Shape Analysis and Statistical Modelling in Brain Imaging

Christopher J. Brignell, MMath

Division of Statistics
School of Mathematical Sciences

Thesis submitted to the University of Nottingham
for the degree of Doctor of Philosophy

March 2007

Contents

Abstract

This thesis considers the registration of shapes, estimation of shape variability and the statistical modelling of human brain magnetic resonance images (MRI). Current shape registration techniques, such as Procrustes analysis, superimpose shapes in order to make inferences regarding the mean shape and shape variability. We apply Procrustes analysis to a subset of the landmarks and give distributional results for the Euclidean distance of a shape from a template. Procrustes analysis is then generalised to minimise a Mahalanobis norm, with respect to a symmetric, positive definite matrix, and the weighted Procrustes estimators for scaling, rotation and translation obtained. This weighted registration criterion is shown, through a simulation study, to reduce the bias and error in maximum likelihood estimates of the mean shape and covariance matrix compared to isotropic Procrustes. A Bayesian Markov chain Monte Carlo algorithm is also presented and shown to be less sensitive to prior information.

We consider two MRI data sets in detail. We examine the first data set for large-scale shape differences between two volunteer groups, healthy controls and schizophrenia patients. The images are registered to a template through modelling the voxel values and we maximise the likelihood over the transformation parameters. Using a suitable labelling and principal components analysis we show schizophrenia patients have less brain asymmetry than healthy controls. The second data set is a sequence of functional MRI scans of an individual's motor cortex taken while they repeatedly press a button. We construct a model with temporal correlations to estimate the trial-to-trial variability in the haemodynamic response using the Expectation-Maximisation algorithm. The response is shown to change with task and through time. For both data sets we compare our techniques with existing software packages and improvements to data pre-processing are suggested. We conclude by discussing potential areas for future research.

List of Figures

List of Tables

Acknowledgements

Firstly, I would like to thank my supervisors, Ian Dryden and Bill Browne, for their encouragement and advice throughout this period of study. Their kindness and support has made this work possible. I would also like to acknowledge my examiners, Merrilee Hurn and Neil Butler, for their suggestions that have improved this work.

In addition, I would like to thank Sean Flynn (University of British Columbia), Bert Park and Stuart Leask (School of Community Health Sciences), and Sue Francis and Peter Wright (Sir Peter Mansfield Magnetic Resonance Centre), for motivating the applications and providing the raw data.

I am also grateful to the staff and research students in the School of Mathematical Sciences for making the University of Nottingham an enjoyable place to work. In particular, I would like to thank my colleagues, Kim Evans, Kelly Handley and Simon Spencer, for their friendship.

Lastly, thanks must go to my family. I am heavily indebted to my parents and sister for enabling me to achieve my potential throughout twenty years of education.

I would like to dedicate this thesis to my wife, Jenny, for her continuous love and support throughout the past three years.

This research was funded by an EPSRC doctoral training account.

> *To whom, then, will you compare God? What image will you compare him to?*
>
> Isaiah 40:18

Chapter 1

Introduction and background

1.1 Introduction

The research presented in this thesis focuses on problems relating to the analysis of shapes. Shape analysis has already proved to be a valuable tool in a wide range of disciplines, such as biology, medicine, archaeology, geography, geology and genetics. Traditionally shape analysis has been conducted by examining ratios of distances. For example, a biologist might use the ratio of a particular bone's length over its width to distinguish between species. However, in selecting certain measurements other information regarding the object's shape is being lost.

More recently, geometrical methods making use of the co-ordinates of an object's distinctive features have been developed. New technology has made the recording of co-ordinates easier and presents new challenges and applications with the analysis of objects captured in digital images. Inference, regarding a mean shape and/or shape variability, is normally made after superimposing the shapes or images on top of each other. However, the most common form of registering shapes to make comparisons places equal importance on each part of the shape. The development of a weighted superimposition method, necessary for brain registration, has received several

contributions (e.g. Goodall, 1991), but a universally applicable method for jointly estimating shape variability and using shape variability to weight the superimposition, has been missing. The first half of this thesis proposes two methods for achieving this. The second half considers practical applications to the human brain, motivated by problems encountered by researchers in neurology. The two topics are related, as the human brain is highly variable in shape, and placing different brain images into a frame of reference for comparison is non-trivial.

For neurological applications, we consider two data sets in detail. The first data set consists of human brain images taken from controls and patients with schizophrenia, with the aim of establishing if there is a statistically significant difference in brain shape between the two groups. Research into schizophrenia's impact on the human brain is well established, as clinicians seek to develop a more robust method for diagnosing the disease. It is believed that anatomical shape analysis will aid earlier diagnosis, which will lead to better treatment.

Secondly, we consider images of one individual's motor cortex taken sequentially over time. The motor cortex is an area of grey matter, towards the posterior of the brain, that is responsible for controlling muscle movements. The aim is to build a statistical model for brain activation in response to a stimuli repeated over a period of time. Brain activation is also highly variable, and establishing how the brain's response changes to each identical stimuli is an important step to understanding the wider problem of how the different areas of the brain divide tasks and interact. Although we develop methods with these data sets in mind, the methods are directly applicable to other diseases and brain regions.

Brain imaging techniques are considered in more detail in the next section. In the rest of this chapter we summarise some existing methods for analysing shapes and assessing shape variability that will be used and extended later in the thesis. We also introduce two techniques, the Expectation-Maximisation (EM) algorithm and Markov chain Monte Carlo

(MCMC) simulation, used in this thesis for estimating the parameters in our statistical models. We then apply some of these existing methods to a simple simulated brain slice to motivate the research that follows in later chapters. A detailed outline of the rest of the thesis concludes this chapter.

1.2 Brain imaging techniques

The application of interest for shape theory in this thesis is the human brain. Important questions regarding how the brain functions remain unanswered at the start of the 21st century. As such, potential neurological applications for shape theory provide motivation for the statistical techniques developed and applied throughout this thesis. The last thirty years have seen the advent of non-invasive brain imaging techniques, which have greatly aided medical understanding by generating three-dimensional images. The most commonly used forms of brain imaging are computed tomography (CAT or CT), magnetic resonance imaging (MRI) and positron emission tomography (PET). In this thesis we treat images as observed data, but this section provides some background on their creation.

CAT scanning involves taking a series of x-ray images from different angles, and then using a computer to solve a series of algebraic equations to build a three dimensional image of where the x-rays were absorbed. MRI uses a large cylindrical magnet around the head to induce a magnetic field, through which radio waves are transmitted. The radio waves cause some of the magnetically-aligned hydrogen nuclei found in the body, to adopt a temporary non-aligned status. The frequency and phase of the reflected radio waves then enables a computer to deconvolve the signal such that the underlying structure can be identified. Each image is a two-dimensional slice, but the rapid collection of a series of slices enables a three-dimensional image to be created. CAT and MRI techniques produce structural images, which are a snapshot of the brain's geometry at the time of scanning. CAT scans show dense materials, such as bone, very clearly whereas MRI scans

will see through bones to produce good contrast between different types of soft tissue.

Functional MRI (fMRI) and PET go further than structural images by producing images showing the areas of the brain activated by a particular task. PET scanning involves injecting a radioactive chemical into the bloodstream. Sensors in the scanner then detect the positron emissions from the radioisotopes and create the image based on the distribution of these emissions. fMRI involves taking a sequence of MRI scans over a period of time to measure the blood-oxygen-level dependent (BOLD) signal. During activity, neurons take up oxygen, and the difference in magnetic properties between oxygenated and deoxygenated blood can be detected in fMRI. Therefore, fMRI scans detect neuronal activity through examining changes in the blood's oxygen level through time, rather than detecting neuronal activity directly. fMRI has largely superseded PET technology for detecting brain activation by lengthening the possible time period of exposure while conducting experiments, although using a particular radioisotope enables PET scans to detect specific brain receptors, which can be useful in the diagnosis of brain diseases.

Regardless of the imaging technique used, the result is a three-dimensional digital image where each three-dimensional pixel (a voxel) is assigned a grey-scale value typically between 0 (black) and 255 (white), which is our data. The resulting statistical analysis will, however, be affected by the initial quality of the image, which is measured in terms of spatial resolution, signal to noise ratio, non-homogeneous noise etc. The image quality is largely determined by the scanner parameters used at source, but can be improved through applying pre-processing tools to reduce noise. These tools might include averaging voxel values across a number of scans, or smoothing through space and/or time. A variety of spatial smoothing techniques exist such as thresholding, median window filtering or Gaussian kernel smoothing. We will primarily use the last of these techniques to improve the signal to noise ratio at the expense of spatial resolution.

1.3 Statistical shape analysis

The ability to quantify, describe and compare the shapes of objects has become of great importance in a wide variety of fields, including biology, medicine and genetics. Following Kendall's (1977, 1984) lead of defining shape as when "we are not interested in the location, orientation or scale", we give the broadly accepted mathematical definition of an object's shape.

Definition 1.3.1 *Shape is all the geometrical information that remains when location, scale and rotational effects are filtered out from an object.*

Size and shape, or form, is similarly defined but with the geometrical information regarding size retained by the object. The geometrical information of an object can be recorded by defining a finite set of points on the object called landmarks.

Definition 1.3.2 *A landmark is a point of correspondence on each object that matches between and within populations.*

The landmarks are chosen to characterise the shape of an object with a few points that might refer to a physical feature or a mathematical property of the object, such as a place of maximum curvature. We will assume throughout this thesis that the correspondence of landmarks between objects is known, although Dryden *et al.* (2006) and Green and Mardia (2006) consider applications of shape theory to cheminformatics and bioinformatics, respectively, where this assumption is relaxed. Information regarding the location of the landmarks is collected in a configuration matrix.

Definition 1.3.3 *The configuration is the set of landmarks on a particular object. The configuration matrix, X, is the $k \times m$ matrix of Cartesian co-ordinates of the k landmarks in m dimensions.*

It is often useful to list the individual parameters of the configuration matrix in a single vector using the vec operator.

Definition 1.3.4 *If A is an $(m \times n)$ matrix, with columns, $A_1, A_2, ..., A_n$, each vectors of length m, then the vector of length mn obtained by stacking the columns on top of one another is denoted,*

$$vec(A) = \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_n \end{bmatrix}.$$

The most common measure of a configuration's size, is the centroid size.

Definition 1.3.5 *The centroid size is given by,*

$$S(X) = \|CX\| = \sqrt{\sum_{i=1}^{k}\sum_{j=1}^{m}(X_{ij} - \bar{X}_j)^2},$$

where X_{ij} is the (i,j)th entry of X, the arithmetic mean of the jth dimension is $\bar{X}_j = \frac{1}{k}\sum_{i=1}^{k} X_{ij}$,

$$C = I_k - \frac{1}{k}1_k 1_k^T,$$

is the centring matrix, $\|X\| = \sqrt{tr(X^T X)}$ is the Euclidean norm, I_k is the $k \times k$ identity matrix, 1_k is the $k \times 1$ vector of ones, and $tr(A)$ is the trace operator that sums the diagonal elements of A.

A configuration matrix can be isotropically scaled (i.e. scaled in all directions equally) by multiplying X with a positive real number. The matrix, X, can be translated by adding a vector of length m to each landmark's co-ordinates, and can be rotated by post-multiplying with a rotation matrix.

Definition 1.3.6 *An $m \times m$ rotation matrix satisfies $\Gamma^T\Gamma = \Gamma\Gamma^T = I_m$ and $|\Gamma| = 1$. The set of all $m \times m$ rotation matrices is known as the special orthogonal group $SO(m)$.*

Definition 1.3.7 *The Euclidean similarity transformations of a configuration matrix, X, are the set of isotropically rescaled, rotated and translated X, i.e.,*

$$\{\beta X\Gamma + 1_k\gamma^T : \beta \in \mathbb{R}^+, \Gamma \in SO(m), \gamma \in \mathbb{R}^m\},$$

where $\beta \in \mathbb{R}^+$ is the scale, Γ is a rotation matrix and γ is the translation m-vector.

Definition 1.3.8 *The rigid-body transformations of a configuration matrix, X, are the set of rotated and translated X, i.e.,*

$$\{X\Gamma + 1_k\gamma^T : \Gamma \in SO(m), \gamma \in \mathbb{R}^m\},$$

where Γ is a rotation matrix and γ is the translation m-vector.

The two sets of transformations differ only in that the scaling transformation is not included in the set of rigid-body transformations.

The analysis of shape configurations based on landmark data from a statistical viewpoint (Mardia and Dryden, 1989b) has led to the formulation of shape distributions (Mardia and Dryden, 1989a) and statistical tests for examining differences between two or more populations (Dryden and Mardia, 1998, Chapter 7). An alternative approach to statistical shape analysis, through the study of inter-landmark distances, has been formulated by Lele and Richtsmeier (2001) and Rao and Suryawanshi (1996). Small (1996) and more comprehensively, Kendall *et al.* (1999), present important theoretical work regarding the distribution of shapes in shape space. However, we approach shape analysis through the framework of Procrustes analysis, as it generally presents easily interpretable findings in Euclidean space. Its primary focus is establishing shape differences, mean shapes and shape variability, and we introduce the field in more detail in the next section.

1.4 Isotropic Procrustes analysis

Procrustes methods involve estimating the transformations to superimpose one configuration on another to minimise a distance criterion. In isotropic Procrustes analysis, the criterion is the Euclidean distance. This is an arbitrary choice and the concepts given here are a basis for extending Procrustes analysis to include a weighted matching criterion in the following chapters.

Procrustes methods can be used to estimate a population's mean shape and shape variability. The adjectives "full" and "partial" are used to distinguish between the estimation of similarity and rigid-body transformations, respectively, in the Procrustes matching. In addition, "ordinary" describes matching one shape to another, whereas "generalised" describes transforming two or more configurations during the matching.

An early review of Procrustes methods was given by Sibson (1978) and more details of Procrustes methods applied to shape data are found in Dryden and Mardia (1998, Chapter 5). In this section we summarise some of standard definitions and results relating to full isotropic Procrustes analysis. Also, we assume without loss of generality, that all configurations are initially centred throughout this chapter, that is the sum of the co-ordinates in each dimension is zero.

The method of full ordinary Procrustes analysis (full OPA) involves the least squares matching of two configurations, X and μ, using the similarity transformations. Estimation of the similarity parameters, β, Γ and γ is carried out by minimising the squared Euclidean distance,

$$D^2_{OPA}(X, \mu) \;\; = \;\; \|\mu - \beta X\Gamma - 1_k\gamma^T\|^2,$$

where $\|Z\|^2 = \mathrm{tr}(Z^TZ)$ is the squared Euclidean norm, γ is a $m \times 1$ location vector, Γ is an $m \times m$ special orthogonal rotation matrix ($\Gamma \in SO(m)$) and $\beta > 0$ is a scale parameter.

Result 1.4.1 *The full ordinary Procrustes solution to the minimisation of*

$D^2_{OPA}(X, \mu)$, where $\|X\| > 0$, $\|\mu\| > 0$, is given by Dryden and Mardia (1998, p84),

$$\begin{aligned} \widehat{\gamma} &= 0, \\ \widehat{\Gamma} &= UV^T, \\ \widehat{\beta} &= \frac{tr(\mu^T X \widehat{\Gamma})}{tr(X^T X)}, \end{aligned}$$

where, $U, V \in SO(m)$ are matrices given by,

$$V \Lambda U^T = \frac{\mu^T X}{\|X\| \|\mu\|},$$

and Λ is a diagonal $m \times m$ matrix of positive elements except possibly the last element.

Kent and Mardia (2001) showed that the Procrustes match is unique when $\lambda_1, ..., \lambda_m$, the diagonal elements of Λ, are non-degenerate and optimally signed. A sequence of numbers is non-degenerate if, when written in non-increasing order, $\lambda_{m-1} + \lambda_m > 0$, and a sequence is optimally signed if all the elements are non-negative except possibly for the smallest in absolute value, $\lambda_1 \geq ... \geq \lambda_{m-1} \geq |\lambda_m| \geq 0$.

Figure 1.1 shows the full ordinary isotropic Procrustes registration of one hand-written digit "3" to a template. Note there are $k = 13$ landmarks in $m = 2$ dimensions. The landmarks were digitised from a sample of hand-written British postcodes. Note how the dotted shape has been translated, rotated and scaled to minimise the total distance between respective pairs of landmarks on itself and the template. Details of the data set can be found in Dryden and Mardia (1998, p13).

The method of full generalised Procrustes analysis (full GPA) involves rescaling, rotating and translating the configurations, X_i, $i = 1, ..., n$, relative to each other to minimise a total sum of squares, and the procedure is

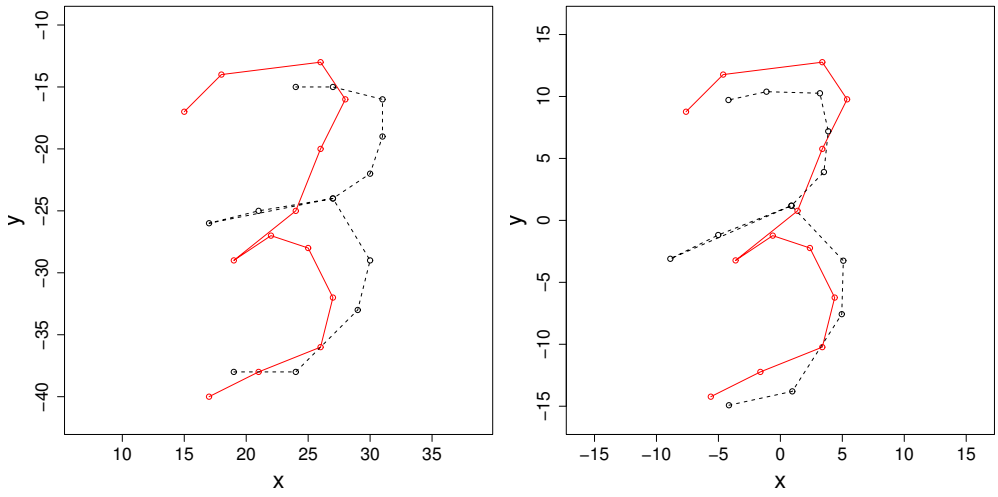

Figure 1.1: The full ordinary isotropic Procrustes registration of one hand-written "3" (dotted) to a template, showing the starting (left) and final (right) positions.

appropriate under the model,

$$X_i = \beta_i(\mu + E_i)\Gamma_i + 1_k\gamma_i^T,$$

where E_i are zero mean $k \times m$ independent random error matrices, μ is the $k \times m$ matrix of the mean configuration and β_i, Γ_i and γ_i are nuisance parameters for scale, rotation and translation. We minimise a quantity proportional to the sum of squared norms of pairwise differences,

$$G(X_1, ..., X_n) = \frac{1}{n}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \|(\beta_i X_i\Gamma_i + 1_k\gamma_i^T) - (\beta_j X_j\Gamma_j + 1_k\gamma_j^T)\|^2, \quad (1.1)$$

subject to the constraint on the size of the average centred shape,

$$S(\bar{X}) = 1,$$

where $\beta_i > 0$, $\Gamma_i \in SO(m)$, $\gamma \in \mathbb{R}^m$, $\|X\| = \sqrt{\mathrm{tr}(X^TX)}$ and $S(X)$ is the

centroid size and,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n}(\beta_i X_i \Gamma_i + 1_k \gamma_i^T),$$

is the centred average shape.

Recall that the shapes, X_i, $i = 1, ..., n$, are initially centred throughout this chapter. The transformation parameters β_i, Γ_i and γ_i are determined by the iterative GPA algorithm, Algorithm 1.4.1, first suggested by Gower (1975) and updated by Ten Berge (1977).

Algorithm 1.4.1 ***Isotropic GPA algorithm***

1. Translations. *Initially, let,*

$$X_i^P = CX_i, i = 1, ..., n,$$

where C is the centring matrix to remove location.

2. Rotations. *Update X_i^P by letting it be the ordinary Procrustes superimposition using rotation only of the current X_i^P on $\bar{X}_{(i)}$, where,*

$$\bar{X}_{(i)} = \frac{1}{n-1}\sum_{j \neq i} X_j^P.$$

Apply this to each of the n configurations in turn until Equation (1.1) cannot be reduced further.

3. Scaling. *Let ϕ be the eigenvector, corresponding to the largest eigenvalue, of the correlation matrix of the vec(X_i^P). Then for $i = 1, ..., n$,*

$$\widehat{\beta}_i = \left(\frac{\sum_{k=1}^{n} \|X_k^P\|^2}{\|X_i^P\|^2}\right)\phi_i,$$

where ϕ_i is the ith value in the vector ϕ.

4. Repetition. *Repeat steps 2 and 3 until Equation (1.1) cannot be reduced further.*

Figure 1.2 shows the full generalised isotropic Procrustes registration of 29 male gorilla skulls. Note there are $k = 8$ landmarks in $m = 2$ dimensions, taken from the midplane of each skull, with the face to the left. Starting at the far left and going clockwise, the landmarks are labelled "prosthion", "nariale", "nasion", "bregma", "lambda", "opisthion", "basion" and "staphylion". The initial registration fixes one landmark, "opisthion", at the origin and the line joining it to another, "basion", to be horizontal. Following registration, better inference can be made regarding the mean shape and the variability in the sample. Note that the shapes have been resized relative to the original configurations. Details of the data set can be found in Dryden and Mardia (1998, p11).

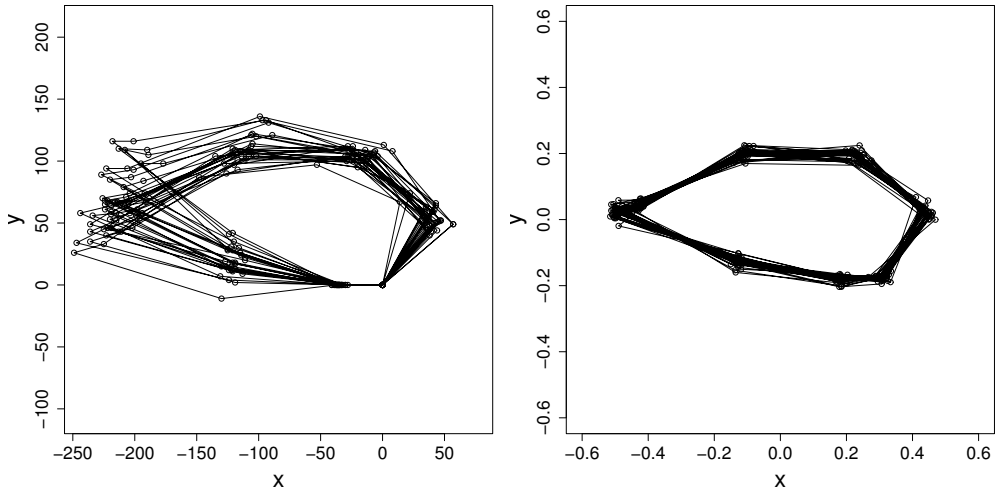

Figure 1.2: The full generalised isotropic Procrustes registration of 29 male gorilla skulls, showing the starting (left) and final (right) positions.

