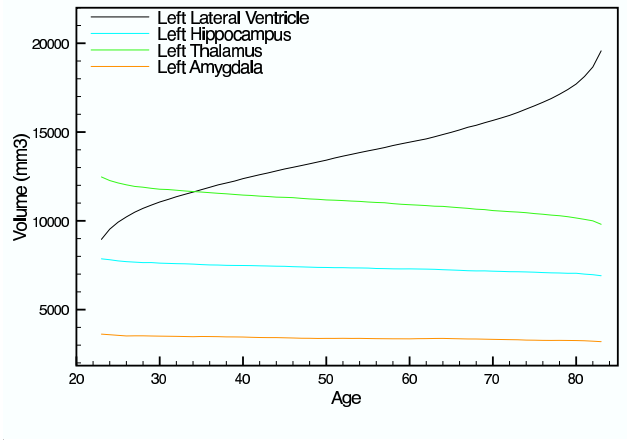


Figure 6: Predicted mean volume given age.

least squares fit to the data.

From a practical viewpoint, the prior added to the covariance matrix improves the conditioning of the sample covariance, allowing the inverse to be calculated. The inverse is required to evaluate the conditional mean and covariance. By expressing the conditional intensity mean as a mode of variation, we can calculate the conditional mean as a linear combination of mode vectors rather than calculating large matrix multiplications. 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 $n_s \times n_s$; this is practically very important as the dimensionality can become very large in 3D.

In summary, advantages of the Bayesian appearance model are: 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; 3) can use the posterior to fit the model, this incorporates shape and intensity priors with the appearance mode without need of arbitrary weighting between them; 4) extends well to incorporating other shapes as priors, not only providing a predicted most-likely guess but also the predicted covariation; and 4) can extend beyond shape and intensity such that other metrics can be incorporated into the model. The disadvantage of the framework is the arbitrary choice of the prior ϵ; though it has been shown that it does not have much impact on the overall fitting. Furthermore, ϵ has some real interpretability as it represents shape or intensity variance. The addition of ϵ provides a means by which we can generalize our models to a larger population, rather than limiting the model to the sampled population.

In the future we wish to further investigate the effects of ϵ on the conditional shape distribution, particularly its effect on the conditional shape priors. ϵ could potentially provide a means to relax the shape priors; this is particularly desirable if we do not have enough data to accurately model all the inter-structure variation in the population. Furthermore, we will be investigating the incorporation of other data such as age, gender, handedness, and pathology into our model and fitting process.

APPENDIX

A Computational Simplifications

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

$$\tilde{z}_1 = U_1 D_1 V_1^T \tag{38}$$

$$\tilde{z}_2 = U_2 D_2 V_2^T \tag{39}$$

We will now express the partitioned covariance and cross-covariance matrices in terms of (38) and (39)

$$\begin{aligned}\Sigma_{11} &= U_1(D_1^2 + 2\epsilon_1^2 I)(n_s-1)^{-1}U_1^T \\ &= U_1 D_{\epsilon_1}^2 U_1^T (n_s-1)^{-1}\end{aligned} \tag{40a}$$

$$\Sigma_{22} = U_2 D_{\epsilon_2}^2 U_2^T (n_s-1)^{-1} \tag{40b}$$

$$\begin{aligned}\Sigma_{12} &= \Sigma_{21}^T \\ &= \tilde{z}_1 \tilde{z}_2^T (n_s-1)^{-1} \\ &= \tilde{z}_1 V_2 D_2^T U_2^T (n_s-1)^{-1}\end{aligned} \tag{40c}$$

Rearranging (27), such that

$$(x_I - \mu_I) = U_I \frac{D_{\epsilon_I}}{\sqrt{(n_s-1)\gamma_v}} b_I \tag{41}$$

where i is the i^{th} partition.