Definition 1.4.1 *Let $\hat{\beta}_i$, $\hat{\Gamma}_i$ and $\hat{\gamma}_i$, be the parameters which minimise Equation (1.1), then the full Procrustes fit of each X_i is given by,*

$$X_i^P = \widehat{\beta}_i X_i \widehat{\Gamma}_i + 1_k \widehat{\gamma}_i^T, \quad i = 1, ..., n.$$

Result 1.4.2 *The full Procrustes estimate of the mean shape is given by,*

$$\widehat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i^P.$$

Analogous definitions and results for isotropic partial Procrustes analysis can also be obtained by omitting the β parameter in the definitions of OPA, GPA, and Procrustes fit, and omitting the constraint on the mean shape in the definition of full GPA.

Alternative registrations or co-ordinate systems exist. Bookstein (1986) proposes removing the similarity transformations in $m = 2$ dimensions by translating, rotating and rescaling such that landmarks 1 and 2 are sent to fixed positions $(0, 0)$ and $(1, 0)$. Dryden and Mardia (1998, p27) redefine Bookstein coordinates by registering landmarks 1 and 2 to $(-\frac{1}{2}, 0)$ and $(\frac{1}{2}, 0)$.

Definition 1.4.2 *Let (x_j, y_j), $j = 1, ..., k$, $k \geq 3$, be landmark co-ordinates in a plane, then after translating, rotating and rescaling landmarks 1 and 2 to $(-\frac{1}{2}, 0)$ and $(\frac{1}{2}, 0)$, the remaining co-ordinates of an object, (u_j^B, v_j^B), $j = 3, ..., k$ are,*

$$\begin{aligned} u_j^B &= \left\{(x_2 - x_1)(x_j - x_1) + (y_2 - y_1)(y_j - y_1)\right\} / D_{12}^2 - \frac{1}{2}, \\ v_j^B &= \left\{(x_2 - x_1)(y_j - y_1) - (y_2 - y_1)(x_j - x_1)\right\} / D_{12}^2, \end{aligned}$$

where $D_{12}^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2 > 0$ and $-\infty < u_j^B, v_j^B < \infty$.

In later chapters we will extend Procrustes methods by seeking to minimise the more general Mahalanobis norm instead of the Euclidean norm.

Definition 1.4.3 *Let A be an $(m \times n)$ matrix and let W be a symmetric $(mn \times mn)$, positive definite, invertible matrix, then the squared Mahalanobis norm, denoted $\|A\|_W^2$, is defined as,*

$$\|A\|_W^2 = vec(A)^T W^{-1} vec(A).$$

The squared Euclidean norm, $\|A\|^2 = tr(A^TA)$, is a special case of the squared Mahalanobis norm, with $W = I_{mn}$.

Letting A be the residual between a shape, X, and a template, μ, then the Mahalanobis distance with respect to W is $\text{vec}(X-\mu)^T W^{-1}\text{vec}(X-\mu)$. A natural choice for W would be the shape covariance matrix, Σ.

Procrustes methods have previously been extended by Goodall (1991) to include weighted least squares (WLS) and iteratively reweighted least squares (IRLS). Within WLS the covariance structure is assumed known, but in IRLS is unknown and estimated. Generally Goodall uses a factored covariance structure, $\Sigma = \Sigma_m \otimes \Sigma_k$, where $\otimes$ is the Kronecker product.

Definition 1.4.4 *Let A be a matrix of dimension $(m \times n)$, with individual entries, a_{ij}, $i = 1, ..., m$, $j = 1, ..., n$; and let B be a $(p \times q)$ matrix. The Kronecker product of A and B is defined as,*

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \dots & a_{1n}B \\ a_{21}B & a_{22}B & \dots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \dots & a_{mn}B \end{bmatrix},$$

which is an $(mp \times nq)$ matrix.

This structure means one can have non-identical variability at each landmark and non-identical variability between the dimensions, but this formulation is not completely general. Goodall's method for when $\Sigma_m = I_m$ replaces configurations X by QX and μ by $Q\mu$, where Q^TQ is the Cholesky decomposition of Σ_k^{-1}. This accounts for the centring, and the rotation is estimated as with isotropic covariance. When $\Sigma_m \neq I_m$, Goodall (1991) claims there is no explicit expression for $X\widehat{\Gamma}$ and so there is no unique minimum of the Mahalanobis norm. Goodall recommends the iterative algorithm of Koschat and Swayne (1991), where they consider a diago-

nal weighting matrix, D, in this case. Goodall's approach to estimating the nuisance parameters for a general covariance structure in two dimensions is to solve a WLS multiple regression problem with $2k$ observations, $\text{vec}(\mu)$ as the response variable and four regression parameters being $\beta\cos\psi, \beta\sin\psi, \gamma_1$ and γ_2. Goodall acknowledges that this method is not useful for size-and-shape analysis. In Chapter 2, we present solutions of the optimal registration parameters for both partial and full, ordinary and generalised, Procrustes analysis. It should also be noted that Gower and Dijksterhuis (2004) introduce a weighted Procrustes measure, $\|WX - \mu\|^2 = \text{tr}(X^TW^TWX - 2X^TW^T\mu + \mu^T\mu)$, where W is a square weighting matrix.

1.5 Principal components analysis

Principal Components Analysis (PCA) is a widely used technique in multivariate analysis (e.g. Mardia *et al.*, 1979) and is often used in shape theory to examine shape variability. The aim of PCA is to summarise the data with a few variables, which are linear combinations of the original variables, while minimising the amount of variability lost. We describe sample PCA here. Let X be a $(p \times n)$ data matrix, with columns containing the vectors, $X_1, ..., X_n$, of length p with sample mean, $\bar{X}$, and sample covariance matrix,

$$S = UDU^T,$$

using the spectral decomposition theorem, where U is an orthogonal matrix of the eigenvectors of S and D is a diagonal matrix of the eigenvalues of S, with eigenvalues $d_1 \geq ... \geq d_p \geq 0$. The jth principal component transformation is defined by,

$$y_j = u_j^T(X - \bar{X}1_n^T), \qquad j = 1, ..., p < n,$$

where u_j is the standardised eigenvector of S corresponding to the jth largest eigenvalue, d_j. Combining the principal components together,

$$Y = U^T(X - \bar{X}1_n^T).$$

The rows of Y are uncorrelated, with the jth row having variability, $\text{Var}(y_j) = d_j$, so the principal components are ordered with the first PC capturing the largest amount of variability. Retaining the first few principal components provides a lower dimensional representation of the data. The ith element of the row y_j, is known as the jth principal component score of the ith individual.

In shape theory, PCA can be applied to the Procrustes residuals, such that the sample covariance matrix is,

$$S = \frac{1}{n}\sum_{i=1}^{n} \text{vec}(X_i^P - \widehat{\mu})\text{vec}(X_i^P - \widehat{\mu})^T,$$

where X_i^P is the Procrustes fit of the ith shape and $\widehat{\mu}$ is the Procrustes estimate of the mean shape. Let the orthonormal eigenvectors of S be u_j, $j = 1, ..., p$, and let the jth principal component score of the ith individual be s_{ij}. Therefore, the ith individual is,

$$\text{vec}(X_i^P) = \text{vec}(\widehat{\mu}) + \sum_{j=1}^{p} s_{ij}u_j,$$

and the percentage of variability captured by the jth PC is,

$$\frac{100d_j}{\sum_{i=1}^{p} d_j}.$$

The effect of the jth PC can often be seen in shape theory by plotting,

$$\text{vec}(X) = \text{vec}(\widehat{\mu}) + cd_j^{1/2}u_j,$$

for a range of values of c. Typically, c varies in the range $[-3, 3]$ to cover the majority of the variability under a multivariate normal model.

1.6 EM algorithm

The Expectation Maximisation (EM) algorithm (Dempster *et al.*, 1977) is a statistical tool for maximum likelihood estimation of model parameters when there are missing data. It can also be used if the form of the likelihood is too complicated for conventional maximum likelihood estimation. It can be applied in either a Bayesian setting to find the maximum *a posteriori* estimator, or in a frequentist approach to find the maximum likelihood estimate. The method works by augmenting the observed data, Y, with additional data, X. One frequentist application of the EM algorithm lets X represent missing data. Alternatively, in this thesis we will let X represent latent or hidden variables such that maximising the likelihood of the complete data set is easier than maximising the likelihood of Y.

Let θ be the model parameters, and let $P(Y|\theta)$ be the probability of the data given the model parameters. The value of θ which maximises P also maximises the log-likelihood, $l(\theta) = \log P(Y|\theta)$ since log is a strictly increasing function. Let θ' be an estimate of the parameters, then for an updated estimate to increase the likelihood we need to maximise, $l(\theta) - l(\theta')$. Let x be a realisation of the latent variables, then,

$$\begin{aligned} P(Y|\theta) &= \sum_x P(Y|x,\theta)P(x|\theta), \\ l(\theta) - l(\theta') &= \log \sum_x P(Y|x,\theta)P(x|\theta) - \log P(Y|\theta'). \end{aligned} \tag{1.2}$$

Jensen's inequality states that if f is a concave function, $\lambda_i \geq 0$ and

$\sum_{i=1}^{n} \lambda_i = 1$ then,

$$f\left(\sum_{i=1}^{n} \lambda_i x_i\right) \geq \sum_{i=1}^{n} \lambda_i f(x_i).$$

Note that the probabilities, $P(x|Y,\theta')$, satisfy the constraints on the λ_i. Given log is a concave function, and $\log(\sum_i a_i) - \log b = \log(\sum_i (a_i/b))$, Equation (1.2) can be rewritten as,

$$\begin{aligned} l(\theta) - l(\theta') &= \log \sum_x \left(P(x|Y,\theta') \frac{P(Y|x,\theta)P(x|\theta)}{P(x|Y,\theta')} \right) - \log P(Y|\theta'), \\ &= \log \sum_x P(x|Y,\theta') \left(\frac{P(Y|x,\theta)P(x|\theta)}{P(x|Y,\theta')P(Y|\theta')} \right), \\ &\geq \sum_x P(x|Y,\theta') \log \left(\frac{P(Y|x,\theta)P(x|\theta)}{P(x|Y,\theta')P(Y|\theta')} \right) = \Lambda(\theta|\theta'). \end{aligned}$$

Hence, $l(\theta) \geq q(\theta|\theta')$, where $q(\theta|\theta') = l(\theta') + \Lambda(\theta|\theta')$. Also, $q(\theta'|\theta') = l(\theta')$ since,

$$\log \left(\frac{P(Y|x,\theta')P(x|\theta')}{P(x|Y,\theta')P(Y|\theta')} \right) = \log \left(\frac{P(Y,x|\theta')}{P(Y,x|\theta')} \right) = 0.$$

So the function $q(\theta|\theta')$ is bounded above by the likelihood function that we wish to maximise, $l(\theta)$, and the two functions are equal for the current estimate, θ'. Therefore, choosing a θ which increases $q(\theta|\theta')$ also increases $l(\theta)$, and to get the greatest possible increase we choose the updated value to maximise $q(\theta|\theta')$. Ignoring terms in $q(\theta|\theta')$ that are constant with respect to θ, the updated value is then,

$$\begin{aligned} \theta^* &= \arg\max \left\{ \sum_x P(x|Y,\theta') \log \left(P(Y|x,\theta)P(x|\theta) \right) \right\}, \\ &= \arg\max \left\{ \sum_x P(x|Y,\theta') \log \left(\frac{P(Y,x,\theta)}{P(x,\theta)} \frac{P(x,\theta)}{P(\theta)} \right) \right\}, \end{aligned}$$

$$
\begin{aligned}
&= \arg\max \left\{ \sum_x P(x|Y,\theta') \log P(Y,x|\theta) \right\}, \\
&= \arg\max \left\{ \mathbb{E}_{X|Y,\theta'} \left[\log P(Y,x|\theta) \right] \right\}.
\end{aligned}
$$

Conventionally, we let $Q(\theta|\theta') = \mathbb{E}_{X|Y,\theta'} [\log P(Y,x|\theta)]$. The EM algorithm is, therefore, an iterative procedure with each iteration consisting of an E-step which determines the conditional expectation of Q and an M-step which maximises the expression with respect to θ. The above derivation shows that at each iteration the likelihood, $l(\theta)$, is non-decreasing. Given the likelihood is bounded above, the algorithm is guaranteed to converge, and at that value $l(\theta) = q(\theta)$. However, the algorithm is not guaranteed to converge to a global maximum and is dependent on the starting estimate of θ. A more comprehensive introduction to the EM algorithm can be found in Tanner (1996).

1.7 Markov chain Monte Carlo simulation

Markov Chain Monte Carlo (MCMC) simulation is a statistical modelling tool that can be used in Bayesian statistics for integrating over high-dimensional probability distributions in order to make inferences about model parameters. MCMC can also be used in other situations as a simulation method. In this thesis, we will use it for Bayesian inference when the posterior distribution cannot be obtained analytically. The method works by approximately drawing dependent samples from the posterior distribution of a parameter, from which inference can be made about the moments of the distribution. MCMC was first formulated by Metropolis *et al.* (1953) with significant additions and improvements from Hastings (1970) and Geman and Geman (1984).

Given observed data, D, and model parameters or missing data, θ, the

joint probability distribution is,

$$P(D, \theta) = P(\theta)P(D|\theta),$$

where $P(\theta)$ is a prior distribution and $P(D|\theta)$ is a likelihood. Therefore, the distribution of θ conditional on the observed data, D, is,

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{\int P(\theta)P(D|\theta)d\theta} = \pi(\theta|D),$$

where $\pi(\theta|D)$ is the posterior distribution of θ that we are interested in. In the rest of this thesis, prior distributions will be denoted with P, likelihoods with L and posterior distributions with π.

Let θ be a vector of k random variables, then Monte Carlo integration evaluates $\mathbb{E}(\theta)$ by drawing samples $\{\theta_t, t = 1, ..., n\}$ from $\pi(\theta|D)$ and approximating,

$$\mathbb{E}(\theta) \approx \frac{1}{n}\sum_{t=1}^{n}\theta_t.$$

Sampling from $\pi(\theta|D)$ could be done by any process that samples the distribution in the correct proportions. MCMC works by constructing $\{\theta_t\}$ from a Markov chain that has $\pi(\theta|D)$ as its stationary distribution. A Markov chain is a sequence of random variables, such that at time t, the value of θ_{t+1} is only dependent on θ_t and independent of the previous samples. The value of θ_{t+1} is sampled from the distribution, $P(\theta_{t+1}|\theta_t)$, which is independent of t. The Markov chain will eventually become effectively independent of its starting state, θ_0, and t, so the distribution of θ_t is invariant or stationary. The first m values of θ_t, which might be dependent on θ_0, are discarded and the rest, which approximately come from the stationary

distribution, are used to give the estimator,

$$\mathbb{E}(\theta) \approx \frac{1}{n-m} \sum_{t=m+1}^{n} \theta_t.$$

Constructing a Markov chain with $\pi(\theta|D)$ as the stationary distribution can be done via the Metropolis-Hastings algorithm, Algorithm 1.7.1. See Chib and Greenberg (1995) for an overview of this technique. At each time, t, a candidate point, θ^*, is sampled from a proposal distribution $q(\theta^*|\theta_t)$ and is accepted as θ_{t+1} with probability,

$$\alpha(\theta_t, \theta^*) = \min\left(1, \frac{\pi(\theta^*)q(\theta_t|\theta^*)}{\pi(\theta_t)q(\theta^*|\theta_t)}\right), \tag{1.3}$$

otherwise $\theta_{t+1} = \theta_t$.

Algorithm 1.7.1 ***The Metropolis-Hastings algorithm***

1. *Initialise θ_0 and let $t = 0$.*

2. *Sample θ^* from $q(\theta^*|\theta_t)$.*

3. *Sample U from the uniform $U(0,1)$ distribution.*

4. *If $U \leq \alpha(\theta_t, \theta^*)$ let $\theta_{t+1} = \theta^*$, otherwise let $\theta_{t+1} = \theta_t$.*

5. *Increment t.*

6. *Repeat steps 2 to 5 n times.*

Gilks *et al.* (1996) provides a complete proof, beyond the scope of this introduction, that the Markov Chain resulting from the Metropolis-Hastings algorithm converges to, and continues to sample from, the specified stationary distribution. In brief, the argument requires us to show that the chain is irreducible, aperiodic and reversible. A chain is irreducible if, given enough iterations, it can reach all interesting parts of its state-space irrespective of its starting point. The aperiodicity requirement prevents the chain from oscillating between a fixed number of states in a regular periodic manner. If a chain satisfies just these two conditions then it has a unique stationary

distribution. The third condition, reversibility, is defined with respect to the distribution, π, and requires the balance equation,

$$\pi(\theta_t)P(\theta_{t+1}|\theta_t) = \pi(\theta_{t+1})P(\theta_t|\theta_{t+1})$$

to be satisfied for all t. If the chain is reversible, as well as irreducible and aperiodic, then the chain's unique stationary distribution is π. The Metropolis-Hastings algorithm can be shown to satisfy these conditions due to the acceptance/rejection step.

The proposal distribution, q, can have almost any form, but must be chosen carefully for the chain to move around the support of π efficiently. If the distance between the proposed value and the current value is typically too large then the majority of proposals will be rejected. If the distance is too small then it will require more samples to move about the entire support of π. Either case will result in slow mixing.

If the proposal distribution is symmetric, $q(\theta^*|\theta_t) = q(\theta_t|\theta^*)$ for all t, then Equation (1.3) simplifies. A special case is the random-walk, where $q(\theta^*|\theta_t) = q(|\theta^* - \theta_t|)$. A common form of $q(\theta^*|\theta_t)$, with q symmetric, is a multivariate normal distribution with mean, θ_t, and constant covariance matrix, Σ.

Sometimes it is more convenient to update each element of θ individually. Let $\theta = (\theta_1, ..., \theta_k)$ be the vector of current parameter values and $\theta_{-i} = (\theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_k)$. In each iteration of the Metropolis-Hastings algorithm the elements of θ are updated in turn. For the ith parameter, we propose a new value, θ_i^*, which we sample from the proposal distribution, $q_i(\theta_i^*|\theta_i, \theta_{-i})$. Then, θ_i is updated to θ_i^* with probability,

$$\alpha(\theta_i, \theta_i^*; \theta_{-i}) = \min\left(1, \frac{\pi(\theta_i^*|\theta_{-i})q_i(\theta_i|\theta_i^*, \theta_{-i})}{\pi(\theta_i|\theta_{-i})q_i(\theta_i^*|\theta_i, \theta_{-i})}\right).$$

All of the MCMC algorithms in this thesis will update parameters in turn within one iteration, although sometimes individual parameters are collated

in matrices and the matrices are updated in turn.

If the full conditional distribution of a parameter given the rest, $\pi(\theta_i|\theta_{-i})$, is known then choosing this distribution as the ith proposal distribution,

$$\pi(\theta_i^*|\theta_{-i}) = q_i(\theta_i^*|\theta_i, \theta_{-i}), \qquad \pi(\theta_i|\theta_{-i}) = q_i(\theta_i|\theta_i^*, \theta_{-i}),$$

results in the proposal being accepted with probability 1. This special case of the Metropolis-Hastings sampling method is known as the Gibbs sampler. If sampling from the full conditional distribution is possible, then using the Gibbs sampler is often computationally quicker and usually mixes more efficiently than alternative Metropolis-Hastings steps.

In practical applications consideration must be given to the choice of prior distribution, based on *a priori* knowledge. If little is known, then using a non-informative prior, such as a uniform distribution over the parameter space, will place more emphasis on the data in the likelihood. Conversely, if a parameter is known to be in a narrow interval, then an informative prior, such as a Gaussian distribution with small variance parameter, will be more appropriate. Care should also be given so that the support of the prior distribution does not extend outside the parameter space. For example, a variance parameter must be positive.

1.8 Motivating example: Asymmetric shape registration

In this section we consider a small motivating simulation study to explore two concepts, shape registration and symmetry, that will occur throughout this thesis. Shape registration, as we have seen through Procrustes methods, is concerned with superimposing one shape on a template to minimise some matching criterion. A shape is symmetric if there is an exact correspon-

dence of its form on opposite sides of a dividing line or plane. In applying shape analysis to the human brain, the object is seen to be approximately symmetric with left and right hemispheres. However, as will be shown later in this thesis, on closer examination the human brain is asymmetric with the hemispheres having different shape and function.

To illustrate asymmetric shape registration, consider a 2-dimensional shape template, Φ, made up of 9 landmarks on a midline and 40 landmarks evenly spaced on the unit circle, where the ith co-ordinate is given by, $\Phi_i = (\sin(\theta_i), \cos(\theta_i))$, where $\theta_i = 2\pi i/40$ for $i = 1, ..., 40$. We perturb this shape to make an asymmetric template, μ, with x co-ordinates,

$$\mu_x = \begin{cases} \sin(\theta_i) + \lambda \sin(2\theta_i) & i = 1, ..., 5, \\ \sin(\theta_i) + \lambda \sin((2\theta_i + \pi)/3) & i = 6, ..., 20, \\ \sin(\theta_i) + \lambda \sin(2\theta_i - \pi) & i = 21, ..., 25, \\ \sin(\theta_i) + \lambda \sin(2(\theta_i + \pi)/3) & i = 26, ..., 40, \end{cases}$$

where λ is a chosen real positive distortion parameter. Clearly, the distortion is greatest when $\theta_i = \pi/4$ or $\theta_i = 5\pi/4$. A data set of n shapes is then simulated from the model, $X_j = \mu + \sigma\epsilon_j$, $j = 1, ..., n$, where $\text{vec}(\epsilon) \sim N(0, I)$ and $\sigma = 0.01$. The templates and an example shape can be seen in Figure 1.3.