A.1 Conditional Mean as a Mode of Variation

Substituting (40a), (40b), (40c), and (41) into (23b),

$$\begin{aligned}\mu_{1|2} = \mu_1 + x_1 V_2 \left[\begin{smallmatrix} D_{2,s} \\ 0 \end{smallmatrix}\right] U_2^T U_2 \left[\begin{smallmatrix} D_{\epsilon_2,s}^{-2} & 0 \\ 0 & \frac{1}{2}\epsilon_2^{-2} I \end{smallmatrix}\right] U_2^T \\ U_2 \left[\begin{smallmatrix} D_{\epsilon_2,s} & 0 \\ 0 & \sqrt{2}\epsilon_2 I \end{smallmatrix}\right] \sqrt{\frac{\gamma_v}{n_s - 1}} b_2\end{aligned} \tag{42}$$

$$\mu_{1|2} = \mu_1 + z_{1dm} \left[\begin{smallmatrix} (V_2 D_{2,s} D_{\epsilon_2,s}^{-1} \sqrt{\frac{\gamma_v}{n_s-1}} b_{2,s}) + 0 \end{smallmatrix} \right]$$

where here $b_{2,s}$ is the upper-left $n_s \times 1$ submatrix of b_2.

$$\mu_{1|2} = \mu_1 + \tilde{z_1}[V_2 D_{2,s} D_{\epsilon_2,s}^{-1} \sqrt{\frac{\gamma_v}{n_s - 1}} b_{2,s}] \tag{43}$$

All matrices within square brackets are of size $n_s \times n_s$ except $b_{2,s}$ which is $n_s \times 1$. If truncating modes at L, only the first L columns of $z_{1dm}[V_2 D_{2,s} D_{\epsilon_2,s}^{-1} \sqrt{\frac{\gamma_v}{n_s-1}}]$ are needed.

A.2 Simplifying Conditional Covariance Operations

We will here define

$$\begin{aligned}\Sigma_{1|2} &= \Sigma_{c11}[\frac{\alpha + (x_2 - \bar{x_2})^T \Sigma_{22}^{-1} (x_2 - \bar{x_2})}{\alpha + k_2}] \\ &= U_{1|2} D_{1|2}^2 U_{1|2}^T\end{aligned} \tag{44}$$

where $U_{1|2}$ are the eigenvectors, and $D_{1|2}^2$ is a diagonal matrix of the eigenvalues.
For notational convenience we will also define

$$\lambda_{c11}^{-1} = \Sigma_{c11} = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \tag{45a}$$

such that

$$U_{1|2} = U_{c11} \tag{45b}$$

$$D_{1|2}^2 = D_{c11}^2 [\frac{\alpha + (x_2 - \bar{x_2})^T \Sigma_{22}^{-1} (x_2 - \bar{x_2})}{\alpha + k_2}] \tag{45c}$$

A.2.1 Simplifying $(x_2 - \mu_2)^T \Sigma_{22}^{-1} (x_2 - \mu_2)$

$$\begin{aligned}&(x_2 - \mu_2)^T \Sigma_{22}^{-1} (x_2 - \mu_2) = \\ &= b_2^T \sqrt{\frac{\gamma_v}{n_s - 1}} D_{\epsilon_2} U_2^T U_2 D_{\epsilon_2}^{-2} U_2^T (n_s - 1) U_2 D_{\epsilon_2} \sqrt{\frac{\gamma_v}{n_s - 1}} b_2 \\ &= b_2^T b_2 \gamma_v\end{aligned} \tag{46}$$

A.2.2 Simplifying Σ_{c11}

$$\begin{aligned}\Sigma_{c11} &= \Sigma_{11} - \Sigma_{21} \Sigma_{22}^{-1} \Sigma_{12} \\ &= U_1 D_{\epsilon_1}^2 U_1^T (\frac{1}{n_s - 1}) - (\frac{1}{n_s - 1}) U_1 D_1 V_1^T \\ &V_2 D_2^T D_{\epsilon_2}^{-2} (n_s - 1) D_2 V_2^T V_1 D_1^T U_1^T (\frac{1}{n_s - 1}) \\ &= U_1 (D_{\epsilon_1}^2 - D_1 V_1^T V_2 D_2^T D_{\epsilon_2}^{-2} D_2 V_2^T V_1 D_1^T) U_1^T (\frac{1}{n_s - 1})\end{aligned} \tag{47}$$

We now define

$$\Sigma_{c11} = U_1 \Sigma_{c11V} U_1^T (\frac{1}{n_s - 1}) \tag{48}$$

such that,

$$\begin{aligned}\Sigma_{c11V} &= D_{\epsilon_1}^2 - D_1 V_1^T V_2 D_2^T D_{\epsilon_2}^{-2} D_2 V_2^T V_1 D_1^T \\ &= \left[\begin{smallmatrix} (D_{\epsilon_1,s}^2 + 2\epsilon_1^2 I & 0 \\ 0 & 2\epsilon_1^2 I \end{smallmatrix}\right] \\ &- \left[\begin{smallmatrix} D_{1,s} \\ 0 \end{smallmatrix}\right] V_1^T V_2 \left[\begin{smallmatrix} D_{2,s} & 0 \end{smallmatrix}\right] \left[\begin{smallmatrix} D_{\epsilon_2,s}^{-2} & 0 \\ 0 & \frac{1}{2}\epsilon_2^{-2} I \end{smallmatrix}\right] \left[\begin{smallmatrix} D_{2,s} \\ 0 \end{smallmatrix}\right] V_2^T V_1 \left[\begin{smallmatrix} D_{1,s} & 0 \end{smallmatrix}\right] \\ &= \left[\begin{smallmatrix} (D_{1,s}^2 + 2\epsilon_1^2 I - D_{1,s} V_1^T V_2 D_{2,s} D_{\epsilon_2,s}^{-2} D_{2,s} V_2^T V_1 D_{1,s}) & 0 \\ 0 & 2\epsilon_1^2 I \end{smallmatrix}\right]\end{aligned} \tag{49}$$

$$\Sigma_{c11V} = \begin{bmatrix} \Sigma_{c11V,s} & 0 \\ 0 & 2\epsilon_1^2 I \end{bmatrix} \tag{50}$$

Express $\Sigma_{c11V,s}$ in terms of its SVD expansion.

$$\Sigma_{c11V,s} = U_{c11V,s} D_{c11V,s}^2 U_{c11V,s}^T \tag{51}$$

Note that $U_{c11V,s}$ and $D_{c11V,s}$ are $n \times n$ matrices.
It now follows that

$$\begin{aligned}\Sigma_{c11V} &= U_{c11V} D_{c11V\epsilon_1}^2 U_{c11V}^T \\ &= \left[\begin{smallmatrix} U_{c11V,s} & 0 \\ 0 & I \end{smallmatrix}\right] \left[\begin{smallmatrix} D_{c11V,s}^2 & 0 \\ 0 & 2\epsilon_1^2 I \end{smallmatrix}\right] \left[\begin{smallmatrix} U_{c11V,s}^T & 0 \\ 0 & I \end{smallmatrix}\right]\end{aligned} \tag{52}$$

Now substituting the new expression for Σ_{c11V} back into Σ_{c11}, we get

$$\begin{aligned}\Sigma_{c11} &= U_1 U_{c11V} D_{c11V\epsilon_1} U_{c11V}^T U_1^T \\ &= U_1 \left[\begin{smallmatrix} U_{c11V,s} & 0 \\ 0 & I \end{smallmatrix}\right] \left[\begin{smallmatrix} D_{c11V,s}^2 & 0 \\ 0 & 2\epsilon_1^2 I \end{smallmatrix}\right] \\ &\left[\begin{smallmatrix} U_{c11V,s}^T & 0 \\ 0 & I \end{smallmatrix}\right] U_1^T (n-1)^{-1}\end{aligned} \tag{53}$$