This scenario reflects the common situation in brain image registration where an asymmetric brain is registered to a symmetric template, often created in studies by averaging left/right flips of a scan. Assuming the errors are isotropic, a registration to the symmetric template would try to minimise,

$$S(X^P, \Phi) = \frac{1}{n} \sum_{j=1}^{n} \|X_j^P - \Phi\|^2,$$

where X_j^P is the jth shape following registration. Ideally, though, a regis-

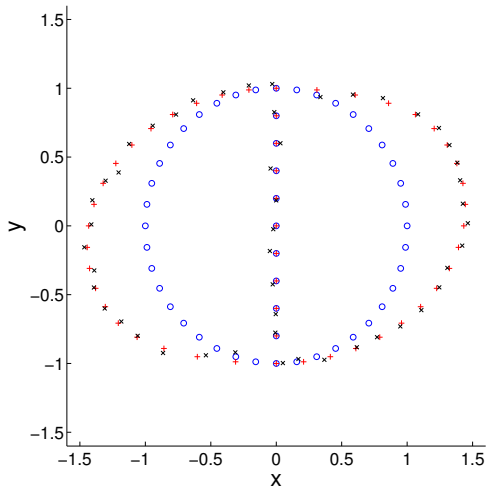

Figure 1.3: A simulated 2D slice of the brain, with a symmetric template, Φ (o), an asymmetric template, μ (+), with $\lambda = 0.5$, and a simulated configuration, X ($\times$), with $\sigma = 0.01$.

tration would minimise the distance to the true model mean,

$$S(X^P, \mu) = \frac{1}{n} \sum_{j=1}^{n} \|X_j^P - \mu\|^2.$$

For a simulated data set of $n = 100$ shapes, consider five different rigid-body registration methods. Two methods use OPA to register the shapes to μ and Φ, where Φ is the symmetric template displayed in Figure 1.3. Three use landmarks on the midline alone: (a) OPA on the midline; (b) a line of best fit registration that minimises the Euclidean norm of the midline x co-ordinates and centres on the y co-ordinates; and (c) endpoint registration, ensuring the top and bottom landmarks in the midline were on the y axis, equidistant from the origin. The distance $S(X^P, \mu)$ was measured following each registration for various values of λ and the results are given in Table 1.1 and can be viewed in the left plot of Figure 1.4.

We measure the asymmetry of the shape by dividing the registered shape

λ	OPA(μ)	OPA(Φ)	OPA(Mid)	Best fit	Endpoint
0.00	0.0094	0.0094	0.0103	0.0111	0.0149
0.02	0.0096	0.0098	0.0105	0.0114	0.0166
0.04	0.0094	0.0102	0.0102	0.0111	0.0161
0.06	0.0096	0.0113	0.0105	0.0112	0.0173
0.08	0.0096	0.0127	0.0105	0.0115	0.0161
0.10	0.0095	0.0144	0.0105	0.0113	0.0158

Table 1.1: $S(X^P, \mu)$ following different registration methods.

into $m = 18$ equally spaced horizontal slices and estimating the area contained between the y axis and the boundary of the shape on the left hand side of the midline, V_{rj}^L, and on the right hand side of the midline, V_{rj}^R, for $r = 1, ..., m$ and $j = 1, ..., n$. The asymmetry function for the jth shape has components,

$$\eta_{rj} = \frac{V_{rj}^R - V_{rj}^L}{V_{rj}^R + V_{rj}^L}.$$

The overall goodness of fit measure is, $S(\eta(X^P), \eta(\mu)) = 1/n \sum_{j=1}^{n} \|\eta(X_j^P) - \eta(\mu)\|^2$, where $\eta(X_j^P)$ is the asymmetry function applied to the jth registered shape. The distance $S(\eta(X^P), \eta(\mu))$ was measured following each registration for various values of λ and the results are given in Table 1.2 and can be viewed in the right plot of Figure 1.4.

λ	OPA(μ)	OPA(Φ)	OPA(Mid)	Best fit	Endpoint
0.00	0.0363	0.0363	0.0365	0.0406	0.0483
0.20	0.0328	0.0363	0.0327	0.0350	0.0430
0.40	0.0328	0.0737	0.0349	0.0376	0.0430
0.60	0.0284	0.1339	0.0309	0.0389	0.0482
0.80	0.0276	0.2121	0.0323	0.0348	0.0412
1.00	0.0269	0.3172	0.0344	0.0398	0.0436

Table 1.2: $S(\eta(X^P), \eta(\mu))$ following different registration methods.

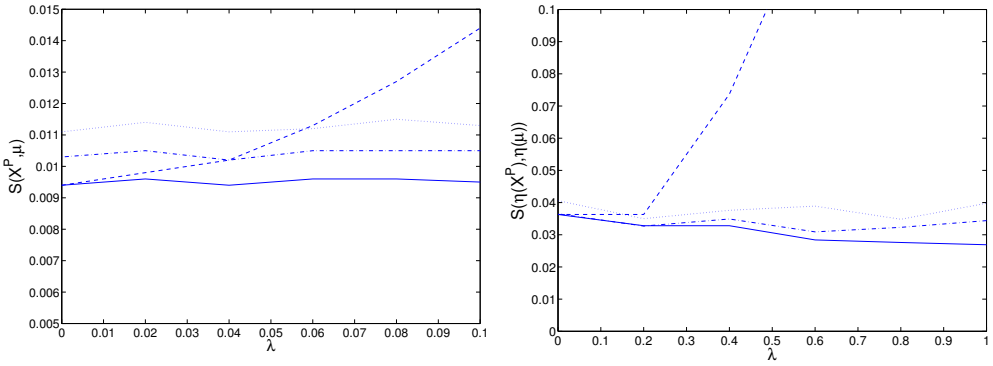

Figure 1.4: The value of $S(X^P, \mu)$ (left) and $S(\eta(X^P), \eta(\mu))$ (right), following registration using OPA onto μ (solid line), OPA onto Φ (dashed line), OPA using midline landmarks (dash/dot), and best fit line (dotted), for different values of λ.

Unsurprisingly, OPA onto μ minimises both the shape difference and the symmetry difference. However, the use of a symmetric template leads to grossly inaccurate registrations, even for very small values of the distortion parameter, λ. The use of midline registration methods consistently produce smaller shape and symmetry differences than using OPA onto Φ when asymmetry is present. OPA onto the midline landmarks produced the least error of the midline methods. This simple simulation study highlights the need for a generalisation of Procrustes methods to cope with data sets where the errors are not isotropic. In addition to template choice, non-isotropic errors might also be caused by systematic errors in landmarking (Glasbey *et al.*, 1995), natural biological variation, or medical deformities, such as brain lesions (Brett *et al.*, 2001). This section also shows that the choice of registration influences the measure of symmetry.

1.9 Outline of the thesis

In Section 1.8 we showed there were advantages to performing Procrustes registration using only the midline landmarks when the template for the other landmarks was incorrect. In Chapter 2 we consider the application of Procrustes methods to a subset of the available landmarks in more detail. We formulate results regarding the distribution of the landmarks in the subset and in the complement when matching to both the correct and incorrect templates. Further, we extend Procrustes methods to enable registration with respect to any symmetric positive definite weighting matrix, Σ.

The choice of Σ will depend on the application and aim of the analysis but in a general setting Σ would represent the variance-covariance matrix of the data. Typically, estimation of shape variability is carried out after isotropic Procrustes registration. However, Chapter 3 seeks to jointly estimate the registration parameters, the mean shape and the shape variability, with the aim of using the estimated shape variability to influence the registration. Both maximum likelihood and Bayesian methods are developed. Their ability to succeed relies on *a priori* information in the form of constraints in the maximum likelihood case, and prior distributions for the Bayesian MCMC algorithm. The methods are compared to isotropic Procrustes methods through a simulation study.

Section 1.8 also introduced the concept of asymmetry commonly found in the human brain. In Chapter 4, we apply a measure of symmetry to a data set consisting of MRI scans. The volunteers in the study were either healthy controls or patients with schizophrenia, and we use asymmetry to establish differences in the shape of the cortical surface between the two patient groups. Comparisons are only possible if the images have been correctly registered to a template, so we develop brain registration algorithms, again using both maximum likelihood and Bayesian frameworks. Both algorithms make use of symmetry around the midline given that asymmetric features on the cortical surface have already been seen to lead to inaccurate

registrations. The technique for analysis we develop is compared to the existing technique of voxel-based morphometry.

In Chapter 5 we consider the more complicated problem of analysing a 4-dimensional data set created by a series of brain images acquired over a few minutes. This presents the additional challenge of registering the data temporally, as well as spatially. Following registration, however, we can analyse brain function by examining the resultant time series of oxygen levels at each voxel. We develop a statistical model for analysing the changes in response to a repeated stimuli, with errors correlated in time as well as space, and use the EM algorithm to maximise the likelihood of the model. The aims are to establish which voxels are activated by the stimulus and to quantify the changes in response through time. Improvements to current data pre-processing methods are also suggested, and the analysis is compared to the results of existing techniques that use statistical linear models and independent components analysis. The creation and analysis of functional imaging data is still developing. The 7T MRI scanner at the University of Nottingham, which came online in September 2005, is one of the two most powerful scanners in Europe. The increased power raises the signal to noise ratio, allowing single trial variability to be examined for the first time.

Lastly, in Chapter 6 we draw some conclusions and discuss potential areas for future research.

Calculations, graphics and results in this thesis have been conducted and produced using software packages R (R Development Core Team, 2005), including the shapes library, and Matlab (MathWorks, Natick, MA, USA). In Chapters 4 and 5, existing software packages, SPM2 (e.g. Friston *et al.*, 1995a) and FSL (Smith *et al.*, 2004), have also been used for data pre-processing and making comparisons between methods.

Chapter 2

Weighted Procrustes analysis

2.1 Introduction

The standard approach to Procrustes analysis, outlined in Chapter 1, assumes that all landmarks are included in the analysis and an equal weighting is given to each. We previously demonstrated in Section 1.8 the need for techniques that do not make these assumptions. In this chapter we develop a weighted Procrustes registration method for shapes with k landmarks in m dimensions. We start by examining a special case, where we divide the landmarks into two sets and use only one subset to estimate the registration parameters. We formulate the distribution of the sum of squared Euclidean distances between a registered shape and a template for both subsets.

This application of isotropic Procrustes is then generalised to allow for a weighted Procrustes technique based on a $km \times km$ matrix, Σ, where the only restriction is that Σ is symmetric positive definite. This method is formulated for matching one shape to a template (ordinary Procrustes) and for registering multiple shapes to each other (generalised Procrustes). An earlier summary of these methods is presented in Brignell *et al.* (2005). A weighted registration is applicable in brain registration where the cortical surface is known to be more variable than other areas. Richmond *et al.*

(2004) consider the registration of molecules where the landmarks (atoms) are defined by charge as well as geometrical location. One method, amongst others, of applying a weighted Procrustes registration could incorporate the compatibility of atom charges in the weighting matrix. The results given in this chapter lay the groundwork for Chapter 3, where we estimate the shape covariance matrix and recursively use this estimate in place of Σ to find optimum registration parameters.

2.2 Subset matching

In conducting standard Procrustes analysis, one might choose to register a shape to a template using only a subset of the available landmarks. Analysis of the subset would give the Procrustes estimates of the registration parameters, which would then be applied to the complement. In this section, we seek to establish the distribution of the sum of squared Euclidean distances, following this registration, between the shape and the template for both the subset and the complement.

Consider a configuration $X = [X_1^T, X_2^T]^T$ which we wish to match to the template, $\mu = [\mu_1^T, \mu_2^T]^T$, by Procrustes analysis (OPA) on the subsets X_1 and μ_1, assuming without loss of generality that μ_1 is centred. Each configuration consists of k landmarks in m dimensions and X_j is $k_j \times m$, where $j = 1, 2$, $k_j > m$ and $k = k_1 + k_2$. We will be matching using rotation and translation, but not scaling. The model considered is $X = \mu + \sigma\epsilon$ where $\epsilon^T 1_k = 0$ and $\epsilon = [\epsilon_1^T, \epsilon_2^T]^T$ is an error matrix and σ is a scalar quantity.

2.2.1 Distribution of the subset landmarks

Proposition 3 of Kent and Mardia (2001) states,

Result 2.2.1 *Let A be a symmetric matrix with spectral decomposition $A = U\Lambda U^T$, where U is orthogonal and $\Lambda = diag(\lambda_i)$ is diagonal, such that the eigenvalues $\{\lambda_i\}$ are optimally signed and nondegenerate. Let $B =$*

$A + \sigma UCU^T = U(\Lambda + \sigma C)U^T$, where C is a general $m \times m$ matrix. Then, for small enough $\sigma > 0$, the rotation matrix Γ maximising $tr(B\Gamma)$ is given by $\Gamma = I + \sigma UDU^T + O(\sigma^2)$, where D has elements,

$$d_{ij} = \begin{cases} (c_{ji} - c_{ij})/(\lambda_i + \lambda_j) & i \neq j, \\ 0 & i = j. \end{cases}$$

The matrix $D = (-C + C^T) * L$, where $(L)_{ij} = 1/(\lambda_i + \lambda_j)$, $1 \leq i, j \leq m$ and $*$ denotes the Hadamard elementwise product. This product is defined such that the ijth entry of $R * S$, where R and S are matrices of the same size, is the product of the individual ijth entries of R and S. To clarify, C is a random $m \times m$ matrix, but U is constant, $U \in O(m)$. The matrix, D, is a function of C and the eigenvalues of $\mu_1^T \mu_1$.

Result 2.2.2 *Following minimisation by rigid-body transformations, the approximate distribution of the sum of Euclidean squared distances between the subset landmarks is given by,*

$$D_S = \left\| \frac{X_1 \widehat{\Gamma}_1 - \mu_1}{\sigma} \right\|^2 \sim \chi^2_{(k_1-1)m - \frac{m(m-1)}{2}}, \tag{2.1}$$

where, without loss of generality, X_1 and μ_1 have been centred and $\widehat{\Gamma}_1$ is the estimator of the minimising rotation.

Proof: Minimising the Euclidean squared distance between X_1 and μ_1, $\|X_1\widehat{\Gamma}_1 - \mu_1\|^2$, is equivalent to maximising the term $\mathrm{tr}(\mu_1^T X_1 \widehat{\Gamma})$. Application of Kent and Mardia's result, using their notation, implies $B = \mu_1^T X_1 = \mu_1^T(\mu_1 + \sigma\epsilon_1)$, $A = \mu_1^T \mu_1 = U\Lambda U^T$ and $\mu_1^T \epsilon_1 = UCU^T$, where $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_m)$.
If $\mathrm{vec}(\epsilon_1) \sim N(0, \sigma^2 I)$, then the distribution of D_S is unaffected by pre-multiplication by $V^T \in O(k)$ and post-multiplication by $W \in O(m)$. Let, $\widetilde{\mu}_1 = V^T \mu_1 W$, then $\widetilde{\mu}_1^T \widetilde{\mu}_1 = W^T \mu_1^T V V^T \mu_1 W = W^T \mu_1^T \mu_1 W = W^T U \Lambda U^T W$. We can, without loss of generality, choose $W = U$, therefore $\widetilde{\mu}_1^T \widetilde{\mu}_1 = \Lambda$.

Further, we can choose V such that $\widetilde{\mu}_1$ only has non-zero entries on the diagonal equivalent to $\text{diag}(\lambda_1^{\frac{1}{2}}, ..., \lambda_m^{\frac{1}{2}})$.
Re-labelling $\widetilde{\mu}_1$ as μ_1, and re-applying Kent and Mardia's result with this choice of μ_1 we see $A = \Lambda$, U is an identity matrix and $\mu_1^T \epsilon_1 = C$. Applying Kent and Mardia's approximation of the rotation matrix to $R = (X_1 \widehat{\Gamma}_1 - \mu_1)/\sigma$ gives,

$$\begin{aligned} \sigma R &= (\mu_1 + \sigma\epsilon_1)(I + \sigma D) - \mu_1 + O(\sigma^2), \\ R &= \epsilon_1 + \mu_1 D + O(\sigma), \\ &\approx \epsilon_1 + \mu_1(C^T * L) - \mu_1(C * L), \end{aligned}$$

assuming σ is small. Given μ_1 only has non-zero entries on the diagonal, then the p,qth element of $(C^T * L)$ and $C * L$ are,

$$(C^T * L)_{pq} = \frac{\mu_{1qq}\epsilon_{1qp}}{\lambda_p + \lambda_q} \qquad (C * L)_{pq} = \frac{\mu_{1pp}\epsilon_{1pq}}{\lambda_p + \lambda_q} \tag{2.2}$$

Considering the p, qth element of R, $1 \leq p, q \leq m$. Then,

$$\begin{aligned} (R)_{pq} &\approx \epsilon_{1pq} + \sum_{i=1}^{m} \mu_{1pi}(C^T * L)_{iq} - \sum_{i=1}^{m} \mu_{1pi}(C * L)_{iq}, \\ &= \epsilon_{1pq} + \mu_{1pp}\left(\frac{\mu_{1qq}\epsilon_{1qp}}{\lambda_p + \lambda_q} - \frac{\mu_{1pp}\epsilon_{1pq}}{\lambda_p + \lambda_q}\right), \\ &= \frac{\lambda_q^{\frac{1}{2}}}{\lambda_p + \lambda_q}\left(\lambda_q^{\frac{1}{2}}\epsilon_{1pq} + \lambda_p^{\frac{1}{2}}\epsilon_{1qp}\right). \end{aligned}$$

For the case, $p > m$, then $(R)_{pq} = \epsilon_{1pq}$, which agrees with Equation (4.1) of Goodall (1995). Following the proof of Goodall (1991), the distributional result in Equation (2.1) is obtained, due to $(R)_{pq} = \lambda_q^{\frac{1}{2}} \lambda_p^{-\frac{1}{2}} (R)_{qp}$ removing $m(m-1)/2$ degrees of freedom and $\epsilon_1^T 1_{k_1} = 0_m$ removing a further m degrees of freedom. □

Using the approximation of Kent and Mardia (2001), we have been able to derive this distributional result, first given by Sibson (1979). We now apply this result to some simulated data in a subset matching context. A template was created with landmarks chosen at random in 2 dimensions. Then the configuration X differed at each landmark from the template by simulated values from a Gaussian distribution with zero mean and standard deviation, $\sigma = 0.01$. From Figure 2.1 it can be seen that the density distribution of the X_1 landmarks given by the histogram follows the distribution specified in Result 2.2.2, where $(k_1 - 1)m - \frac{1}{2}m(m-1) = 2k_1 - 3$ when $m = 2$.

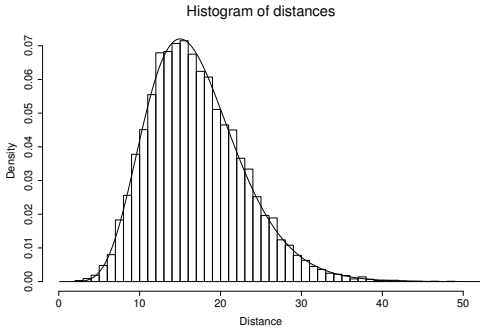

Figure 2.1: Histogram shows D_S with $k_1 = 10$, $\sigma = 0.01$. Line shows $\chi^2_{2k_1-3}$.

2.2.2 Distribution of landmarks in the complement

Now consider the rest of the configuration, after applying the same centring and rotation matrices determined from, and applied to, the subset.

Result 2.2.3 *Following minimisation by rigid-body transformations of X_1 to μ_1, the approximate distribution of the sum of Euclidean squared distances between landmarks in the complement is given by,*

$$D_C = \left\| \frac{X_2\widehat{\Gamma}_1 - \mu_2}{\sigma} \right\|^2 \sim \sum_{i=1}^{k_2 m} \eta_i U_i^2, \tag{2.3}$$

where, without loss of generality, X_2 and μ_2 have been translated using the centring of X_1 and μ_1, respectively, $U_i^2 \sim \chi_1^2$, and η_i are the eigenvalues of $P_2 C_{1km} C_{1km}^T P_2^T$, where $C_{1km} = C_{1k} \otimes I_m$, $C_{1k} = I_k - \frac{1}{k} 1_k (1_{k_1}^T 0_{k_2}^T)$ and P_2 is a $k_2 m \times km$ matrix such that $(X_2 \widehat{\Gamma} - \mu_2)/\sigma = P_2 vec([\epsilon_1^T, \epsilon_2^T])$.

Proof: Using the result from Kent and Mardia (2001) and the simplification given in Result 2.2.2, we express the difference between the rotated, translated configuration and the template, in the complement, as $Q = (X_2 \widehat{\Gamma}_1 - \mu_2)/\sigma$. Hence,

$$\begin{aligned} \sigma Q &= (\mu_2 + \sigma\epsilon_2)(I + \sigma D) - \mu_2 + O(\sigma^2), \\ Q &= \epsilon_2 + \mu_2 D + O(\sigma), \\ &\approx \epsilon_2 + \mu_2 (C^T * L) - \mu_2 (C * L), \end{aligned}$$

assuming σ is small.
Applying Equation (2.2), we consider the p,qth element of Q, $p = 1, ..., k_2$, $q = 1, ..., m$,

$$\begin{aligned} (Q)_{pq} &\approx \epsilon_{2pq} + \sum_{i=1}^{m} \mu_{2pi} (C^T * L)_{iq} - \sum_{i=1}^{m} \mu_{2pi} (C * L)_{iq}, \\ &= \epsilon_{2pq} + \sum_{i=1}^{m} \mu_{2pi} \left(\frac{\epsilon_{1qi} \mu_{1qq}}{\lambda_i + \lambda_q} \right) - \sum_{i=1}^{m} \mu_{2pi} \left(\frac{\mu_{1ii} \epsilon_{1iq}}{\lambda_i + \lambda_q} \right), \\ &= \epsilon_{2pq} + \sum_{i=1}^{m} \frac{\mu_{2pi}}{\lambda_i + \lambda_q} \left(\lambda_q^{\frac{1}{2}} \epsilon_{1qi} - \lambda_i^{\frac{1}{2}} \epsilon_{1iq} \right). \end{aligned}$$

We write, $\text{vec}(Q) = P_2 \text{vec}([\epsilon_1^T, \epsilon_2^T])$, where P_2 is a $k_2 m \times km$ matrix. Assume, without loss of generality, that μ_1 and X_1 are centred, by initially premultiplying the configurations by a centring matrix, $C_{1k} = I_k - \frac{1}{k} 1_k (1_{k_1}^T 0_{k_2}^T)$, to centre on the subset rather than the whole configuration. Therefore, $\text{vec}(\epsilon) \sim N(0, C_{1km} C_{1km}^T)$, where $C_{1km} = C_{1k} \otimes I_m$, and $\text{vec}(Q) \sim N(0, P_2 C_{1km} C_{1km}^T P_2^T = \Sigma_2)$.

Let $\Sigma_2 = R\Lambda R^T$, by the spectral decomposition theorem, and $Y = \Sigma_2^{-\frac{1}{2}}\text{vec}(Q) \sim N(0, \Sigma_2^{-\frac{1}{2}}\Sigma_2\Sigma_2^{-\frac{1}{2}} = I)$, $U = R^TY \sim N(0, RR^T = I)$, then using standard results regarding quadratic forms (Mathai and Provost, 1992),

$$\|\text{vec}(Q)\|^2 = \text{vec}(Q)^T\text{vec}(Q) \approx Y^T\Sigma_2 Y = Y^T R\Lambda R^T Y = U^T\Lambda U = \sum_{i=1}^{k_2 m} \eta_i U_i^2,$$

where $U_i \sim N(0,1)$ and η_i are the eigenvalues of Σ_2. The result then follows. $\square$.