Define

$$U_1 = \left[\begin{smallmatrix} U_{1,s} & U_{1,s_2} \end{smallmatrix}\right] \tag{54}$$

where $U_{1,s}$ and U_{1,s_2} are $k_1 \times n_s$ and $k_1 \times (k_1 - n_s)$ submatrices of U_1 respectively.

$$\begin{aligned}\Sigma_{c11} =& \left[\begin{smallmatrix} U_{1,s} & U_{1,s_2} \end{smallmatrix}\right] \left[\begin{smallmatrix} U_{c11V,s} & 0 \\ 0 & I \end{smallmatrix}\right] \left[\begin{smallmatrix} D_{c11V,s}^2 & 0 \\ 0 & 2\epsilon_1^2 I \end{smallmatrix}\right] \\ & \left[\begin{smallmatrix} U_{c11V,s}^T & 0 \\ 0 & I \end{smallmatrix}\right] \left[\begin{smallmatrix} U_{1,s} & U_{1,s_2} \end{smallmatrix}\right]^T (n-1)^{-1} \\ =& \left[\begin{smallmatrix} U_{1,s} U_{c11V,s} & U_{1,s_2} \end{smallmatrix}\right] \left[\begin{smallmatrix} D_{c11V,s}^2 & 0 \\ 0 & 2\epsilon_1^2 I \end{smallmatrix}\right] \\ & \left[\begin{smallmatrix} U_{1,s} U_{c11V,s} & U_{1,s_2} \end{smallmatrix}\right]^T (n_s - 1)^{-1}\end{aligned} \tag{55}$$

From earlier $U_{1|2} = U_{c11}$

$$\begin{aligned}\Sigma_{c11} =& \left[\begin{smallmatrix} U_{1|2,s} & U_{1|2,s_2} \end{smallmatrix}\right] \left[\begin{smallmatrix} D_{c11V,s}^2 & 0 \\ 0 & 2\epsilon_1^2 I \end{smallmatrix}\right] \\ & \left[\begin{smallmatrix} U_{1|2,s} & U_{1|2,s_2} \end{smallmatrix}\right]^T (n_s - 1)^{-1}\end{aligned} \tag{56}$$

$$\begin{aligned}V_{1|2} =& \Sigma_{1|2} \frac{(n_s - \frac{1}{n_s} + k_2)}{(n_s - \frac{1}{n_s} + k_2 - 2)} \\ =& \left[\begin{smallmatrix} U_{1|2,s} & U_{1|2,s_2} \end{smallmatrix}\right] \left[\begin{smallmatrix} D_{c11V,s}^2 & 0 \\ 0 & 2\epsilon_1^2 I \end{smallmatrix}\right] \left[\begin{smallmatrix} U_{1|2,s} & U_{1|2,s_2} \end{smallmatrix}\right]^T \\ & \frac{(\alpha + k_2)(\alpha + (x_2 - \bar{x}_2)^T \Sigma_{22}^{-1} (x_2 - \bar{x}_2))}{(\alpha + k_2 - 2)(n_s - 1)(\alpha + k_2)} \\ =& \left[\begin{smallmatrix} U_{1|2,s} & U_{1|2,s_2} \end{smallmatrix}\right] \left[\begin{smallmatrix} D_{c11V,s}^2 & 0 \\ 0 & 2\epsilon_1^2 I \end{smallmatrix}\right] \left[\begin{smallmatrix} U_{1|2,s} & U_{1|2,s_2} \end{smallmatrix}\right]^T \\ & \frac{(\alpha + (x_2 - \bar{x_2})^T \Sigma_{22}^{-1} (x_2 - \bar{x_2}))}{(\alpha + k_2 - 2)(n_s - 1)}\end{aligned} \tag{57}$$

where $\alpha = n_s - \frac{1}{n_s}$

The first n eigenvectors (order by eigenvalues) of the conditional covariance is $U_{1,s}U_{c11V,s}$, which are of dimensions $k_1 \times n_s$ and $n_s \times n_s$ respectively. The eigenvectors greater than n_s are U_{1,s_2}. To arrive at the expression for $U_{1|2}$ 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 $\tilde{z_1}$ and $\tilde{z_2}$.