The linear combination of χ^2 variables can be approximated using a Satterthwaite approximation, see Johnson *et al.* (1994) for details. Given, $\mathbb{E}(D_C) = \sum_i \eta_i$ and $\text{Var}(D_C) = 2\sum_i \eta_i^2$, we can equate these with the first two moments of $a\chi_r^2$, namely ar and $2a^2r$, to give $D_C \approx a\chi_r^2$ where,

$$a = \frac{\sum_{i=1}^{k_2 m} \eta_i^2}{\sum_{i=1}^{k_2 m} \eta_i}, \qquad r = \frac{(\sum_{i=1}^{k_2 m} \eta_i)^2}{\sum_{i=1}^{k_2 m} \eta_i^2},$$

which can be solved numerically. In the case $m = 2$, a and r can be calculated using,

$$\begin{aligned}
\sum_{i=1}^{k_2 m} \eta_i &= \frac{2k_2(k_1+1)}{k_1} + \frac{(k_1-1)}{k_1(\lambda_1+\lambda_2)} \sum_{p=1}^{k_2}\sum_{q=1}^{2} \mu_{2pq}^2, \\
\sum_{i=1}^{k_2 m} \eta_i^2 &= \frac{1}{k_1^2}\Bigg(2k_2(k_2 + 2k_1 + k_1^2) + \frac{2k_1(k_1-1)}{\lambda_1+\lambda_2} \sum_{p=1}^{k_2}\sum_{q=1}^{2} \mu_{2pq}^2 \\
&+ \frac{2(k_1-1)}{\lambda_1+\lambda_2} \left(\sum_{q=1}^{2} \left(\sum_{p=1}^{k_2} \mu_{2pq} \right)^2 \right) + \frac{(k_1-1)^2}{(\lambda_1+\lambda_2)^2} \left(\sum_{p=1}^{k_2}\sum_{q=1}^{2} \mu_{2pq}^2 \right)^2 \Bigg),
\end{aligned}$$

where $\lambda_1 = \mu_{111}^2$ and $\lambda_2 = \mu_{122}^2$, the only non-zero entries of μ_1.

Returning to the simulation, the density distribution of the X_2 landmarks was plotted as a histogram, see Figure 2.2. The exact values of η

were calculated and in the first column, the line shows the density of the linear combination of χ^2 variables. Note that this line is not smooth because the density is based on 10,000 simulations rather than an exact distribution. In the second column the line shows the Satterthwaite approximation, based on the totals of the eigenvalues. It can be seen that the approximation is very good, although with low values of k_1 the result is less obvious.

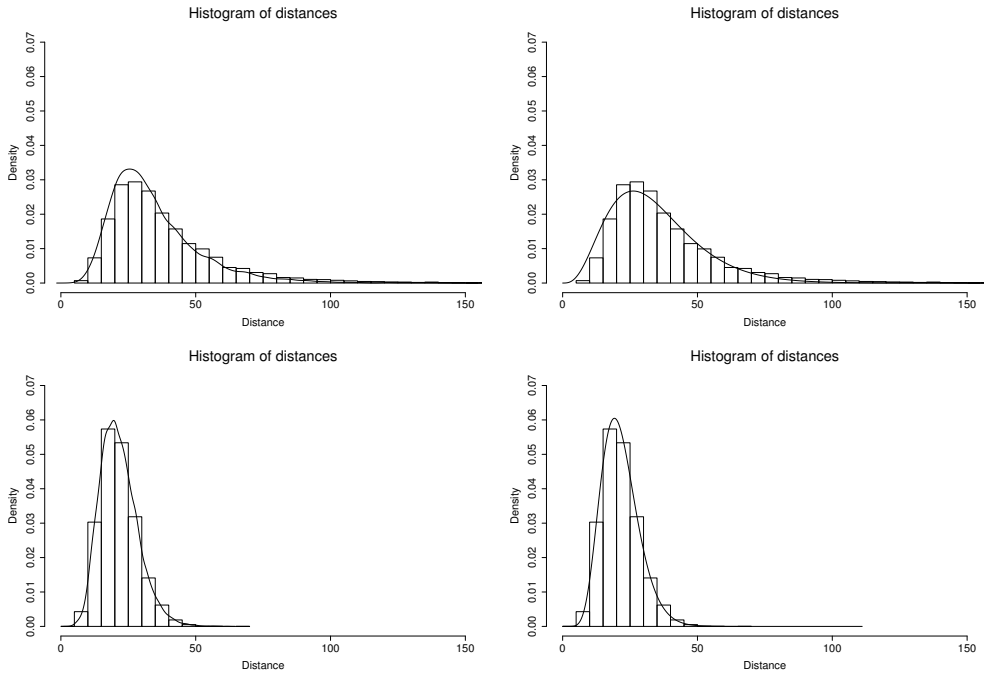

Figure 2.2: Histograms show D_C with $k_2 = 10$, $\sigma = 0.01$. The rows show the distribution with $k_1 = 3, 30$ respectively. In the first column, the line shows the distribution of $\sum \eta_i U_i^2$. In the second column it shows the distribution of $a\chi_r^2$.

Different combinations of k_1 and k_2 were used in the simulations. The number of X_2 landmarks has no influence on the distribution of X_1 landmarks. However, the X_2 distribution is affected by the value of k_1, see Figure 2.2. Table 2.1 shows the results of simulations carried out with $k_2 = 10$. In each case, the first three eigenvalues are shown in the table, and the

remaining eigenvalues, $\eta_4 = ... = \eta_{20} = 1$. Note that as $k_1 \to \infty$, $\eta_i \to 1$. Therefore, $\sum \eta_i \to k_2 m$ and $\sum \eta_i^2 \to k_2 m$. Hence, $a \to 1$ and $r \to k_2 m$. This implies $\mathbb{E}(D_C) \to k_2 m$ and $\text{Var}(D_C) \to 2k_2 m$.

k_1	η_1, η_2, η_3	$\sum \eta_i$	$\sum \eta_i^2$	a	r	$\mathbb{E}(a\chi_r^2)$	$\text{Var}(a\chi_r^2)$
3	10.53, 4.33, 4.09	36.0	163.3	4.5	7.9	35.55	320.0
10	5.62, 2.00, 1.98	26.6	56.4	2.1	12.5	26.25	110.3
30	1.81, 1.33, 1.19	21.3	23.5	1.1	19.4	21.34	46.9
50	1.54, 1.20, 1.18	20.9	22.2	1.1	19.7	21.67	47.7

Table 2.1: The approximate expected value and variance for $D_C \approx a\chi_r^2$ for different values of k_1 given $k_2 = 10$.

In the simulations we tried different values for the variance of the normal distribution from which the errors were simulated. As expected, for low values of σ our results held but as $\sigma \to \infty$ the distribution of $\|(X_1\Gamma_1 - \mu_1)/\sigma\|^2 \to \chi^2_{2k_1-2}$ because X_1 is centred, see Figure 2.3. Our expected distribution of $\|(X_2\Gamma_1 - \mu_2)/\sigma\|^2$ consistently over-estimates the correct mean for high values of σ, see Table 2.2

σ	$\bar{D}_S$	$\mathbb{E}(\chi^2_{2k_1-3})$	$\bar{D}_C$	$\mathbb{E}(\sum \eta_i \chi_1^2)$	$\mathbb{E}(a\chi_r^2)$
0.01	17.02	17.00	24.72	24.43	24.54
0.1	17.06	17.00	25.02	24.66	24.68
1	17.04	17.00	26.47	25.75	25.79
10	17.11	17.00	23.30	25.14	25.34
100	17.90	17.00	22.07	24.77	24.85
1000	18.00	17.00	22.00	24.18	24.18

Table 2.2: The mean value of D_S and D_C from simulations compared to theoretical values and an approximate theoretical value, for $k_1 = 10$, $k_2 = 10$.

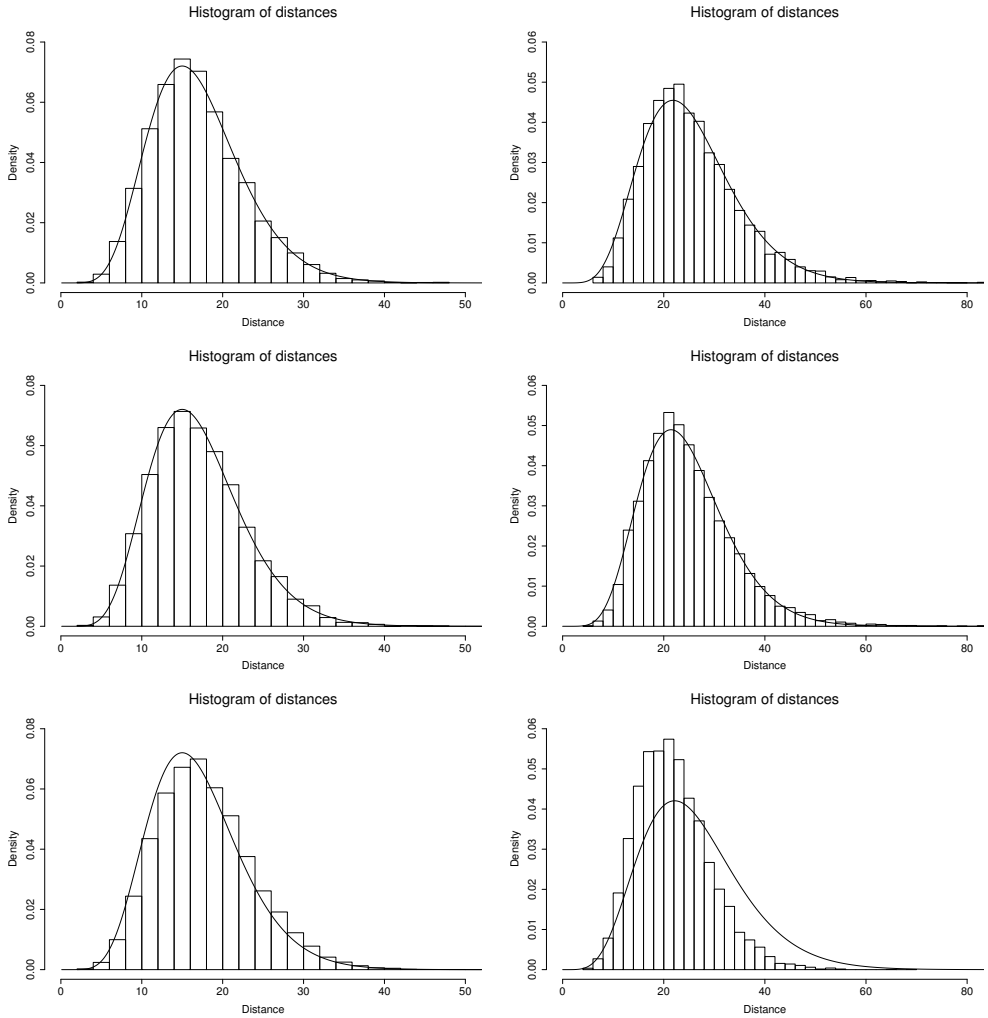

Figure 2.3: Histograms show distribution with $k_1 = 10$, $k_2 = 10$. The rows show the distribution with $\sigma = 0.01, 1, 10$ respectively. In the first column, the histogram shows distribution of D_S and the line shows the distribution of $\chi^2_{2k_1-3}$. In the second column, the histogram shows distribution of D_C and the line shows the distribution of $a\chi^2_r$.

2.2.3 Distribution of landmarks with respect to an alternative complement template

We extend the simulation to include a 2 dimensional slice of a brain-like configuration, where there are 9 landmarks forming a midline and 40 landmarks forming a circle to represent the surface of the brain, as shown in Figure 1.3. Using the midline landmarks as X_1 and the circle as X_2, the distribution followed our expectations. The brain surface was then perturbed to make it asymmetric. Frequently, a patient's brain is asymmetric, but it may be desirable to measure the distance to a symmetric template. In this case, allow μ to represent the asymmetric template about which points are perturbed, and let Φ be the circle.

Result 2.2.4 *Following minimisation by rigid-body transformations of X_1 to μ_1, the approximate distribution of the sum of Euclidean squared distances between landmarks in the complement of the shape, X_2, and the complement of an alternative template, Φ_2, is given by,*

$$D_C = \left\| \frac{X_2\widehat{\Gamma}_1 - \Phi_2}{\sigma} \right\|^2 \sim \sum_{i=1}^{k_2 m} \eta_i V_i^2, \tag{2.4}$$

where $V_i^2 \sim \chi_1^2(\zeta_i)$, η_i are the eigenvalues of $\Sigma_2 = P_2 C_{1km} C_{1km}^T P_2^T$ and each ζ_i is a function of $\mu_2 - \Phi_2$ and Σ.

Proof: Using the result from Kent and Mardia (2001) and the simplification given in Result 2.2.2, we express the difference between the rotated, translated configuration and the alternative template, in the complement, as $Q = (X_2\widehat{\Gamma}_1 - \Phi_2)/\sigma$. Hence,

$$\begin{aligned} \sigma Q &= (\mu_2 + \sigma\epsilon_2)(I + \sigma D) - \Phi_2 + O(\sigma^2), \\ Q &= \frac{\mu_2 - \Phi_2}{\sigma} + (\epsilon_2 + \mu_2 D) + O(\sigma). \end{aligned}$$

From the proof of Result 2.2.3 we know,

$$\text{vec}(Q) \approx \Delta + P_2\text{vec}([\epsilon_1^T\epsilon_2^T]) \sim N(\Delta, P_2C_{1km}C_{1km}^TP_2^T = \Sigma_2),$$

where $\Delta = \text{vec}\left(\frac{\mu_2-\Phi_2}{\sigma}\right)$. Let $Z = \Sigma_2^{-\frac{1}{2}}(\text{vec}(Q) - \Delta) \sim N(0, I)$, $\Sigma_2 = R\Lambda R^T$ by the spectral decomposition theorem, and $U = R^TZ \sim N(0, RR^T = I)$. Then, $\text{vec}(Q) = \Sigma_2^{\frac{1}{2}}(Z + \Sigma_2^{-\frac{1}{2}}\Delta)$, and,

$$\begin{aligned}
\text{vec}(Q)^T\text{vec}(Q) &\approx (Z + \Sigma_2^{-\frac{1}{2}}\Delta)^TR\Lambda R^T(Z + \Sigma_2^{-\frac{1}{2}}\Delta), \\
&= (R^TZ + R^T\Sigma_2^{-\frac{1}{2}}\Delta)^T\Lambda(R^TZ + R^T\Sigma_2^{-\frac{1}{2}}\Delta), \\
&= (U + \lambda)^T\Lambda(U + \lambda), \\
&= \sum_{i=1}^{k_2m}\eta_iV_i^2,
\end{aligned}$$

where $V_i \sim N(\lambda_i, 1)$ and λ_i is the ith element of $\lambda = R^T\Sigma_2^{-\frac{1}{2}}\Delta$. Therefore, $V_i^2 \sim \chi_1^2(\zeta_i)$ where $\zeta_i = \lambda_i^2$. □

Given the complexity of the non-central chi-squared distribution, it seems sensible to approximate this result using a scaled central chi-squared distribution (Johnson *et al.*, 1995), i.e. the Satterthwaite approximation. If we let $\sum\eta_i\chi_1^2(\zeta_i) = a\chi_r^2$, we can calculate values for a and r. Equating results for the first two moments gives,

$$a = \frac{\sum\eta_i^2(1 + 2\zeta_i)}{\sum\eta_i(1 + \zeta_i)}, \qquad r = \frac{(\sum\eta_i(1 + \zeta_i))^2}{\sum\eta_i^2(1 + 2\zeta_i)}.$$

An improved approximation would be of the form $c\chi_f^2 + b$, Equating the first three moments gives,

$$\begin{aligned}
\sum_i\eta_i(1 + \zeta_i) &= cf + b, \qquad 2\sum_i\eta_i^2(1 + 2\zeta_i) = 2c^2f, \\
8\sum_i\eta_i^3(1 + 3\zeta_i) &= 8c^3f,
\end{aligned}$$

which can be solved to give,

$$\begin{aligned} c &= \frac{\sum \eta_i^3(1+3\zeta_i)}{\sum \eta_i^2(1+2\zeta_i)}, \qquad f = \frac{(\sum \eta_i^2(1+2\zeta_i))^3}{(\sum \eta_i^3(1+3\zeta_i))^2}, \\ b &= \sum \eta_i(1+\zeta_i) - \frac{(\sum \eta_i^2(1+2\zeta_i))^2}{\sum \eta_i^3(1+3\zeta_i)}. \end{aligned}$$

This method was applied to our simulated 2-dimensional brain slice. Due to the midline having the same landmarks in the asymmetric template and the circle, the distribution of X_1 landmarks was the same as before. Figure 2.4 shows that the theoretical distribution for D_C seems to be fairly accurate, with the other distributions being good approximations. Using three parameters produces a noticeable benefit.

The distributional results presented in this section could allow us to quantify theoretically the efficiency of the subset matching estimator and quantify its variance.

2.3 Covariance weighted OPA

In Chapter 1 full isotropic ordinary Procrustes analysis was defined as obtaining the Procrustes estimators of translation, rotation and scale to minimise the Euclidean distance, $D^2_{OPA}(X,\mu) = \|\mu - X^P\|^2$. Euclidean distance is a special case of the Mahalanobis norm $\|\mu - X^P\|^2_\Sigma = \text{vec}(\mu - X^P)^T\Sigma^{-1}\text{vec}(\mu - X^P)$ with $\Sigma = I_{km}$. In Section 2.3.1 partial covariance weighted Procrustes analysis is defined with respect to the Mahalanobis norm and expressions for the minimising translation and rotation are given. This is extended to include scaling in Section 2.3.2, before consideration is given to some special cases.

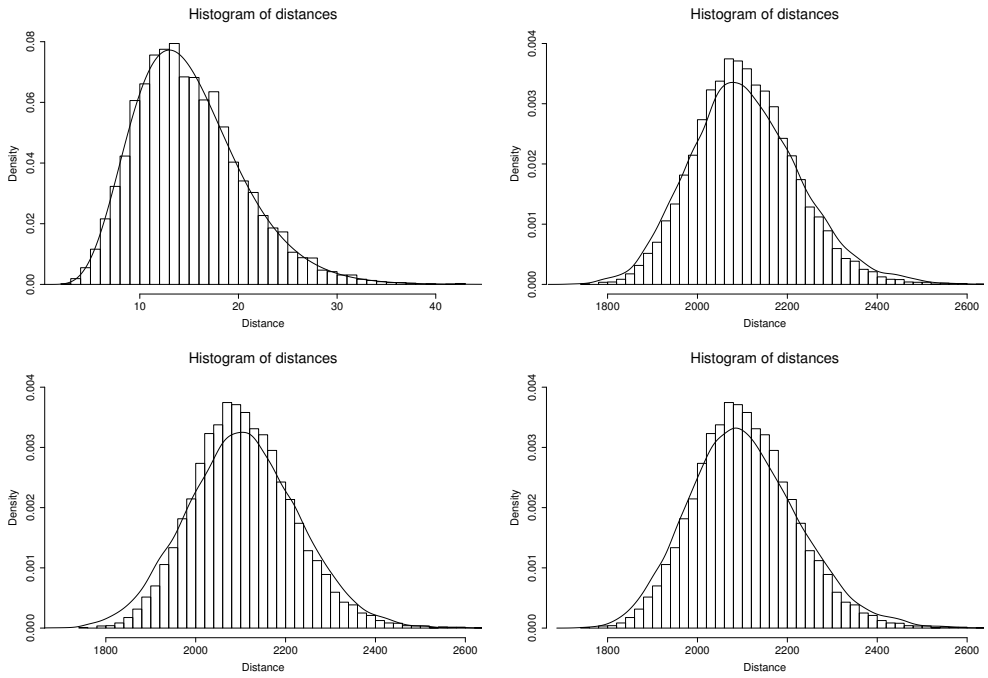

Figure 2.4: Top left, the histogram shows D_S with the expected distribution. The other histograms shows D_C with $k_1 = 9$, $k_2 = 40$, $\sigma = 0.01$. Top right, the line shows the distribution of $\sum \eta_i \chi^2_1(\zeta_i)$, simulated using $N(\eta_i\sqrt{\zeta_i}, \eta_i)$. The bottom row shows the approximate distribution using the first two moments on the left and the first three moments on the right.

2.3.1 Partial covariance weighted OPA

Definition 2.3.1 *The method of partial covariance weighted ordinary Procrustes analysis (partial CW OPA) involves the least squares matching of one configuration to another using rigid-body transformations. Estimation of the translation and rotation parameters, γ and Γ, is carried out by minimising the Mahalanobis norm,*

$$D^2_{pCWP}(X, \mu; \Sigma) = \|\mu - X\Gamma - 1_k\gamma^T\|^2_\Sigma, \tag{2.5}$$

where Σ $(km \times km)$ is a symmetric positive definite matrix, γ is a $m \times 1$ location vector and Γ is an $m \times m$ special orthogonal rotation matrix.

The translation which minimises Equation (2.5) is given by Result 2.3.1. In general, the minimising rotation is solved numerically, however when $m = 2$ there is only one rotation angle and a solution is given by Result 2.3.2.

Result 2.3.1 *Given two configuration matrices, X and μ, and a symmetric positive definite matrix, Σ, the translation, as a function of rotation, which minimises the Mahalanobis norm, $D^2_{pCWP}(X, \mu; \Sigma)$ is,*

$$\widehat{\gamma} = [(I_m \otimes 1_k)^T\Sigma^{-1}(I_m \otimes 1_k)]^{-1}(I_m \otimes 1_k)^T\Sigma^{-1}\text{vec}(\mu - X\Gamma). \tag{2.6}$$

Proof: Let $v = \text{vec}(\mu - X\Gamma)$ then,

$$\begin{aligned} D^2_{pCWP}(X, \mu; \Sigma) &= (v - (I_m \otimes 1_k)\gamma)^T\Sigma^{-1}(v - (I_m \otimes 1_k)\gamma), \\ &= v^T\Sigma^{-1}v - 2v^T\Sigma^{-1}(I_m \otimes 1_k)\gamma \\ &\quad +\gamma^T(I_m \otimes 1_k)^T\Sigma^{-1}(I_m \otimes 1_k)\gamma. \end{aligned}$$

The minimising translation is found by setting the first derivative equal to zero,

$$\frac{dD^2_{pCWP}}{d\gamma} = -2(I_m \otimes 1_k)^T\Sigma^{-1}v + 2(I_m \otimes 1_k)^T\Sigma^{-1}(I_m \otimes 1_k)\gamma = 0.$$

The second derivative is clearly positive because Σ^{-1} is positive definite. Therefore D^2_{pCWP} is minimised when $\gamma = [(I_m \otimes 1_k)^T\Sigma^{-1}(I_m \otimes 1_k)]^{-1}(I_m \otimes 1_k)\Sigma^{-1}v$. $\square$

Result 2.3.2 *Assuming that* $m = 2$, *let* $A = [(I_m \otimes 1_k)^T\Sigma^{-1}(I_m \otimes 1_k)]^{-1}(I_m \otimes 1_k)^T\Sigma^{-1}$, *and denote the partitioned submatrices as,*

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \qquad X = \begin{bmatrix} X_1 & X_2 \end{bmatrix}, \qquad \mu = \begin{bmatrix} \mu_1 & \mu_2 \end{bmatrix},$$

where A_{ij} *have dimension* $(1 \times k)$ *and* X_i, μ_i *have dimension* $(k \times 1)$ *for* $i, j = 1, 2$, *then given two configuration matrices,* X *and* μ, *and a symmetric positive definite matrix* Σ, *the rotation which minimises the Mahalanobis norm,* $D^2_{pCWP}(X, \mu; \Sigma)$ *is given by,*

$$\begin{aligned} \cos\widehat{\theta} &= \frac{S(2\lambda - 2Q) + TR}{(2\lambda - 2P)(2\lambda - 2Q) - R^2}, \\ \sin\widehat{\theta} &= \frac{T(2\lambda - 2P) + SR}{(2\lambda - 2P)(2\lambda - 2Q) - R^2}, \end{aligned} \tag{2.7}$$

where,

$$\begin{aligned} P &= \begin{bmatrix} (X_1 + 1_k\delta_1) \\ (X_2 + 1_k\delta_2) \end{bmatrix}^T \Sigma^{-1} \begin{bmatrix} (X_1 + 1_k\delta_1) \\ (X_2 + 1_k\delta_2) \end{bmatrix}, \\ Q &= \begin{bmatrix} (X_2 - 1_k\zeta_1) \\ -(X_1 + 1_k\zeta_2) \end{bmatrix}^T \Sigma^{-1} \begin{bmatrix} (X_2 - 1_k\zeta_1) \\ -(X_1 + 1_k\zeta_2) \end{bmatrix}, \end{aligned}$$

$$
\begin{aligned}
R &= -2\begin{bmatrix} (X_1 + 1_k\delta_1) \\ (X_2 + 1_k\delta_2) \end{bmatrix}^T \Sigma^{-1} \begin{bmatrix} (X_2 - 1_k\zeta_1) \\ -(X_1 + 1_k\zeta_2) \end{bmatrix}, \\
S &= -2\begin{bmatrix} (X_1 + 1_k\delta_1) \\ (X_2 + 1_k\delta_2) \end{bmatrix}^T \Sigma^{-1} \begin{bmatrix} (\mu_1 - 1_k\alpha_1) \\ (\mu_2 - 1_k\alpha_2) \end{bmatrix}, \\
T &= 2\begin{bmatrix} (X_2 - 1_k\zeta_1) \\ -(X_1 + 1_k\zeta_2) \end{bmatrix}^T \Sigma^{-1} \begin{bmatrix} (\mu_1 - 1_k\alpha_1) \\ (\mu_2 - 1_k\alpha_2) \end{bmatrix}, \\
\alpha_i &= A_{i1}\mu_1 + A_{i2}\mu_2, \\
\delta_i &= -A_{i1}X_1 - A_{i2}X_2, \\
\zeta_i &= A_{i1}X_2 - A_{i2}X_1,
\end{aligned}
\tag{2.8}
$$

and λ *is the real root less than* $\frac{1}{2}\left(P + Q - \sqrt{(P-Q)^2 + R^2}\right)$ *of the quartic equation,*

$$
\begin{aligned}
16\lambda^4 - 32(P+Q)\lambda^3 \\
+[16(P^2+Q^2) + 64PQ - 4(S^2+T^2) - 8R^2]\lambda^2 \\
+[8R^2(P+Q) - 32PQ(P+Q) + 8(QS^2 + PT^2 - STR)]\lambda \\
+16P^2Q^2 + R^4 - R^2(S^2+T^2) + 4RST(P+Q) \\
-4P^2T^2 - 4Q^2S^2 - 8PQR^2 &= 0.
\end{aligned}
\tag{2.9}
$$

Proof: From Equation (2.6) the minimising translation is $\widehat{\gamma} = A\text{vec}(\mu - X\Gamma)$, so for $m = 2$,

$$
\begin{bmatrix} \widehat{\gamma}_1 \\ \widehat{\gamma}_2 \end{bmatrix} = \begin{bmatrix} \alpha_1 + \delta_1\cos\theta + \zeta_1\sin\theta \\ \alpha_2 + \delta_2\cos\theta + \zeta_2\sin\theta \end{bmatrix}, \text{ because, } \Gamma = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}.
$$

Therefore,

$$
\begin{aligned}
&\text{vec}(\mu - X\Gamma - 1_k\gamma^T) \\
&\quad = \begin{bmatrix} (\mu_1 - 1_k\alpha_1) - (X_1 + 1_k\delta_1)\cos\theta + (X_2 - 1_k\zeta_1)\sin\theta \\ (\mu_2 - 1_k\alpha_2) - (X_2 + 1_k\delta_2)\cos\theta - (X_1 + 1_k\zeta_2)\sin\theta \end{bmatrix},
\end{aligned}
$$

and $D^2_{pCWP}(X,\mu;\Sigma) = C + P\cos^2\theta + Q\sin^2\theta + R\cos\theta\sin\theta + S\cos\theta + T\sin\theta$ where,

$$C = \begin{bmatrix} (\mu_1 - 1_k\alpha_1) \\ (\mu_2 + 1_k\alpha_2) \end{bmatrix}^T \Sigma^{-1} \begin{bmatrix} (\mu_1 - 1_k\alpha_1) \\ (\mu_2 + 1_k\alpha_2) \end{bmatrix}.$$

Let λ be the real Lagrangian multiplier to enforce the constraint $\cos^2\theta + \sin^2\theta = 1$ and let $L = D^2_{pCWP}(X,\mu;\Sigma) + \lambda(1 - \cos^2\theta - \sin^2\theta)$. Then,

$$\begin{aligned} \frac{\partial L}{\partial(\cos\theta)} &= 2(P-\lambda)\cos\theta + R\sin\theta + S = 0, \\ \frac{\partial L}{\partial(\sin\theta)} &= 2(Q-\lambda)\sin\theta + R\cos\theta + T = 0, \\ \frac{\partial L}{\partial\lambda} &= 1 - \cos^2\theta - \sin^2\theta = 0. \end{aligned}$$

Solving the first two equations simultaneously and substituting the solutions in the third, gives the expressions for $\cos\theta$, $\sin\theta$ and the quartic equation respectively. To show this is a minimum of D^2_{pCWP}, consider the matrix of second derivatives,

$$S^* = \begin{bmatrix} \frac{\partial^2 L}{\partial(\cos^2\theta)} & \frac{\partial^2 L}{\partial(\cos\theta)\partial(\sin\theta)} \\ \frac{\partial^2 L}{\partial(\cos\theta)\partial(\sin\theta)} & \frac{\partial^2 L}{\partial(\sin^2\theta)} \end{bmatrix} = \begin{bmatrix} 2(P-\lambda) & R \\ R & 2(Q-\lambda) \end{bmatrix}.$$

Let $\xi_1 \geq \xi_2$ be the eigenvalues of S^*. Then, $|S^* - \xi_i I| = (\xi_i + 2\lambda)^2 - 2(P+Q)(\xi_i + 2\lambda) + 4PQ - R^2$, so $(\xi_i + 2\lambda) = P + Q \pm \sqrt{(P-Q)^2 + R^2}$. Given Σ^{-1} is positive definite, $P > 0$ and $Q > 0$, then ξ_2 is strictly positive if $P + Q - 2\lambda - \sqrt{(P-Q)^2 + R^2} > 0$ which is true if the constraint on λ is satisfied. $\square$

Note that a unique solution of Equation (2.9) that satisfies the constraint may not exist and it may be necessary to evaluate $D^2_{pCWP}(X,\mu;\Sigma)$ for several choices of λ or use numerical methods. In our experience, however,

this is rarely required.

Figure 2.5 shows the partial ordinary covariance weighted Procrustes registration of one second thoracic mouse vertebra to another for three different weighting matrices, Σ. Note there are $k = 6$ landmarks in $m = 2$ dimensions taken at points of maximum curvature from a cross-section. Starting at the far left and going clockwise, the landmarks are numbered: 4, 3, 2, 6, 1, 5. The weighting matrices used are:

$$\Sigma_1 = I_{km}, \qquad \Sigma_2 = I_m \otimes \Sigma_K, \qquad \Sigma_3 = \Sigma_M \otimes \Sigma_K, \tag{2.10}$$

where,

$$\Sigma_M = \begin{bmatrix} 10 & 0 \\ 0 & 0.1 \end{bmatrix}, \qquad \Sigma_K = \begin{bmatrix} 0.1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 10 & 0 & 0 & 0 \\ 0 & 0 & 0 & 10 & 0 & -9 \\ 0 & 0 & 0 & 0 & 10 & 0 \\ 0 & 0 & 0 & -9 & 0 & 10 \end{bmatrix}.$$

The first example is the same as isotropic OPA, giving the same weighting to all co-ordinates and minimising the total distance between all pairs of landmarks. The second example weights landmarks one and two more heavily than the others, and this is reflected in these landmarks being closely matched at the expense of introducing large variability elsewhere. The third example goes a stage further, and weights the y direction more heavily than the x direction. This forces a large rotation to ensure landmarks 1 and 2 have similar y values on both the template and the dotted image.

2.3.2 Full covariance weighted OPA

Definition 2.3.2 *The method of full covariance weighted ordinary Procrustes analysis (full CW OPA) involves the least squares matching of one*

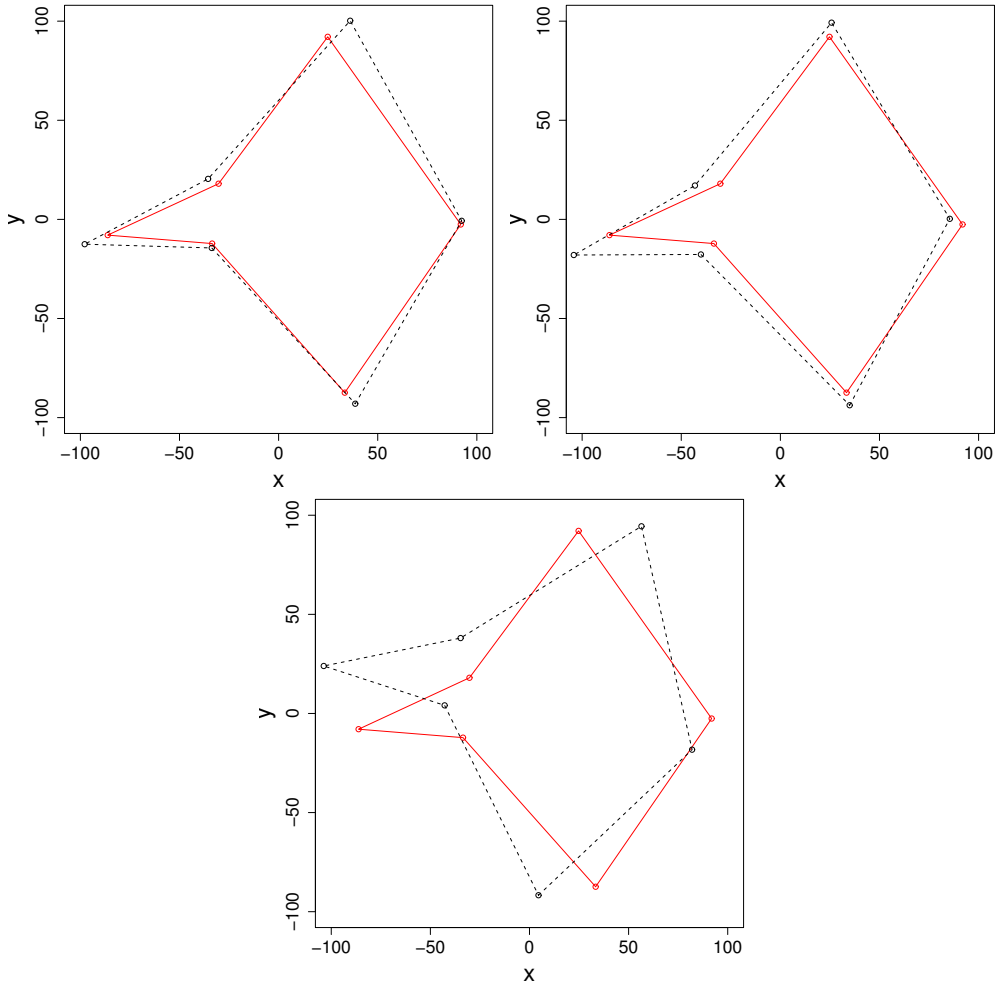

Figure 2.5: The partial ordinary covariance weighted Procrustes registration of a mouse vertebra (dotted) to a template, using Σ_1 (top left), Σ_2 (top right) and Σ_3 (bottom).

configuration to another using similarity transformations. Estimation of the translation, rotation and scaling parameters, γ, Γ and β, is carried out by minimising the Mahalanobis norm

$$D^2_{CWP}(X, \mu; \Sigma) = \|\mu - \beta X\Gamma - 1_k\gamma^T\|^2_\Sigma, \tag{2.11}$$

where Σ $(km \times km)$ is a symmetric positive definite matrix, γ is a $m \times 1$ location vector, Γ is an $m \times m$ special orthogonal rotation matrix and $\beta > 0$ is a scale parameter.

In general, the minimising rotation is solved numerically and the minimising translation and scaling are given by Result 2.3.3. However, when $m = 2$ all the similarity transformation parameters can be obtained by Result 2.3.4.

Result 2.3.3 *Given two configuration matrices, X and μ, and a symmetric positive definite matrix Σ, the translation and scaling, as a function of rotation, which minimise the Mahalanobis norm, $D^2_{CWP}(X, \mu; \Sigma)$ are,*

$$\begin{bmatrix} \widehat{\gamma} \\ \widehat{\beta} \end{bmatrix} = B^{-1} \begin{bmatrix} (I_m \otimes 1_k)^T\Sigma^{-1}\text{vec}(\mu) \\ \text{vec}(X\Gamma)^T\Sigma^{-1}\text{vec}(\mu) \end{bmatrix}, \tag{2.12}$$

where,

$$B = \begin{bmatrix} (I_m \otimes 1_k)^T\Sigma^{-1}(I_m \otimes 1_k) & (I_m \otimes 1_k)^T\Sigma^{-1}\text{vec}(X\Gamma) \\ \text{vec}(X\Gamma)^T\Sigma^{-1}(I_m \otimes 1_k) & \text{vec}(X\Gamma)^T\Sigma^{-1}\text{vec}(X\Gamma) \end{bmatrix}.$$

Proof: Let $v = \text{vec}(\mu)$ and $\xi = \text{vec}(X\Gamma)$ then,

$$\begin{aligned} D^2_{CWP}(X, \mu; \Sigma) &= (v - \beta\xi - (I_m \otimes 1_k)\gamma)^T\Sigma^{-1}(v - \beta\xi - (I_m \otimes 1_k)\gamma), \\ &= v^T\Sigma^{-1}v - 2\beta\xi^T\Sigma^{-1}v - 2v^T\Sigma^{-1}(I_m \otimes 1_k)\gamma + \beta^2\xi^T\Sigma^{-1}\xi \\ &\quad +2\beta\xi^T\Sigma^{-1}(I_m \otimes 1_k)\gamma + \gamma^T(I_m \otimes 1_k)^T\Sigma^{-1}(I_m \otimes 1_k)\gamma. \end{aligned}$$

This implies,

$$\begin{aligned}\frac{dD^2_{CWP}}{d\gamma} &= -2(I_m \otimes 1_k)^T\Sigma^{-1}v + 2\beta(I_m \otimes 1_k)^T\Sigma^{-1}\xi \\ &\quad +2(I_m \otimes 1_k)^T\Sigma^{-1}(I_m \otimes 1_k)\gamma, \\ \frac{dD^2_{CWP}}{d\beta} &= -2\xi^T\Sigma^{-1}v + 2\beta\xi^T\Sigma^{-1}\xi + 2\xi^T\Sigma^{-1}(I_m \otimes 1_k)\gamma.\end{aligned}$$

Therefore the minimum is at the solution of,

$$B\begin{bmatrix}\gamma \\ \beta\end{bmatrix} = \begin{bmatrix}(I_m \otimes 1_k)^T\Sigma^{-1}\mathrm{vec}(\mu) \\ \mathrm{vec}(X\Gamma)^T\Sigma^{-1}\mathrm{vec}(\mu)\end{bmatrix}.$$

The matrix of second derivatives is clearly positive because Σ^{-1} is positive definite. □

Result 2.3.4 *Assuming that* $m = 2$, *let* $A = [(I_m \otimes 1_k)^T\Sigma^{-1}(I_m \otimes 1_k)]^{-1}(I_m \otimes 1_k)^T\Sigma^{-1}$, *and define* P, Q, R, S *and* T *as in Equations (2.8), then given two configuration matrices,* X *and* μ, *and a symmetric positive definite matrix,* Σ, *the similarity transformation parameters which minimise the Mahalanobis norm,* $D^2_{CWP}(X, \mu; \Sigma)$ *are given by,*

$$\begin{aligned}\gamma &= A\mathrm{vec}(\mu - \beta X\Gamma), & \beta &= +\sqrt{\psi_1^2 + \psi_2^2}, \\ \cos\theta &= \frac{\psi_1}{\sqrt{\psi_1^2 + \psi_2^2}}, & \sin\theta &= \frac{\psi_2}{\sqrt{\psi_1^2 + \psi_2^2}},\end{aligned}$$

where,

$$\psi_1 = \frac{RT - 2QS}{4PQ - R^2}, \qquad \psi_2 = \frac{RS - 2PT}{4PQ - R^2}. \tag{2.13}$$

Proof: Replacing $\cos\theta$ with $\beta\cos\theta$ and $\sin\theta$ with $\beta\sin\theta$ in the proof of Result 2.3.2 gives $D^2_{CWP}(X, \mu; \Sigma) = C + P\beta^2\cos^2\theta + Q\beta^2\sin^2\theta + R\beta^2\cos\theta\sin\theta + S\beta\cos\theta + T\beta\sin\theta$. Let $\psi_1 = \beta\cos\theta$ and $\psi_2 = \beta\sin\theta$,

then,

$$\frac{dD^2_{CWP}}{d\psi_1} = 2P\psi_1 + R\psi_2 + S, \qquad \frac{dD^2_{CWP}}{d\psi_2} = 2Q\psi_2 + R\psi_1 + T.$$

Setting these expressions equal to zero and solving them simultaneously gives the required expressions for ψ_1 and ψ_2. Solving $\psi_1 = \beta\cos\theta$ and $\psi_2 = \beta\sin\theta$ subject to the constraint that $\cos^2\theta + \sin^2\theta = 1$ gives the rotation and scale parameters. Given these, the translation is obtained by letting $v = \text{vec}(\mu - \beta X\Gamma)$ in the proof of Result 2.3.1. □

2.3.3 Special case: $\Sigma = I_{km}$

If $\Sigma = I_{km}$, then covariance weighted OPA should give the same result as isotropic OPA. Note that when $m = 2$ we can represent the $k \times 2$ configuration matrix X of real co-ordinates as the $k \times 1$ vector of complex co-ordinates, with the real and imaginary components representing the x and y co-ordinates respectively. In this notation, assuming X and μ are centred, the full Procrustes fit for isotropic OPA has the solution,

$$\widehat{\gamma} = 0, \qquad \widehat{\theta} = \arg(X^*\mu), \qquad \widehat{\beta} = \frac{(X^*\mu\mu^*X)^{1/2}}{X^*X},$$

where X^* denotes the transpose of the complex conjugate of X.

Result 2.3.5 *If $\Sigma = I_{km}$ and assuming without loss of generality that X and μ are centred, then the covariance weighted OPA transformation parameter estimates are (I) $\widehat{\gamma} = 0$, (II) $\widehat{\beta} = tr(\mu^T X\widehat{\Gamma})/tr(X^T X)$ and for $m = 2$, (III) $\widehat{\gamma} = 0$, (IV) $\widehat{\theta} = arg(X^*\mu)$ and (V) $\widehat{\beta} = (X^*\mu\mu^*X)^{1/2}/(X^*X)$.*

Proof: If $\Sigma = I_{km}$, then Equation (2.6) gives,

$$\widehat{\gamma} = \frac{1}{k}(I_m \otimes 1_k)^T \text{vec}(\mu - X\Gamma).$$

X and μ centred implies $(I_m \otimes 1_k)^T \mathrm{vec}(X) = 0 = (I_m \otimes 1_k)^T \mathrm{vec}(\mu)$. Therefore $\widehat{\gamma} = 0$. If $\Sigma = I_{km}$, and X and μ are centred, then Equation (2.12) gives,

$$\begin{bmatrix} \gamma \\ \beta \end{bmatrix} = \begin{bmatrix} kI_m & 0_k \\ 0_k^T & \mathrm{tr}((X\Gamma)^T X\Gamma) \end{bmatrix}^{-1} \begin{bmatrix} 0_k \\ \mathrm{tr}(\mu X\Gamma) \end{bmatrix},$$

which has solutions $\widehat{\gamma} = 0$ and $\widehat{\beta} = \mathrm{tr}(\mu X\Gamma)/\mathrm{tr}(X^T X)$, therefore statements (I) and (II) are true.
For $m = 2$, $\widehat{\gamma} = 0$, as shown above for general m. Referring to the notation of Result 2.3.2, if $\Sigma = I_{km}$, then $A_{11} = A_{22} = \frac{1}{k} 1_k^T$ and $A_{12} = A_{21} = 0_k^T$. Then, from Equations (2.8), if X and μ are centred, then $\alpha_i = \delta_i = \zeta_i = 0$, for $i = 1, 2$, and P, Q, R, S and T simplify to,

$$\begin{aligned} P = Q &= X_1^T X_1 + X_2^T X_2, & R &= -2(X_1^T X_2 - X_2^T X_1) = 0, \\ S &= -2(X_1^T \mu_1 + X_2^T \mu_2), & T &= 2(X_2^T \mu_1 - X_1^T \mu_2). \end{aligned}$$

CASE 1: PARTIAL OPA. With $P = Q$ and $R = 0$, the quartic equation, Equation (2.9), reduces to,

$$\begin{aligned} 16\lambda^4 - 64P\lambda^3 + [96P^2 - 4(S^2 + T^2)]\lambda^2 + [8P(S^2 + T^2) - 64P^3]\lambda & \\ +16P^4 - 4P^2(S^2 + T^2) &= 0, \\ \Longrightarrow \qquad 16(\lambda - P)^4 - 4(S^2 + T^2)(\lambda - P)^2 &= 0, \end{aligned}$$

which has solutions $\lambda = P$, $\lambda = P \pm (S^2 + T^2)^{1/2}/2$. The constraint in Result 2.3.2 implies $\lambda = P - (S^2 + T^2)^{1/2}/2$. Therefore,

$$\cos\theta = \frac{S}{2\lambda - 2P} = \frac{-S}{(S^2 + T^2)^{1/2}}, \qquad \sin\theta = \frac{T}{2\lambda - 2P} = \frac{-T}{(S^2 + T^2)^{1/2}},$$

and $\tan\theta = T/S = (X_1^T \mu_2 - X_2^T \mu_1)/(X_1^T \mu_1 + X_2^T \mu_2)$. Note that $X^* \mu = (X_1^T \mu_1 + X_2^T \mu_2) + (X_1^T \mu_2 - X_2^T \mu_1)i$. Therefore $\theta = \arg(X^* \mu)$.