A.2.3 Simplifying the calculation of $(x_1 - \mu_1)^T \Sigma_{1|2}^{-1}(x_1 - \mu_1)$

It is worth noting that this calculation would typically be done at run time, so it is important to simplify. $(x_1 - \mu_1)$ is a $k_1 \times 1$ matrix. It is straight forward to show that

$$\Sigma_{c11}^{-1} = \begin{bmatrix} U_{1|2,s} & U_{1|2,s_2} \end{bmatrix} \begin{bmatrix} D_{c11V,s}^{-2} & 0 \\ 0 & \frac{1}{2}\epsilon_1^{-2} I \end{bmatrix} \begin{bmatrix} U_{1|2,s} & U_{1|2,s_2} \end{bmatrix}^T (n-1) \quad (58)$$

$$\begin{aligned}
&(x_1-\mu_1)^T \Sigma_{1|2}^{-1}(x_1 - \mu_1) \\
&=(x_1 - \mu_1)^T \begin{bmatrix} U_{1|2,s} & U_{1|2,s_2} \end{bmatrix} \begin{bmatrix} D_{c11V,s}^{-2} & 0 \\ 0 & \frac{1}{2}\epsilon_1^{-2} I \end{bmatrix} \\
&\quad \begin{bmatrix} U_{1|2,s} & U_{1|2,s_2} \end{bmatrix}^T (x_1 - \mu_1)(n_s - 1) \\
&= \begin{bmatrix} (x_1-\mu_1^T)U_{1|2,s} & (x_1-\mu_1)^T U_{1|2,s_2} \end{bmatrix} \begin{bmatrix} D_{c11V,s}^{-2} & 0 \\ 0 & \frac{1}{2}\epsilon_1^{-2} I \end{bmatrix} \\
&\quad \begin{bmatrix} (x_1-\mu_1)^T U_{1|2,s} & (x_1-\mu_1)^T U_{1|2,s_2} \end{bmatrix}^T (n_s - 1) \\
&=(x_1 - \mu_1)^T U_{1|2,s} D_{c11V,s}^{-2} U_{1|2,s}^T (x_1 - \mu_1) \\
&\quad + \frac{1}{2}\epsilon_1^{-2}(x_1 - \mu_1)^T U_{1|2,s_2} U_{1|2,s_2}^T (x_1 - \mu_1))(n_s - 1) \\
&=[(x_1 - \mu_1)^T U_{1|2,s} D_{c11V,s}^{-2} U_{1|2,s}^T (x_1 - \mu_1) \\
&\quad + \frac{1}{2}\epsilon_1^{-2}(x_1 - \mu_1)^T (x_1 - \mu_1)](n_s - 1)
\end{aligned} \quad (59)$$

Note the dimensions of the matrices. $(x_1 - \mu_1)$ is $k_1 \times n_s$, $U_{1|2,s}$ is $k_1 \times n_s$, and $D_{c11V,s}^{-2}$ 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.

B Simplification of the Posterior

The full expression we wish to maximize is,

$$\ln p(x_s \mid x_I, z = \ln p(x_I \mid x_s, z) + \ln p(x_s \mid z) - \ln p(x_I \mid z) \quad (60)$$