CASE 2: FULL OPA. With $P = Q$ and $R = 0$, Equations (2.13) reduce to $\psi_1 = -S/2P$ and $\psi_2 = -T/2P$. Therefore, $\psi_2/\psi_1 = \sin\theta/\cos\theta = \tan\theta = T/S$ and so $\theta = \arg(X^*\mu)$ as in case 1. Further,

$$\begin{aligned}\widehat{\beta} &= \sqrt{\psi_1^2 + \psi_2^2} = \frac{\sqrt{S^2+T^2}}{2P}, \\ &= \frac{\sqrt{(X_1^T\mu_1 + X_2^T\mu_2)^2 + (X_2^T\mu_1 - X_1^T\mu_2)^2}}{(X_1^TX_1 + X_2^TX_2)} = \frac{\sqrt{X^*\mu\mu^*X}}{X^*X},\end{aligned}$$

and so statements (III), (IV) and (V) are also true. □

2.3.4 Special case: $\Sigma = I_m \otimes \Sigma_k$

This special case will be considered in two ways. Firstly, expressions for the similarity transformation parameters will be derived by minimising $D^2_{CWP}(X, \mu; I_m \otimes \Sigma_k)$, then the results given in Sections 2.3.1 and 2.3.2 will be shown to simplify and give equivalent expressions under this constraint. Assume, without loss of generality, that X and μ are located such that $1_k^T\Sigma_k^{-1}X = 0 = 1_k^T\Sigma_k^{-1}\mu$. Then,

$$\begin{aligned}D^2_{CWP}(X, \mu; I_m \otimes \Sigma_k) &= \|\mu - \beta X\Gamma - 1_k\gamma^T\|^2_{I_m\otimes\Sigma_k}, \\ &= \mathrm{vec}(\mu - \beta X\Gamma - 1_k\gamma^T)^T(I_m \otimes \Sigma_k^{-1}) \\ &\quad \times \mathrm{vec}(\mu - \beta X\Gamma - 1_k\gamma^T), \\ &= \mathrm{tr}\left[(\mu - \beta X\Gamma - 1_k\gamma^T)^T\Sigma_k^{-1}(\mu - \beta X\Gamma - 1_k\gamma^T)\right], \\ &= \mathrm{tr}(\mu^T\Sigma_k^{-1}\mu) + \beta^2\mathrm{tr}(X^T\Sigma_k^{-1}X) \\ &\quad -2\beta\mathrm{tr}(\mu^T\Sigma_k^{-1}X\Gamma) + \mathrm{tr}(\gamma 1_k^T\Sigma_k^{-1}1_k\gamma^T) \\ &\quad -2\mathrm{tr}(\gamma 1_k^T\Sigma_k^{-1}\mu^T) + 2\beta\mathrm{tr}(\gamma 1_k^T\Sigma_k^{-1}X\Gamma).\end{aligned}$$

Clearly, the last two terms are zero, and the requirement of Σ to be positive definite implies $1_k^T\Sigma_k^{-1}1_k > 0$. Therefore, D^2_{CWP} is minimised if

$\widehat{\gamma} = 0$ and D^2_{CWP} simplifies to,

$$D^2_{CWP}(X, \mu; I_m \otimes \Sigma_k) = \mathrm{tr}(\mu^T \Sigma_k^{-1} \mu) + \beta^2 \mathrm{tr}(X^T \Sigma_k^{-1} X) - 2\beta \mathrm{tr}(\mu^T \Sigma_k^{-1} X \Gamma).$$

The minimising rotation is $\widehat{\Gamma} = UV^T$, where $\mu^T \Sigma_k^{-1} X = V \Lambda U^T$, by the same argument as for isotropic OPA. For the minimising scaling, note that the first derivative is,

$$\frac{dD^2_{CWP}}{d\beta} = 2\beta \mathrm{tr}(X^T \Sigma_k^{-1} X) - 2\mathrm{tr}(\mu^T \Sigma_k^{-1} X \Gamma).$$

The second derivative is positive because Σ_k^{-1} is positive definite, so the minimum of D^2_{CWP} is at $\widehat{\beta} = \mathrm{tr}(\mu^T \Sigma_k^{-1} X \Gamma)/\mathrm{tr}(X^T \Sigma_k^{-1} X)$. Note that if X and μ are replaced by $X_Q = QX$ and $\mu_Q = Q\mu$ respectively, where $\Sigma_k^{-1} = Q^T Q$ is the Cholesky decomposition, then the estimates of $\widehat{\Gamma}$ and $\widehat{\beta}$ for matching X to μ with $\Sigma = I_m \otimes \Sigma_k$ are equivalent to the estimates of $\widehat{\Gamma}$ and $\widehat{\beta}$ for matching X_Q to μ_Q with $\Sigma = I_{km}$, as claimed by Goodall (1991).

Result 2.3.6 *If Σ is of the form $I_m \otimes \Sigma_k$, where $\Sigma_k(k \times k)$ is a symmetric positive definite matrix, and assume without loss of generality that X and μ are located such that $1_k^T \Sigma_k^{-1} X = 0 = 1_k^T \Sigma_k^{-1} \mu$, then*

$$\widehat{\gamma} = 0, \qquad \widehat{\beta} = \frac{(S^2 + T^2)^{1/2}}{2P},$$
$$\cos\widehat{\theta} = \frac{-S}{(S^2 + T^2)^{1/2}}, \qquad \sin\widehat{\theta} = \frac{-T}{(S^2 + T^2)^{1/2}},$$

where,

$$\begin{aligned} P &= X_1^T \Sigma_k^{-1} X_1 + X_2^T \Sigma_k^{-1} X_2, \\ S &= -2(X_1^T \Sigma_k^{-1} \mu_1 + X_2^T \Sigma_k^{-1} \mu_2), \\ T &= 2(X_2^T \Sigma_k^{-1} \mu_1 + X_1^T \Sigma_k^{-1} \mu_2). \end{aligned}$$

Proof: If $\Sigma = I_m \otimes \Sigma_k$, then the similarity transformation estimates of

Result 2.3.4 can be simplified. For the translation,

$$\begin{aligned}\widehat{\gamma} &= [(I_m \otimes 1_k)^T (I_m \otimes \Sigma_k)^{-1} \\ &\quad \times (I_m \otimes 1_k)]^{-1} (I_m \otimes 1_k)^T (I_m \otimes \Sigma_k)^{-1} \mathrm{vec}(\mu - \beta X \Gamma), \\ &= [I_m \otimes (1_k^T \Sigma_k^{-1} 1_k)]^{-1} [I_m \otimes (1_k^T \Sigma_k^{-1})] \mathrm{vec}(\mu - \beta X \Gamma), \\ &= [I_m \otimes (1_k^T \Sigma_k^{-1} 1_k)^{-1} (1_k^T \Sigma_k^{-1})] \mathrm{vec}(\mu - \beta X \Gamma).\end{aligned}$$

Therefore, $\widehat{\gamma}^T = (1_k^T \Sigma_k^{-1} 1_k)^{-1} 1_k^T \Sigma_k^{-1} (\mu - \beta X \Gamma)$, which is zero given $1_k^T \Sigma_k^{-1} X = 0 = 1_k^T \Sigma_k^{-1} \mu$. Referring to the notation of Result 2.3.2, if $\Sigma = I_m \otimes \Sigma_k$ then $A_{11} = A_{22} = (1_k^T \Sigma_k^{-1} 1_k)^{-1} 1_k^T \Sigma_k^{-1}$ and $A_{12} = A_{21} = 0_k^T$. Then, from Equations (2.8), if X and μ are located such that $1_k^T \Sigma_k^{-1} X = 0 = 1_k^T \Sigma_k^{-1} \mu$, then $\alpha_i = \delta_i = \zeta_i = 0$, for $i = 1, 2$, and P, Q, R, S and T simplify to,

$$\begin{aligned}P = Q &= X_1^T \Sigma_k^{-1} X_1 + X_2^T \Sigma_k^{-1} X_2, \\ R &= -2(X_1^T \Sigma_k^{-1} X_2 - X_2^T \Sigma_k^{-1} X_1) = 0, \\ S &= -2(X_1^T \Sigma_k^{-1} \mu_1 + X_2^T \Sigma_k^{-1} \mu_2), \\ T &= 2(X_2^T \Sigma_k^{-1} \mu_1 - X_1^T \Sigma_k^{-1} \mu_2).\end{aligned}$$

The minimising rotation and scaling can then be obtained by following the arguments of CASE 1 and CASE 2 in the proof of Result 2.3.5. Further, if X and μ are replaced by $X_Q = QX$ and $\mu_Q = Q\mu$ respectively, where $\Sigma_k^{-1} = Q^T Q$ is the Cholesky decomposition, then the expressions for P, S and T are equivalent to those given in Result 2.3.5 for the isotropic case. □

2.4 Covariance weighted GPA

2.4.1 Definition and algorithm

In Section 2.3 covariance weighted ordinary Procrustes analysis was defined as obtaining the Procrustes estimators to minimise the Mahalanobis norm $D^2_{CWP}(X, \mu; \Sigma) = \|\mu - X^P\|^2_\Sigma$, where one configuration, X, is translated, rotated and possibly scaled with respect to a reference configuration, μ. In this section CW OPA is extended to define covariance weighted generalised Procrustes analysis for a more general data set with $n \geq 2$ configurations, $X_1, X_2, ..., X_n$. This allows inferences to be made regarding the sample mean shape.

Definition 2.4.1 *The method of full covariance weighted generalised Procrustes analysis (full CW GPA) involves the least squares matching of n configurations relative to each other using similarity transformations, and the procedure is appropriate under the model,*

$$X_i = \beta_i(\mu + E_i)\Gamma_i + 1_k\gamma_i^T,$$

where E_i are zero mean $k \times m$ independent random error matrices, μ is the $k \times m$ matrix of the mean configuration and γ_i, Γ_i and β_i are nuisance parameters for translation, rotation and scale. A quantity proportional to the sum of squared Mahalanobis norms of pairwise differences,

$$G_{CWP}(X_1, ..., X_n; \Sigma) = \frac{1}{n}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \left\|(\beta_i X_i\Gamma_i + 1_k\gamma_i^T) - (\beta_j X_j\Gamma_j + 1_k\gamma_j^T)\right\|^2_\Sigma, \qquad (2.14)$$

is minimised subject to the constraint on the size of the centred average shape,

$$\|\bar{X}\|^2 = 1,$$

where $\Gamma_i \in SO(m)$, $\beta_i > 0$ *and*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n}(\beta_i X_i \Gamma_i + 1_k \gamma_i^T).$$

Partial covariance weighted generalised Procrustes analysis can be similarly defined using rigid-body transformations and without a constraint on the size of the mean shape. Minimising the sum of squared Mahalanobis norms of pairwise differences is equivalent to minimising the distance between each configuration and the mean of the configurations,

$$\begin{aligned}
&G_{CWP}(X_1, ..., X_n; \Sigma) \\
&= \inf_{\beta_i, \Gamma_i, \gamma_i} \frac{1}{n}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \left\|(\beta_i X_i \Gamma_i + 1_k \gamma_i^T) - (\beta_j X_j \Gamma_j + 1_k \gamma_j^T)\right\|_\Sigma^2, \\
&= \inf_{\beta_i, \Gamma_i, \gamma_i} \sum_{i=1}^{n} \left\|(\beta_i X_i \Gamma_i + 1_k \gamma_i^T) - \frac{1}{n}\sum_{j=1}^{n}(\beta_j X_j \Gamma_j + 1_k \gamma_j^T)\right\|_\Sigma^2, \\
&= \inf_{\beta_i, \Gamma_i, \gamma_i} \sum_{i=1}^{n} \left\|(\beta_i X_i \Gamma_i + 1_k \gamma_i^T) - \bar{X}\right\|_\Sigma^2.
\end{aligned}$$

An alternative is to minimise the distance from the ith shape to the mean of the rest of the sample,

$$\begin{aligned}
&G_{CWP}(X_1, ..., X_n; \Sigma) \\
&= \inf_{\beta_i, \Gamma_i, \gamma_i} \sum_{i=1}^{n} \left\|(\beta_i X_i \Gamma_i + 1_k \gamma_i^T) - \frac{1}{n}\sum_{j=1}^{n}(\beta_j X_j \Gamma_j + 1_k \gamma_j^T)\right\|_\Sigma^2, \\
&= \inf_{\beta_i, \Gamma_i, \gamma_i} \sum_{i=1}^{n} \left\|\frac{n-1}{n}(\beta_i X_i \Gamma_i + 1_k \gamma_i^T) - \frac{1}{n}\sum_{j=1, j\neq i}^{n}(\beta_j X_j \Gamma_j + 1_k \gamma_j^T)\right\|_\Sigma^2, \\
&= \inf_{\beta_i, \Gamma_i, \gamma_i} \sum_{i=1}^{n} \frac{(n-1)^2}{n^2} \left\|(\beta_i X_i \Gamma_i + 1_k \gamma_i^T) - \bar{X}_{-i}\right\|_\Sigma^2,
\end{aligned}$$

where,

$$\bar{X}_{-i} = \frac{1}{n-1} \sum_{j=1, j\neq i}^{n} (\beta_j X_j \Gamma_j + 1_k \gamma_j^T). \tag{2.15}$$

Once the transformation parameters that minimise Equation (2.14) have been identified, it is possible to make the following definitions.

Definition 2.4.2 *The full covariance weighted Procrustes fit of the each X_i is given by,*

$$X_i^P = \widehat{\beta}_i X_i \widehat{\Gamma}_i + 1_k \widehat{\gamma}_i^T, \tag{2.16}$$

where $\widehat{\gamma}_i$, $\widehat{\Gamma}_i$ and $\widehat{\beta}_i$ are the minimising parameters of Equation (2.14), for $i = 1, ..., n$.

Definition 2.4.3 *The full covariance weighted Procrustes estimate of the mean shape is given by $\widehat{\mu}$ where*

$$\widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i^P, \tag{2.17}$$

where X_i^P is the full covariance weighted Procrustes fit of X_i, $i = 1, ..., n$.

The partial covariance weighted Procrustes fit and the partial covariance weighted Procrustes estimate of the mean shape can be similarly defined by omitting the β parameter. The similarity transformation parameters which minimise Equation (2.14) can be obtained by the following algorithm.

Algorithm 2.4.1 ***Covariance weighted GPA algorithm***

1. Initial registration: *Given symmetric positive definite matrix, Σ, and n shapes, $X_1, X_2, ..., X_n$, calculate $\bar{X}$ as the mean shape resulting from the isotropic GPA algorithm.*

2. Centre, scale and orientate mean shape: *Centre and scale*

$\bar{X}$ such that $S(\bar{X}) = 1$, and apply the same translation and scaling to each X_i. Further, apply an identical rotation to $\bar{X}, X_1, ..., X_n$ to minimise $G_{CWP}(X_1, ..., X_n; \Sigma)$.

3. CW OPA: *For $i = 1, ..., n$ register X_i to $\bar{X}_{-i}$ using covariance weighted OPA to minimise $D^2_{CWP}(X_i, \bar{X}_{-i}; \Sigma)$.*

4. Repetition. *Repeat steps (2) and (3) until $G_{CWP}(X_1, ..., X_n; \Sigma)$ cannot be reduced further.*

The algorithm is guaranteed to converge because $G_{CWP}(X_1, ..., X_n; \Sigma)$ is non-increasing at each step and bounded below. Step 2 of the algorithm translates and rotates all the shapes relative to the axes to speed up the rate of convergence. The location of the mean shape is arbitrary, but the orientation of the mean shape is not. If Σ has been defined with respect to some user-specified reference configuration, μ_0, then minimising $D_{OPA}(\bar{X}, \mu_0)$ or $D^2_{CWP}(X_i, \mu_0; \Sigma)$ instead of $G_{CWP}(X_1, ..., X_n; \Sigma)$ at step 2 keeps the shapes in that frame of reference. Alternatively, if Σ has been chosen with respect to the axes, then the rotation in step 2 can be calculated numerically for $m \geq 3$, or by Result 2.4.1 for $m = 2$.

Result 2.4.1 *For $m = 2$, let the rotation matrix, $\Gamma = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$, then,*

$$G_{CWP}(X_1, ..., X_n; \Sigma) = \inf_{\beta_i, \Gamma_i, \gamma_i, \Gamma} \sum_{i=1}^{n} \|(\beta_i X_i \Gamma_i + 1_k \gamma_i^T - \bar{X})\Gamma\|^2_\Sigma,$$

is minimised when θ is a solution of

$$\tan(2\theta) = \frac{2r}{p - q},$$

where,

$$p = \sum_{i=1}^{n} \mathrm{vec}(R_i)^T \Sigma^{-1} \mathrm{vec}(R_i), \qquad q = \sum_{i=1}^{n} \mathrm{vec}(R_i^\perp)^T \Sigma^{-1} \mathrm{vec}(R_i^\perp),$$

$$
\begin{aligned}
r &= \sum_{i=1}^{n} \mathrm{vec}(R_i)^T \Sigma^{-1} \mathrm{vec}(R_i^{\perp}), \\
R_i &= (\beta_i X_i \Gamma_i + 1_k \gamma_i^T - \bar{X}) = \begin{bmatrix} R_{i1} & R_{i2} \end{bmatrix}, \qquad R_i^{\perp} = \begin{bmatrix} -R_{i2} & R_{i1} \end{bmatrix}.
\end{aligned}
$$

Proof:

$$
\begin{aligned}
G_{CWP}(X_1, ..., X_n; \Sigma) &= \sum_{i=1}^{n} \mathrm{vec}(R_i\Gamma)^T \Sigma^{-1} \mathrm{vec}(R_i\Gamma), \\
&= \sum_{i=1}^{n} \begin{bmatrix} R_{i1}\cos\theta - R_{i2}\sin\theta \\ R_{i2}\cos\theta + R_{i1}\sin\theta \end{bmatrix}^T \Sigma^{-1} \begin{bmatrix} R_{i1}\cos\theta - R_{i2}\sin\theta \\ R_{i2}\cos\theta + R_{i1}\sin\theta \end{bmatrix}, \\
&= p\cos^2\theta + q\sin^2\theta + 2r\cos\theta\sin\theta, \\
\frac{dG_{CWP}}{d\theta} &= (q-p)\sin 2\theta + 2r\cos 2\theta.
\end{aligned}
$$

Therefore, the minimum of G_{CWP} is when θ is a solution of this last equation. □

Figure 2.6 shows the partial generalised covariance weighted Procrustes registration of 30 second thoracic mouse vertebrae for three different weighting matrices, Σ. Two of the vertebrae were first considered in Section 2.3.1. The three weighting matrices used are given in Equation (2.10). The first example is the same as isotropic GPA, giving the same weighting to all coordinates and showing approximately the same variability at each landmark. The second example weights the first two landmarks more heavily than the others, and this is reflected in these landmarks forming lines pointing towards each other, and more variability being introduced elsewhere, particularly at landmark 4. The third example weights the y direction more heavily than the x direction. This forces large rotations to ensure landmarks 1 and 2 have similar y values across all configurations, and gives large variability to the other landmarks.

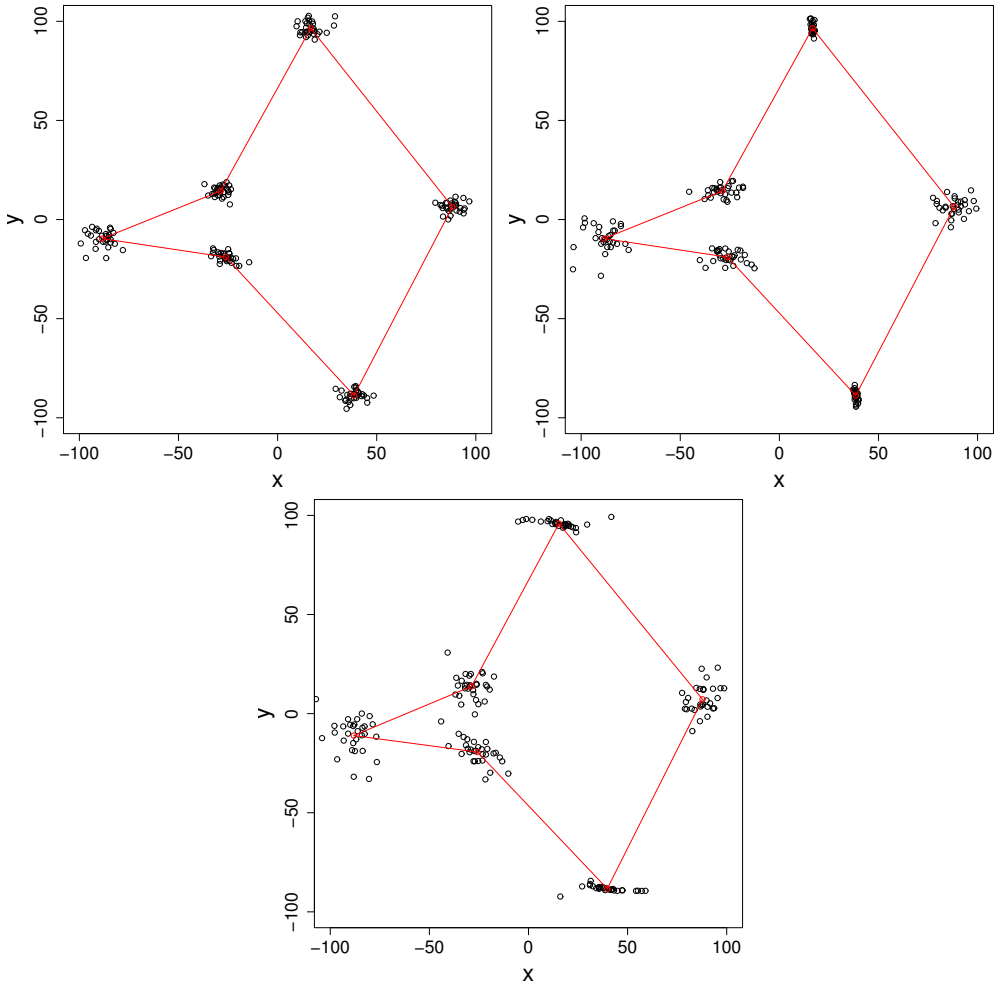

Figure 2.6: The partial generalised covariance weighted Procrustes registration of 30 mice vertebra, using Σ_1, Σ_2 and Σ_3. The mean shape is shown with a solid line.

2.4.2 Relating CW GPA to the multivariate normal distribution

If the data set $X_1, X_2, ..., X_n$ comes from a multivariate normal distribution with unknown mean shape, μ, and known covariance matrix, Σ, that is $\text{vec}(X_i) \sim N_{km}(\text{vec}(\mu), \Sigma)$ then the CW GPA algorithm can be used to maximise the likelihood of the model, and provide the maximum likelihood estimate of the mean shape.

Result 2.4.2 *Maximising the likelihood of the multivariate normal model,* $\text{vec}(X_i) \sim N_{km}(\text{vec}(\mu), \Sigma)$ *for a given* Σ *is equivalent to minimising* $G_{CWP}(X_1, ..., X_n; \Sigma)$*, and the maximum likelihood estimate of the mean shape is,*

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i^P, \tag{2.18}$$

where X_i^P *is the full covariance weighted Procrustes fit of* X_i*.*

Proof: The log-likelihood of the multivariate normal model, $\text{vec}(X_i) \sim N_{km}(\text{vec}(\mu), \Sigma)$, where X_i are shapes invariant under Euclidean similarity transformations, is,

$$\begin{aligned}\log L(X_1, ..., X_n; \mu, \Sigma) = &-\frac{n}{2}\log|2\pi\Sigma| \\ &-\frac{1}{2}\sum_{i=1}^{n}\text{vec}(\beta_i X_i\Gamma_i + 1_k\gamma_i^T - \mu)^T\Sigma^{-1}\text{vec}(\beta_i X_i\Gamma_i + 1_k\gamma_i^T - \mu).\end{aligned}$$

Therefore, the maximum likelihood estimate of the mean shape is the solution of,

$$\frac{d\log L}{d\mu} = \sum_{i=1}^{n}\Sigma^{-1}\text{vec}(\beta_i X_i\Gamma_i + 1_k\gamma_i^T) - n\Sigma^{-1}\mu = 0.$$

Hence, $\widehat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n}(\beta_i X_i \Gamma_i + 1_k \gamma_i^T)$ and,

$$\begin{aligned} \log L &= -\frac{n}{2}\log|2\pi\Sigma| - \frac{1}{2}\inf_{\beta_i,\Gamma_i,\gamma_i}\sum_{i=1}^{n}\|\beta_i X_i\Gamma_i + 1_k\gamma_i^T - \bar{X}\|_\Sigma^2, \\ &= -\frac{n}{2}\log|2\pi\Sigma| - \frac{1}{2}G_{CWP}(X_1,...,X_n;\Sigma). \end{aligned}$$

Therefore, minimising G_{CWP} is equivalent to maximising $L(X_1,...,X_n;\mu,\Sigma)$. □

If Σ is diagonal, with variance σ_1^2 for a subset of the landmarks, and variance σ_2^2 for the remainder, then as $\sigma_2 \to \infty$, the distribution could be approximated using the subset matching results shown earlier, with the mean shape forming the template.

2.5 Discussion

In this chapter, we have presented some distributional results for shape differences when isotropic Procrustes methods are applied to a subset of the landmarks, and extended Procrustes methods to define mean shapes and fitted shapes in the context of a weighted superimposition matrix, Σ. Importantly, the latter has been presented for both partial and full registrations, and the results shown to agree with isotropic Procrustes methods in the case $\Sigma = I$. Procrustes methods have previously been criticised by Lele (1993) for not producing consistent estimates of mean shapes. Kent and Mardia (1997) show that the full isotropic Procrustes estimate of the mean shape is consistent in the presence of isotropic errors, i.e. $\widehat{\mu} \to \mu$ as $n \to \infty$, for the case $m = 2$. The partial estimate is also consistent for shape up to a scale factor if the errors are assumed to be Gaussian as well as isotropic. Future work will need to investigate whether or not covariance weighted GPA produces a consistent estimator for μ under Gaussian errors with covariance matrix, Σ.

However, this must be seen within the context of shape theory and other

estimators of the mean shape. Lele's (1993) own estimate of the mean shape, using Euclidean distance matrix analysis (EDMA), is inconsistent under certain circumstances (Goodall, 1995). Mardia and Dryden (1994) evaluate the bias of several estimators through a simulation study and find most estimators perform well when the data contains low amounts of isotropic shape variability. They also show that several well-known estimators, such as Bookstein's (1986) estimate of the mean shape, are biased for larger amounts of variability. Typically, shape analysis is applied to data sets where the natural variability in the data is much greater than the bias in the estimators (Kent and Mardia, 1997), so this is not too concerning. For example, Kent (1994) demonstrated that if the standard deviation of measurements, σ, under a Gaussian model is small compared to the size of the shape, then the bias of the isotropic Procrustes estimator is of order σ^2.

Fortunately, Procrustes/maximum-likelihood estimates continue to perform well in the presence of large isotropic variability, and in Chapter 3 we will consider a simulation study, comparing the covariance weighted and isotropic Procrustes estimators in the presence of non-isotropic variability. Dryden and Mardia (1991) also consider maximum likelihood estimates of shape variability, with particularly stringent constraints on the form of the covariance matrix, through the use of numerical routines. In Chapter 3, we start by developing an algorithm for the maximum likelihood estimate of a more general form.

Chapter 3

Estimating shape variability

3.1 Introduction

In Chapter 2 it was shown that isotropic Procrustes analysis could be extended to produce Procrustes estimates of the shape transformation parameters based on a symmetric weighting matrix, Σ, of dimension $km \times km$, where k is the number of landmarks and m is the number of dimensions. The matrix is constrained to be positive definite ($\Sigma > 0$) so that the transformation parameters are defined. A matrix of this form is also a covariance matrix, so it would be natural to use the variability of the data as the inverse of the weighting matrix. In this chapter we develop methods for estimating shape variability, subject to Σ being positive definite, in conjunction with the mean shape, μ, and the transformation parameters, β_i, Γ_i and γ_i. Clearly, Σ is dependent on the other parameters and vice-versa, so we develop iterative procedures using conditional maximum likelihood (ML) estimation in Section 3.2 and Markov Chain Monte Carlo (MCMC) simulation in Section 3.4. Both methods are dependent on *a priori* information, to enforce the constraint in the ML algorithm or to specify prior distributions in the MCMC algorithm, and we assess both methods through a simulation study. The MCMC method is extended to include the possibility of missing

data in Section 3.6.

3.2 Maximum likelihood estimation

In this section we derive a maximum likelihood estimator (MLE), under a multivariate Gaussian model, for the shape covariance matrix, Σ, given the matrix is positive definite. Firstly, the unconstrained maximum likelihood estimates of Σ, for various structures are given. Estimation of transformation parameters is dependent on Σ being positive definite, so Σ is estimated with this restriction, and possible constraints on the elements of Σ are considered in Section 3.2.2. Lastly, an algorithm is given for estimating the transformation parameters, mean shape and covariance matrix, subject to the constraints.

3.2.1 Covariance matrix parameterisation

Let, $X_1, X_2, ..., X_n$, be $k \times m$ configurations at arbitrary locations in Euclidean space, such that the shapes have a multivariate normal distribution $\text{vec}(X_i) \sim N_{km}(\text{vec}(\mu), \Sigma)$, where μ and Σ are both unknown. The log-likelihood of the model, where the shapes are invariant under Euclidean similarity transformations is,

$$\log L(X_1, ..., X_n; \mu, \Sigma) = -\frac{n}{2}\log|2\pi\Sigma| \\ -\frac{1}{2}\sum_{i=1}^{n}\text{vec}(\beta_i X_i\Gamma_i + 1_k\gamma_i^T - \mu)^T\Sigma^{-1}\text{vec}(\beta_i X_i\Gamma_i + 1_k\gamma_i^T - \mu).$$

We will maximise the likelihood of the model by iteratively estimating the similarity transformation parameters, μ and Σ in turn, each conditional on the current estimates of the remaining parameters. Therefore, we can use the estimators of the transformation parameters given in Chapter 2 and the

conditional maximum likelihood estimate of the mean shape is,

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n}(\beta_i X_i \Gamma_i + 1_k \gamma_i^T).$$

The application may require the covariance matrix to follow a particular parameterisation. Restricting the form of the covariance matrix reduces the number of parameters to be estimated, and hence reduces the number of constraints necessary to force $\Sigma > 0$. Let X_i^P be the ith configuration following a similarity transformation using the current maximum likelihood estimates, then the maximum likelihood estimate of Σ is given below for a variety of parameterisations. Details can be found in Mardia *et al.* (1979) and Dutilleul (1999).

1. Unconstrained case. The MLE of Σ for the model $\text{vec}(X_i) \sim N_{km}(\text{vec}(\mu), \Sigma)$ is,

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\text{vec}(X_i^P - \bar{X})\text{vec}(X_i^P - \bar{X})^T.$$

2. Factored case. The MLEs of Σ_m $(m \times m)$ and Σ_k $(k \times k)$ for the model $\text{vec}(X_i) \sim N_{km}(\text{vec}(\mu), \Sigma = \Sigma_m \otimes \Sigma_k)$ are,

$$\begin{aligned}
\widehat{\Sigma}_m &= \frac{1}{nk}\sum_{i=1}^{n}(X_i^P - \bar{X})^T \widehat{\Sigma}_k (X_i^P - \bar{X}), \\
\widehat{\Sigma}_k &= \frac{1}{nm}\sum_{i=1}^{n}(X_i^P - \bar{X}) \widehat{\Sigma}_m (X_i^P - \bar{X})^T.
\end{aligned}$$

3. Isotropic factored case. The MLE of Σ_k, $(k \times k)$ for the model $\text{vec}(X_i) \sim N_{km}(\text{vec}(\mu), \Sigma = I_m \otimes \Sigma_k)$ is,

$$\widehat{\Sigma}_k = \frac{1}{nm}\sum_{i=1}^{n}(X_i^P - \bar{X})(X_i^P - \bar{X})^T.$$

4. Block diagonal case. The MLEs of Σ_{kj}, $(k \times k)$, for $j = 1, ..., m$, for the model $\text{vec}(X_i) \sim N_{km}(\text{vec}(\mu), \Sigma = \text{diag}(\Sigma_{k1}, ..., \Sigma_{km}))$ are,

$$\widehat{\Sigma}_{kj} = \frac{1}{n}\sum_{i=1}^{n}(X_{ij}^P - \bar{X}_j)(X_{ij}^P - \bar{X}_j)^T,$$

where the subscript j denotes the jth column.

5. Scaled case. The MLE of Σ for the model $\text{vec}(X_i) \sim N_{km}(\text{vec}(\mu), \Sigma = p\Sigma_0)$, where p is a scalar and Σ_0 is known, is,

$$\widehat{\Sigma} = \frac{1}{km}\text{tr}\left(\frac{\Sigma_0^{-1}}{n}\sum_{i=1}^{n}\text{vec}(X_i^P - \bar{X})\text{vec}(X_i^P - \bar{X})^T\right)\Sigma_0.$$

This factored parameterisation of case 2 is often applicable in shape studies (for example, Goodall, 1991) as it separates variability between dimensions from variability between landmarks. This makes interpretation easier, greatly reduces the number of parameters to estimate and makes matrix inversion computationally quicker. This model is particularly suited to datasets where variability between dimensions is homogeneous at each landmark, perhaps due to the method of recording the coordinates. For example, a magnetic resonance scanner may have less spatial resolution between image slices than within an image slice. Mitchell *et al.* (2003) propose a likelihood ratio test for determining if a covariance matrix can be factored.

The solutions for case 2 have to be solved iteratively and can only be solved up to a multiplicative constant given $\Sigma_m \otimes \Sigma_k = c\Sigma_m \otimes (1/c)\Sigma_k$. As a result, some applications use the simplification of case 3 because this estimator has the desirable properties of being unique and having a non-iterative solution. The model assumes equal, independent variability between the dimensions, which could be considered valid if the coordinates of the overlaid shapes form a spherical cluster surrounding each coordinate of the mean shape. This model is also appealing if the estimation of only Σ_k is of primary interest.

Case 4 also estimates the variability with the constraint of preserving independence between dimensions, but allows for a different correlation structure between landmarks in each dimension. This is a generalisation of case 3 but involves the estimation of many more parameters than the previous two cases. Consequently, many studies would not benefit from an application of this model. Case 5 estimates $\widehat{\Sigma}$ as proportional to a given matrix, Σ_0. The specification of Σ_0 could result from *a priori* knowledge. Other parameterisations of the covariance matrix exist in the literature due to the specific nature of the data being considered. For example, Hurn *et al.* (2001) only consider the outline of shapes and, hence, apply a model similar to case 3 but constrain Σ_k to be a first order cyclic Markov structure. Despite the reduction in parameters that alternative parameterisations offer, we consider the most general, unconstrained, case in detail. The estimation of shape variability for alternative models should be simpler, with less constraints necessary to ensure the estimate is positive definite.

3.2.2 Estimating the covariance matrix

Let the spectral decomposition of the unconstrained maximum likelihood estimator be, $\Sigma = UDU^T$, where D is a diagonal matrix of eigenvalues, $d_1 \geq d_2 \geq ... \geq d_p$. Note that $p = km$ in our application to shape analysis. It has been widely noted (e.g. Dey and Srinivasan, 1985) that while this estimator is unbiased, its eigenvalues are more widely dispersed than the eigenvalues of the population covariance matrix. Consequently, a number of alternative estimators have been proposed. Let n be the number of observations, $X_1, ..., X_n$, then some alternatives artificially adjust the eigenvalues by replacing d_i with λ_i where,

$$\lambda_i = \frac{nd_i}{n+p+1-2i}, \qquad \text{or} \qquad \lambda_i = \frac{nd_i}{n-p+1+2d_i\sum_{j\neq i} 1/(d_i-d_j)},$$

as proposed by Dey and Srinivasan (1985) and Stein (1977), respectively. Another method for improving the maximum likelihood estimator is to shrink the estimator, Σ, towards the identity matrix, I, using a linear combination of the two. Haff (1980) and Ledoit and Wolf (2004) proposed,

$$\begin{aligned}\widehat{\Sigma} &= \frac{(np-2n-2)\bar{d}}{pn^2}I + \frac{n}{n+1}\Sigma, \\ \widehat{\Sigma} &= \frac{b^2\bar{d}}{d^2}I + \frac{a^2}{d^2}\Sigma,\end{aligned}$$

respectively, where $\bar{d}$ is the arithmetic mean of the diagonal elements of Σ, $d^2 = \|\Sigma - \bar{d}I\|^2/p$, $b^2 = \min(\bar{b}^2, d^2)$, $a^2 = d^2 - b^2$ and,

$$\bar{b}^2 = \frac{1}{n^2}\sum_{i=1}^{n} \|\text{vec}(X_i)\text{vec}(X_i)^T - \Sigma\|^2/p.$$

Unfortunately, the proposed methods for adjusting the eigenvalues do not remove singularities in the eigenvalues and the methods involving shrinking will, on repeated application in a conditional maximum likelihood algorithm, produce estimates of shape variability similar to those based on isotropic Procrustes. Therefore, we develop a constrained maximum likelihood estimator with $\widehat{\Sigma}$ positive definite, so that estimates of the transformation parameters of translation, rotation and scaling are constrained.

In general, we would require r columns of Σ to be defined or constrained, where r is the number of transformation parameters to be estimated. For example, $r = m(m-1)/2 + m + 1$, for the case of full Procrustes analysis. Given Σ is symmetric, this equates to $c = r(2km - r + 1)/2$ individual entries of Σ. In general, this means constraints can be written in the form $A\text{vech}(\Sigma) = b$ where the vech operator lists the $s = km(km+1)/2$ distinct elements of Σ and A is a $c \times s$ matrix and b is a vector of length c. If Σ is constrained by one of the parameterisations of Section 3.2.1 then less constraints are necessary.

Suppose $X = \beta\mu\Gamma + 1_k\gamma^T$, then after full Procrustes registration of the

shape, variability due to translation, rotation and scaling has been removed. Therefore, a natural method of constraining Σ in shape analysis is to specify the amount of variability in the r directions. In this section we present such a method for the typical cases of $m = 3$ and $m = 2$, although the method can easily be extrapolated to higher dimensions.

Let $m = 3$, then $r = 7$, and assuming the angles of rotations, θ_x, θ_y and θ_z are small, cos() and sin() can be approximated using the Taylor expansions to the first order, so,

$$\begin{aligned}
\Gamma &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_x & \sin\theta_x \\ 0 & -\sin\theta_x & \cos\theta_x \end{bmatrix} \begin{bmatrix} \cos\theta_y & 0 & \sin\theta_y \\ 0 & 1 & 0 \\ -\sin\theta_y & 0 & \cos\theta_y \end{bmatrix} \begin{bmatrix} \cos\theta_z & \sin\theta_z & 0 \\ -\sin\theta_z & \cos\theta_z & 0 \\ 0 & 0 & 1 \end{bmatrix}, \\
&= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & \theta_x \\ 0 & -\theta_x & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & \theta_y \\ 0 & 1 & 0 \\ -\theta_y & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & \theta_z & 0 \\ -\theta_z & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} + O(\theta_x^2, \theta_y^2, \theta_z^2).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathrm{vec}(X) &= \mathrm{vec}\left(\begin{bmatrix} 1_k\gamma_x & 1_k\gamma_y & 1_k\gamma_z \end{bmatrix} \right) + \beta \mathrm{vec}\left(\begin{bmatrix} \mu_x & \mu_y & \mu_z \end{bmatrix} \Gamma \right), \\
&= \begin{bmatrix} 1_k\gamma_x \\ 1_k\gamma_y \\ 1_k\gamma_z \end{bmatrix} + \beta \begin{bmatrix} \mu_x - \mu_y\theta_z - \mu_z\theta_y + O(\theta^2) \\ \mu_x\theta_z + \mu_y - \mu_z\theta_x + O(\theta^2) \\ \mu_x\theta_y + \mu_y\theta_x + \mu_z + O(\theta^2) \end{bmatrix}, \\
&= \begin{bmatrix} 1_k & 0_k & 0_k \\ 0_k & 1_k & 0_k \\ 0_k & 0_k & 1_k \end{bmatrix} \begin{bmatrix} \gamma_x \\ \gamma_y \\ \gamma_z \end{bmatrix} + \beta \begin{bmatrix} \mu_x \\ \mu_y \\ \mu_z \end{bmatrix} \\
&\quad + \beta \begin{bmatrix} 0_k & -\mu_z & -\mu_y \\ -\mu_z & 0_k & \mu_x \\ \mu_y & \mu_x & 0_k \end{bmatrix} \begin{bmatrix} \theta_x \\ \theta_y \\ \theta_z \end{bmatrix} + 0(\theta^2), \\
&= \gamma_x v_1 + \gamma_y v_2 + \gamma_z v_3 + \beta v_4 + \beta\theta_x v_5 + \beta\theta_y v_6 + \beta\theta_z v_7 + O(\theta^2), \quad (3.1)
\end{aligned}$$

where,

$$v_1 = \begin{bmatrix} 1_k \\ 0_k \\ 0_k \end{bmatrix}, \quad v_2 = \begin{bmatrix} 0_k \\ 1_k \\ 0_k \end{bmatrix}, \quad v_3 = \begin{bmatrix} 0_k \\ 0_k \\ 1_k \end{bmatrix}, \quad v_4 = \begin{bmatrix} \mu_x \\ \mu_y \\ \mu_z \end{bmatrix},$$
$$v_5 = \begin{bmatrix} 0_k \\ -\mu_z \\ \mu_y \end{bmatrix}, \quad v_6 = \begin{bmatrix} -\mu_z \\ 0_k \\ \mu_x \end{bmatrix}, \quad v_7 = \begin{bmatrix} -\mu_y \\ \mu_x \\ 0_k \end{bmatrix}.$$

Therefore, there is approximately no variability in the direction of v_j for $j = 1, ..., r$ following isotropic Procrustes registration. These vectors are linearly independent and, using the Gram-Schmidt process, can be transformed into r orthonormal vectors, $u_j(\mu)$, which is simplified if the mean shape is centred,

$$u_1 = \frac{v_1}{\|v_1\|}, \quad u_2 = \frac{v_2}{\|v_2\|}, \quad u_3 = \frac{v_3}{\|v_3\|}, \quad u_4 = \frac{v_4}{\|v_4\|}, \quad u_5 = \frac{v_5}{\|v_5\|},$$
$$u_6 = \frac{v_6 - (v_6^T u_5)u_5}{\|v_6 - (v_6^T u_5)u_5\|}, \quad u_7 = \frac{v_7 - (v_7^T u_5)u_5 - (v_7^T u_6)u_6}{\|v_7 - (v_7^T u_5)u_5 - (v_7^T u_6)u_6\|}.$$

Letting $u_j(\mu)$ form r eigenvectors of Σ, and specifying a strictly positive amount of variability, σ_j^2, for the direction $u_j(\mu)$, $j = 1, ..., r$, will be sufficient to ensure Σ is positive definite, even if σ_j^2 is near zero. Of course, if partial Procrustes registration is used then it is unnecessary to specify σ_4^2.