The full expression for $\ln p(x_I \mid x_s, z)$ is,

$$\begin{aligned}
&\ln p(x_I \mid x_s, z) = \ln \left(\frac{\Gamma(\frac{1}{2}(\alpha_{I|s} + k_I))}{\Gamma(\frac{1}{2}\alpha_{I|s})(\alpha_{I|s}\pi)^{k_I/2}} \left|\lambda_{I|s}\right|^{1/2} \right) \\
&+ \ln \left(\left[1 + \frac{1}{\alpha_{I|s}}(x_I - \mu_{I|s})^t \lambda_{I|s}(x_I - \mu_{I|s}) \right]^{\frac{-(\alpha_{I|s}+k_I)}{2}} \right)
\end{aligned} \quad (61)$$

where $\alpha_{I|s}$ is the degree of freedom form the conditional distribution of x_I given x_s. Now simplifying the expression,

$$\begin{aligned}
&\ln p(x_I \mid x_s, z) = \ln\left(\frac{\Gamma(\frac{1}{2}(\alpha_{I|s}+k_I))}{\Gamma(\frac{1}{2}\alpha_{I|s})(\alpha_{I|s}\pi)^{k_I/2}}\right) + \ln\left(\left|\lambda_{I|s}\right|^{1/2}\right) \\
&-\frac{(\alpha_{I|s}+k_I)}{2}\ln\left(1+\frac{1}{\alpha_{I|s}}(x_I-\mu_{I|s})^t\lambda_{I|s}(x_I-\mu_{I|s})\right) \\
&= C+\frac{1}{2}\ln\left|\lambda_{I|s}\right| \\
&-\frac{(\alpha_{I|s}+k_I)}{2}\ln\left(1+\frac{1}{\alpha_{I|s}}(x_I-\mu_{I|s})^t\lambda_{I|s}(x_I-\mu_{I|s})\right) \\
&= C+\frac{1}{2}\ln\left|\lambda_{cII}\left(\frac{\alpha_{I,s}+k_s}{\alpha_{I,s}+b_s^Tb_s\gamma_v}\right)\right| - \frac{(\alpha_{I,s}+k_s+k_I)}{2} \\
&\ln(1+\frac{1}{\alpha_{I,s}+k_s} \\
&(x_I-\mu_{I|s})^t\lambda_{cII}\left(\frac{\alpha_{I,s}+k_s}{\alpha_{I,s}+b_s^Tb_s\gamma_v}\right)(x_I-\mu_{I|s})) \\
&= C+\frac{k_I}{2}\ln\left(\frac{\alpha_{I,s}+k_s}{\alpha_{I,s}+b_s^Tb_s\gamma_v}\right) \\
&-\frac{(\alpha_{I,s}+k_s+k_I)}{2} \\
&\ln\left(1+\frac{1}{\alpha_{I,s}+b_s^Tb_s\gamma_v}(x_I-\mu_{I|s})^t\lambda_{cII}(x_I-\mu_{I|s})\right)
\end{aligned} \tag{62}$$

The full expression for $\ln p(x_s \mid z)$ is

$$\begin{aligned}
&\ln p(x_s \mid z) = \ln\left(\frac{\Gamma(\frac{1}{2}(\alpha_{s+k_s}))}{\Gamma(\frac{1}{2}\alpha_s)(\alpha_s\pi)^{k_s/2}}\left|\lambda_s\right|^{1/2}\right) \\
&+\ln\left(\left[1+\frac{1}{\alpha_s}(x_I-\mu_s)^t\lambda_s(x_s-\mu_s)\right]^{\frac{-(\alpha_s+k_s)}{2}}\right) \\
&= C-\frac{(\alpha_s+k_s)}{2}\ln\left(1+\frac{1}{\alpha_s}(x_I-\mu_s)^t\lambda_s(x_s-\mu_s)\right) \\
&= C-\frac{(\alpha_s+k_s)}{2}\ln\left(1+\frac{1}{\alpha_s}b_s^Tb_s\gamma_v\right)
\end{aligned} \tag{63}$$

The expression for $\ln p(x_I \mid z)$ is given by

$$\begin{aligned}
&\ln p(x_I \mid z) = \ln\left(\frac{\Gamma(\frac{1}{2}(\alpha_{s+k_I}))}{\Gamma(\frac{1}{2}\alpha_I)(\alpha_I\pi)^{k_I/2}}\left|\lambda_I\right|^{1/2}\right) \\
&+\ln\left(\left[1+\frac{1}{\alpha_I}(x_I-\mu_I)^t\lambda_I(x_I-\mu_I)\right]^{\frac{-(\alpha_I+k_I)}{2}}\right) \\
&= C-\frac{(\alpha_I+k_I)}{2}\ln\left(1+\frac{1}{\alpha_I}(x_I-\mu_I)^t\lambda_I(x_I-\mu_I)\right)
\end{aligned} \tag{64}$$

By substituting (62), (63), and (64) into (60), we get an expression for the full posterior given by

$$\begin{aligned}
=&C + \frac{k_I}{2}\ln\left(\frac{\alpha_{I,s}+k_s}{\alpha_{I,s}+b_s^T b_s \gamma_v}\right) \\
&- \frac{(\alpha_{I,s}+k_s+k_I)}{2} \\
&\ln\left(1+\frac{1}{\alpha_{I,s}+b_s^T b_s\gamma_v}(x_I-\mu_{I|s})^t\lambda_{cII}(x_I-\mu_{I|s})\right) \\
&+ \frac{(\alpha_s+k_s)}{2}\ln\left(1+\frac{1}{\alpha_s}b_s^T b_s\gamma_v\right) \\
&- \frac{(\alpha_I+k_I)}{2}\ln\left(1+\frac{1}{\alpha_I}(x_I-\mu_I)^t\lambda_I(x_I-\mu_I)\right)
\end{aligned} \tag{65}$$

C Model Registration

A x has a multivariate Student distribution, $St_k(x,\mu,\lambda,\alpha)$, and $y = Ax$ such that A is an $m \times k$ matrix of real numbers such that $m \leq k$ and $A\lambda^{-1}A^t$ is non-singular, then y has a distribution given by $St_k(x, A\mu, (A\lambda^{-1}A^t)^{-1}, \alpha)$ [1]. Expressing λ in terms of its eigenvalues and eigenvectors,

$$\lambda = UD^{-2}U^t \tag{66}$$

It follows that the precision matrix for y is given by

$$\begin{aligned}
(A\lambda^{-1}A^t)^{-1} &= (AUD^2U^tA^t)^{-1} \\
&= (AU)D^{-2}(UA)^t
\end{aligned} \tag{67}$$

AU being the registered eigenvectors; the original eigenvalues remain unchanged.

D Calculation of Conditional Shape Prior

Analogous to appendix A.2.1, it can be shown that

$$(x_{1|2}-\mu_{1|2})^T\Sigma_{1|2}^{-1}(x_{1|2}-\mu_{1|2}) = b_{1|2}^T b_{1|2}\gamma_v \tag{68}$$

In our application we are search for the optimal value of b_1 our current shape model, and we know b_2 from our previous fit. Two $n_s \times n_s$ matrices can be calculated that can map b_1 and b_2 to $b_{1|2}$; given $b_{1|2}$ we only need to calculate the inner-product of the vector.

We parameterize the shape x_1 in terms of its conditional distribution, $p(x_1 \mid x_2)$, given by

$$\begin{aligned}
x_1 =&\mu_{1|2} + \\
&U_{1|2}D_{\epsilon_{c11}}\sqrt{\frac{\alpha+b_2^T b_2\gamma_v}{\alpha+k_2}}\sqrt{\frac{\gamma_v}{n_s-1}}b_{c1}
\end{aligned} \tag{69}$$

Parameterizing the shape x_1 in terms of $p(x_1)$, we obtain

$$x_1 = \mu_1 + U_1 D_{\epsilon_1}\sqrt{\frac{\gamma_v}{n_s-1}}b_1 \tag{70}$$

γ_v is only a function of the number in the training set.

Equating 69 and 70, we get

$$\begin{aligned}
&\mu_1 + U_1 D_{\epsilon_1}\sqrt{\frac{\gamma_v}{n-1}}b_1 = \mu_{1|2} \\
&\quad + U_{1|2}D_{\epsilon_{c11}V}\sqrt{\frac{\alpha+b_{x_2}^T b_{x_2}\gamma_v}{\alpha+k_2}}\sqrt{\frac{\gamma_v}{n-1}}b_{c1}
\end{aligned} \tag{71}$$

From earlier, $U_{1|2} = U_1 U_{c11V}$ and
$\mu_{1|2} = \mu_1 + \tilde{z_1} V_2 D_2 D_{\epsilon_2}^{-1} \sqrt{\frac{\gamma_v}{n-1}} b_2$.

Now substituting into 71,

$$\begin{aligned} \mu_1 + U_1 D_{\epsilon_1} \sqrt{\frac{\gamma_v}{n_s - 1}} b_1 = \mu_1 + \tilde{z_1} V_2 D_2 D_{\epsilon_2}^{-1} \sqrt{\frac{\gamma_v}{n_s - 1}} b_2 \\ + U_1 U_{c11V} D_{\epsilon_{c11V}} \sqrt{\frac{\alpha + b_2^t b_2 \gamma_v}{\alpha + k_2}} \sqrt{\frac{\gamma_v}{n_s - 1}} b_{c1} \end{aligned} \tag{72}$$

Rearranging,

$$\begin{aligned} b_{c1} = & \sqrt{\frac{\alpha + k_2}{\alpha + b_2^T b_2 \gamma_v}} \\ & \left(D_{\epsilon_{c11V}}^{-1} U_{c11V}^t D_{\epsilon_1} b_1 - D_{\epsilon_{c11V}}^{-1} U_{c11V}^t D_1 V_1^t V_2 D_2 D_{\epsilon_2}^{-1} b_2 \right) \end{aligned} \tag{73}$$

Since all the singular values of $\tilde{z_1}$ and $\tilde{z_2}$ above n_s are zero, we can simplify to get

$$\begin{aligned} b_{c1} = & \sqrt{\frac{\alpha + k_2}{\alpha + b_2^T b_2 \gamma_v}} \\ & (D_{\epsilon_{c11V,s}}^{-1} U_{c11V,s}^t D_{\epsilon_1} b_{1,s} \\ & - D_{\epsilon_{c11V},s}^{-1} U_{c11V,s}^t D_1 V_1^t V_2 D_2 D_{\epsilon_2,s}^{-1} b_{2,s}) \end{aligned} \tag{74}$$

E Acknowledgment

The authors would like to thank the ESPRC and the IBIM grant for funding as well as the BBSRC. We would also like to thank Christian Haselgrove at the Centre for Morphometric Analysis for helping retrieve the Data. We would also like to thanks Tim Cootes for his valuable input.

References

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

[2] C.S. Bretherton, C. Smith, and J.M. Wallace. An intercomparison of methods for finding coupled patterns in climate data. *Journal of Climate*, 5:541–560, June 1992.

[3] 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.

[4] Olivier Colliot, Oscar Camara, and Isabelle Bloch. Integration of fuzzy spatial relations in deformable models—application to brain mri segmentation. *Pattern Recognition*, 39(1401-1414), 2006.

[5] 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.

[6] 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.

[7] 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).

[8] 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.

[9] 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.

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

[11] S. Lobregt and M.A. Viergever. A discrete dynamic contour model. *IEEE Transactions on Medical Imaging*, 14(1):12–24, March 1995.

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

[13] 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.

[14] 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.

[15] Anil Rao, Tim Cootes, and Daniel Rueckert. Hierarchical statistical shape analysis and prediction of subcortical brain structures. In *2006 Conference on Computer Vision and Pattern Recognition Workshop*, 2006.

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

[17] 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.