For $m = 2$, then $r = 4$ in the case of full Procrustes analysis, and Equation (3.1) simplifies to $\text{vec}(X) = \gamma_x u_1 + \gamma_y u_2 + \beta u_3 + \beta\theta u_4 + O(\theta^2)$ where,

$$u_1 = \frac{1}{\sqrt{k}} \begin{bmatrix} 1_k \\ 0_k \end{bmatrix}, \qquad u_2 = \frac{1}{\sqrt{k}} \begin{bmatrix} 0_k \\ 1_k \end{bmatrix},$$
$$u_3 = \frac{1}{\|\mu\|} \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \qquad u_4 = \frac{1}{\|\mu\|} \begin{bmatrix} -\mu_y \\ \mu_x \end{bmatrix}.$$

To parameterise a maximum likelihood estimator of Σ with these constraints, let Σ^* be the covariance matrix $\widehat{\Sigma}$, but with variability removed in the directions on which constraints have been imposed,

$$\Sigma^* = U\widehat{\Sigma}U^T, \tag{3.2}$$

where U is the projection matrix, $U = I_{km} - \sum_{j=1}^{r} u_j(\mu)u_j(\mu)^T$. Combining Σ^* with our constrained eigenvalues and eigenvectors will produce a positive definite covariance matrix,

$$\Sigma = \Sigma^* + \sum_{j=1}^{r} \sigma_j^2 u_j(\mu)u_j(\mu)^T.$$

3.2.3 Estimating the transformation parameters

In Section 3.2.2 we showed how constraints could be imposed on $\widehat{\Sigma}$ to ensure it was positive definite. If we parameterise Σ with eigenvectors aligned to the axes, as the constraints imply, then the estimate of the translation is simplified by Result 3.2.1 that follows. The rest of the transformation parameters, the mean shape and covariance matrix can then be estimated with the use of an algorithm.

Result 3.2.1 *Let* $\Sigma = \sum_{j=1}^{r} \sigma_j^2 u_j(\mu)u_j(\mu)^T + \sum_{j=1}^{km-r} \lambda_j \nu_j \nu_j^T$ *where* u_j *are orthonormal vectors on which variability has been constrained to be* σ_j^2*, and* λ_j *and* ν_j *are the remaining eigenvalues and eigenvectors of* Σ *orthogonal to* $u_j(\mu)$*. Then, assuming* m *of the vectors* $u_j(\mu)$ *are the columns of* $(I_m \otimes 1_k)/\sqrt{k}$*,* X_i *are centred and* $\mathrm{vec}(X_i) \sim N_{km}(\mathrm{vec}(\mu), \Sigma)$*, the likelihood is maximised when the translation parameter,* γ*, is zero.*

Proof: The likelihood for the model is,

$$L(X_1, ..., X_n; \mu, \Sigma) = |2\pi\Sigma|^{-\frac{n}{2}}$$
$$\times \exp\left(-\frac{1}{2}\sum_{i=1}^{n} \mathrm{vec}(\beta_i X_i \Gamma_i - 1_k\gamma_i^T - \mu)^T \Sigma^{-1} \mathrm{vec}(\beta_i X_i \Gamma_i - 1_k\gamma_i^T - \mu)\right).$$

Let the m columns of $(I_m \otimes 1_k)$ be 1_j, γ_{ij} be the jth element of the translation vector for shape X_i, for $j = 1, ..., m$, and $v_i = \text{vec}(\beta_i X_i \Gamma_i - \mu)$ then the log-likelihood simplifies as follows,

$$\begin{aligned}
\log L &= -\frac{n}{2}\log|2\pi\Sigma| - \frac{1}{2}\sum_{i=1}^{n}\left(v_i - \sum_{j=1}^{m}\gamma_{ij}1_j\right)^T \Sigma^{-1}\left(v_i - \sum_{j=1}^{m}\gamma_{ij}1_j\right), \\
&= -\frac{n}{2}\log|2\pi\Sigma| - \frac{1}{2}\sum_{i=1}^{n}\left(v_i^T\Sigma^{-1}v_i\right) \\
&\quad -\frac{1}{2}\sum_{i=1}^{n}\left(-2\left(\sum_{j=1}^{m}\gamma_{ij}1_j^T\right)\Sigma^{-1}v_i + \left(\sum_{j=1}^{m}\gamma_{ij}1_j^T\right)\Sigma^{-1}\left(\sum_{j=1}^{m}1_j\gamma_{ij}\right)\right).
\end{aligned}$$

Now, $1_j^T\Sigma^{-1} = \frac{k}{\sigma_j^2}1_j^T 1_j 1_j^T$ as all the eigenvectors of Σ are orthogonal to 1_j except the one proportional to 1_j. Therefore,

$$\begin{aligned}
\log L &= -\frac{n}{2}\log|2\pi\Sigma| - \frac{1}{2}\sum_{i=1}^{n}\left(v_i^T\Sigma^{-1}v_i\right) \\
&\quad -\frac{1}{2}\sum_{i=1}^{n}\left(-2\sum_{j=1}^{m}\frac{k\gamma_{ij}}{\sigma_j^2}1_j^T 1_j 1_j^T v_i + \sum_{j=1}^{m}\frac{k\gamma_{ij}^2}{\sigma_j^2}1_j^T 1_j 1_j^T 1_j\right).
\end{aligned}$$

Given all X_i are centred, then μ is centred by Equation (2.18), and $1_j^T\text{vec}(\beta_i X_i\Gamma_i - \mu) = 0$. Therefore, the maximising translation is clearly $\gamma_{ij} = 0$ for all $i = 1, ..., n$ and $j = 1, ..., m$. □

This leads to the following algorithm, which combines CW OPA estimates of the transformation parameters, and maximum likelihood estimates of the mean shape and covariance matrix, to maximise the likelihood in the space orthogonal to the constraint vectors, for the model $\text{vec}(X_i) \sim N_{km}(\text{vec}(\mu), \Sigma)$.

Algorithm 3.2.1 ***Covariance weighted maximum likelihood estimation (CWMLE)***

1. Centre shapes: *Centre* $X_1, X_2, ..., X_n$.

2. Evaluate mean shape and covariance matrix: *Calculate the maximum likelihood estimates of the mean shape, $\widehat{\mu}$, and the covariance matrix, $\widehat{\Sigma}$. Note that the mean shape will also be centred. Evaluate the projection matrix, $U = I_{km} - \sum_{j=1}^{r} u_j(\widehat{\mu})u_j^T(\widehat{\mu})$.*

3. Evaluate log-likelihood: *Calculate the variability in $\widehat{\Sigma}$ orthogonal to the constraint vectors, $\Sigma^* = U\widehat{\Sigma}U^T$. Replace the last r eigenvalues of Σ^* with σ_j^2 for $j = 1, ..., r$, and label it Σ. Evaluate the log-likelihood as,*

$$\begin{aligned}\log L &= -\frac{n}{2}\sum_{j=1}^{r}\log(2\pi\sigma_j^2) - \frac{n}{2}\sum_{j=1}^{km-r}\log(2\pi\lambda_j) \\ &\quad -\frac{1}{2}\sum_{i=1}^{n}\left(vec(\beta_i X_i\Gamma_i - \mu)^T\Sigma^{-1}vec(\beta_i X_i\Gamma_i - \mu)\right),\end{aligned}$$

where λ_j is the jth eigenvalue of Σ^*.

4. CW OPA: *Estimate β_i and Γ_i using CW OPA to register X_i to $\bar{X}_{-i}$, as defined in Equation (2.15), using Σ as the covariance matrix.*

5. Repetition: *Repeat steps (2)-(4) and identify the parameters that give the highest log-likelihood.*

Different starting points for the data may be tried to see if the likelihood can be further improved. Possible starting points may include the output of isotropic GPA or Bookstein registrations of a baseline, see Bookstein (1986).

The convergence properties of the algorithm are dependent on the choice of values for σ_j^2, $j = 1, ..., r$ but the algorithm is almost guaranteed to converge for low values. Choosing small values for σ_j^2, $j = 1, ..., r$ causes the majority of variability in the directions of scaling and rotations to be minimised at each iteration. Consequently, the value of Σ will, given enough iterations, converge and the algorithm essentially reduces to the covariance weighted GPA algorithm of Section 2.4 which is guaranteed to converge. Algorithm 3.2.1 will also generally converge for moderate values of σ_j^2, $j = 1, ..., r$ but for large values, however, each configuration is almost unconstrained to move in the direction of the transformations. For example, a large value

of σ_j^2 corresponding to a rotation could give each shape a large amount of freedom to rotate between successive iterations. In the absence of any other influential constraint each shape can "spin" independently and convergence becomes impossible.

Figure 3.1 shows the registration of 30 second thoracic mouse vertebrae, first considered in Section 2.3.1, following the isotropic partial GPA algorithm and the partial CWMLE algorithm. Note that if σ_j^2 is small then the registration following CWMLE is almost identical to that obtained by isotropic GPA. However, if σ_j^2 is allowed to increase then variability is allowed to remain in the direction of rotation, and lines of landmarks form at tangents to circles centred at the origin.

3.3 Simulation study

The methods of isotropic partial GPA and partial CWMLE were applied to three data sets, originally suggested by Lele (1993), each generated from different multivariate normal distributions,

Example 1: $\text{vec}(X) \sim N_{km}(\text{vec}(\mu_A), \Sigma_A = I_{km})$,

Example 2: $\text{vec}(X) \sim N_{km}(\text{vec}(\mu_B), \Sigma_B = I_m \otimes \Sigma_K)$,

Example 3: $\text{vec}(X) \sim N_{km}(\text{vec}(\mu_C), \Sigma_C = \Sigma_M \otimes \Sigma_K)$,

where,

$$\mu_A = \mu_B = \mu_C = \begin{bmatrix} 0 & 5 \\ 40 & 0 \\ 0 & -5 \\ -40 & 0 \end{bmatrix},$$

$$\Sigma_M = \begin{bmatrix} 0.001 & 0 \\ 0 & 1 \end{bmatrix}, \qquad \Sigma_K = \begin{bmatrix} 0.01 & 0 & 0 & 0 \\ 0 & 10 & 0 & -9.999 \\ 0 & 0 & 0.01 & 0 \\ 0 & -9.999 & 0 & 10 \end{bmatrix}.$$

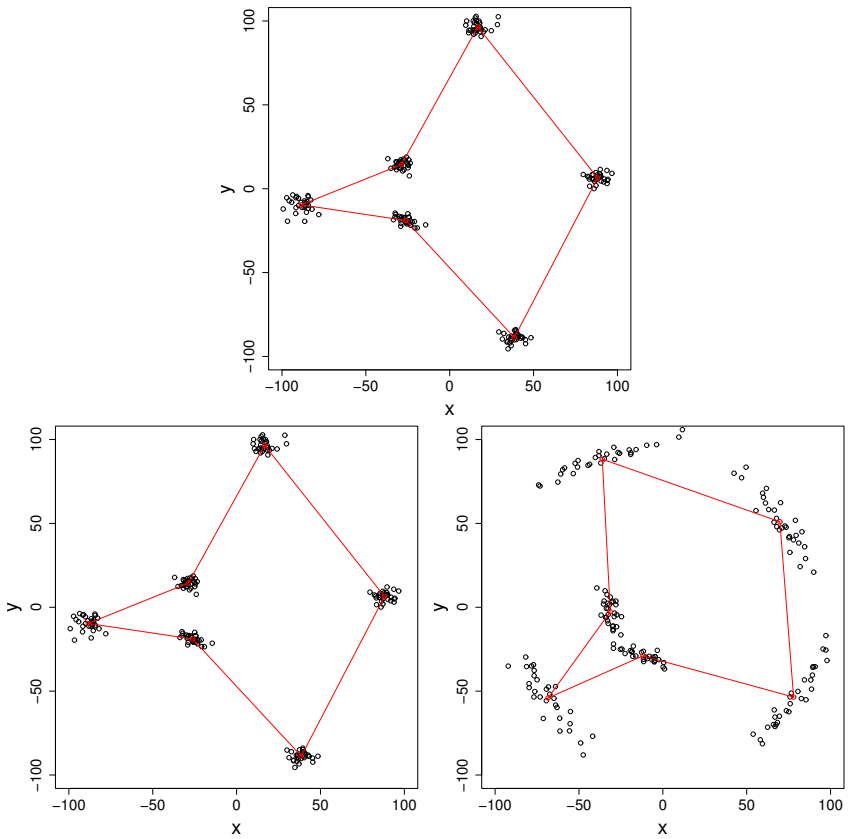

Figure 3.1: The partial covariance weighted Procrustes registration of 30 mice vertebra, using isotropic partial GPA (top) and partial CWMLE (bottom) under the constraint $\sigma_j^2 = 40$ (left), and $\sigma_j^2 = 400$ (right). The mean shape is shown with a solid line.

This makes,

$$\Sigma_C = \begin{bmatrix} 10^{-5} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.01 & 0 & -0.01 & 0 & 0 & 0 & 0 \\ 0 & 0 & 10^{-5} & 0 & 0 & 0 & 0 & 0 \\ 0 & -0.01 & 0 & 0.01 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.01 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 10 & 0 & -9.999 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.01 & 0 \\ 0 & 0 & 0 & 0 & 0 & -9.999 & 0 & 10 \end{bmatrix}.$$

To simulate data from $\text{vec}(X_i) \sim N_{km}(\text{vec}(\mu), \Sigma)$, we calculate $\text{vec}(X_i) = \text{vec}(\mu) + \Sigma^{1/2} W_i$, where W_i is a $km \times 1$ vector such that $W_i \sim N_{km}(0, I_{km})$, for $i = 1, ..., n$. Note that $\Sigma^{1/2}$ is the symmetric square root of Σ, such that $\Sigma^{1/2} = \Gamma \Lambda^{1/2} \Gamma^T$ where Γ is the matrix of eigenvectors of Σ and $\Lambda^{1/2}$ is the diagonal matrix of square roots of eigenvalues of Σ.

Removing the effects of translation and rotation from each of the covariance matrices, using Equation (3.2), gives:

$$\Sigma_A^* = \begin{bmatrix} 0.742 & -0.25 & -0.242 & -0.25 & 0 & 0.062 & 0 & -0.062 \\ -0.25 & 0.75 & -0.25 & -0.25 & 0 & 0 & 0 & 0 \\ -0.242 & -0.25 & 0.742 & -0.25 & 0 & -0.062 & 0 & 0.062 \\ -0.25 & -0.25 & -0.25 & 0.75 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.75 & -0.25 & -0.25 & -0.25 \\ 0.062 & 0 & -0.062 & 0 & -0.25 & 0.258 & -0.25 & 0.242 \\ 0 & 0 & 0 & 0 & -0.25 & -0.25 & 0.75 & -0.25 \\ -0.062 & 0 & 0.062 & 0 & -0.25 & 0.242 & -0.25 & 0.258 \end{bmatrix},$$

$$\Sigma_B^* = \begin{bmatrix} 0.158 & -0.001 & -0.155 & -0.001 & 0 & 0.02 & 0 & -0.02 \\ -0.001 & 10.001 & -0.001 & -9.998 & 0 & 0 & 0 & 0 \\ -0.155 & -0.001 & 0.158 & -0.001 & 0 & -0.02 & 0 & 0.02 \\ -0.001 & -9.998 & -0.001 & 10.001 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.006 & -0.001 & -0.004 & -0.001 \\ 0.02 & 0 & -0.02 & 0 & -0.001 & 0.004 & -0.001 & -0.001 \\ 0 & 0 & 0 & 0 & -0.004 & -0.001 & 0.006 & -0.001 \\ -0.02 & 0 & 0.02 & 0 & -0.001 & -0.001 & -0.001 & 0.004 \end{bmatrix},$$

$$\Sigma_C^* = \begin{bmatrix} 0.151 & 0 & -0.151 & 0 & 0 & 0.019 & 0 & -0.019 \\ 0 & 0.01 & 0 & -0.01 & 0 & 0 & 0 & 0 \\ -0.151 & 0 & 0.151 & 0 & 0 & -0.019 & 0 & 0.019 \\ 0 & -0.01 & 0 & 0.01 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.006 & -0.001 & -0.004 & -0.001 \\ 0.019 & 0 & -0.019 & 0 & -0.001 & 0.004 & -0.001 & -0.001 \\ 0 & 0 & 0 & 0 & -0.004 & -0.001 & 0.006 & -0.001 \\ -0.019 & 0 & 0.019 & 0 & -0.001 & -0.001 & -0.001 & 0.004 \end{bmatrix}.$$

3.3.1 Single simulations

For each example, 130 configurations were sampled from these distributions. The isotropic partial GPA algorithm, given in Chapter 1, and the partial CWMLE algorithm, Algorithm 3.2.1, were applied with the starting point for the latter being the registration given by isotropic partial GPA. Let $s_A \approx 5$, $s_B \approx 20$ and $s_c \approx 0.5$, be the sum of the eigenvalues of $\widehat{\Sigma}_A$, $\widehat{\Sigma}_B$ and $\widehat{\Sigma}_C$ estimated following the isotropic partial GPA algorithm. We choose $\sigma_1^2 = \sigma_2^2 = 10^{-5}$ for each example. We choose *a priori* σ_3^2 to be $\sigma_A^2 = 10^{-5}$, $\sigma_B^2 = 20$ and $\sigma_C^2 = 5$ for the three examples, respectively, as we expect the ideal registration for the three examples to have no $(\sigma_A^2/(\sigma_A^2 + s_A) \approx 0)$, moderate $(\sigma_B^2/(\sigma_B^2 + s_B) \approx 0.5)$ and high $(\sigma_C^2/(\sigma_C^2 + s_C) \approx 0.9)$ proportion of variability in the direction of rotation. The values of μ, Σ and Σ^* that gave the highest likelihood of the model for each example/algorithm are given below. Plots of the initial data, the registered data and the first two principal components of Σ are shown in Figures 3.2 to 3.7. We discuss the findings in Section 3.3.2.

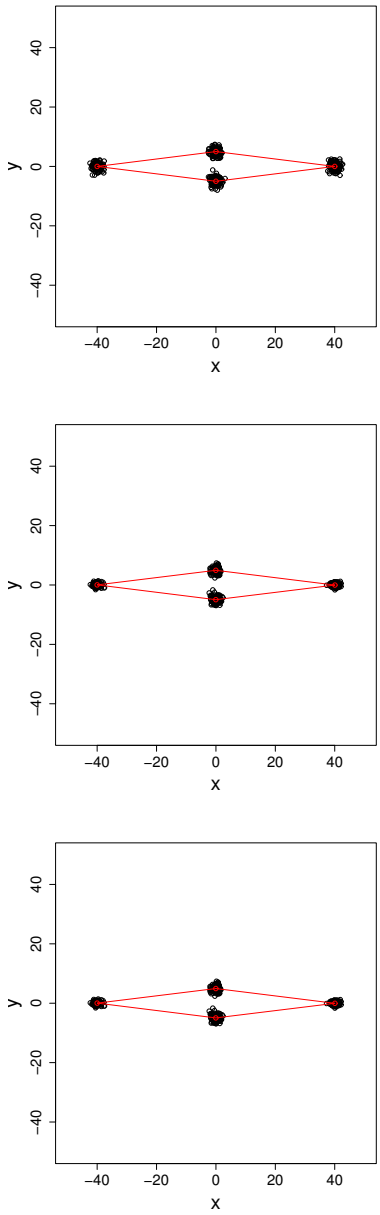

Figure 3.2: The data for example 1 (top), following the isotropic partial GPA algorithm (middle) and partial CWMLE algorithm (bottom). The mean shape is shown with a solid line.

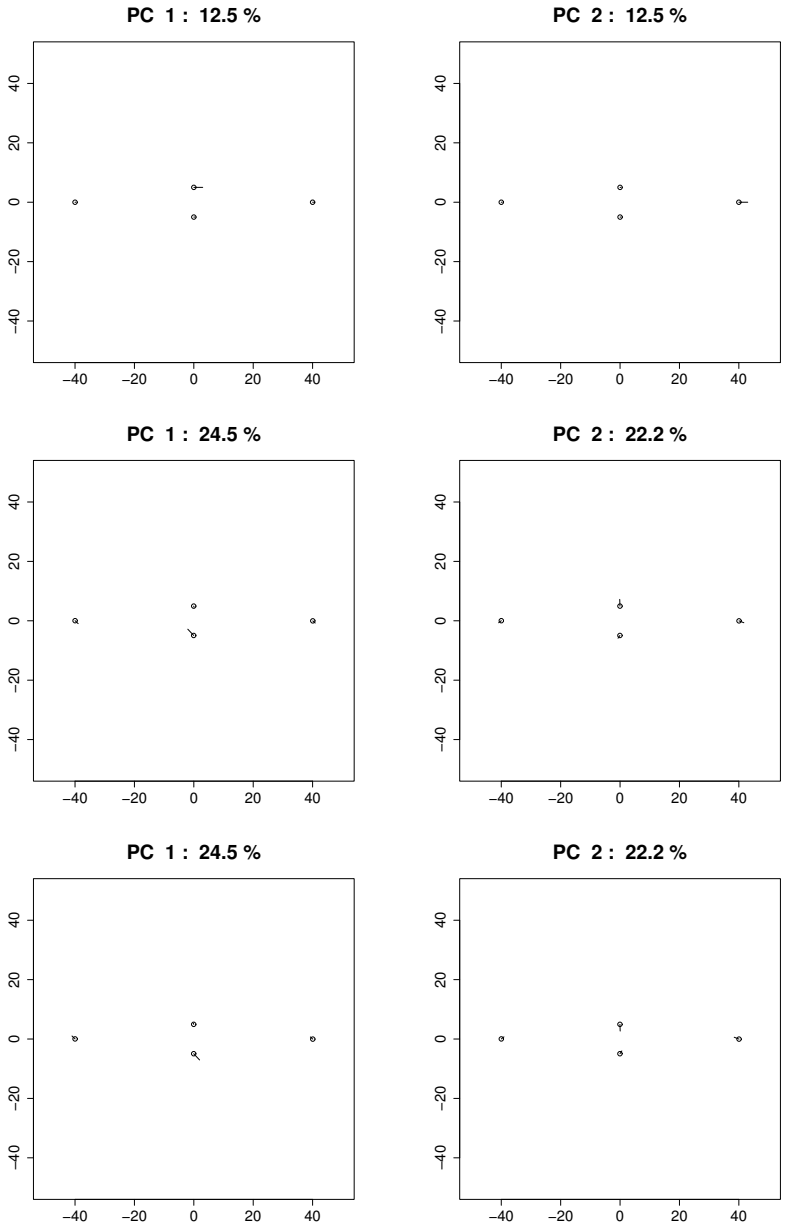

Figure 3.3: The first two principal components of Σ_A (top) and for the maximum likelihood estimate, $\widehat{\Sigma}_A$, following isotropic partial GPA (middle) and partial CWMLE (bottom). The mean shape plus three times the principal component vectors are shown.

Example 1: Isotropic GPA

$$\widehat{\mu}_A = \begin{bmatrix} -0.035 & 4.942 \\ 40.059 & -0.029 \\ -0.04 & -4.945 \\ -39.985 & 0.031 \end{bmatrix},$$

$$\widehat{\Sigma}_A = \begin{bmatrix} 0.666 & -0.288 & -0.218 & -0.16 & 0.048 & 0.029 & 0.005 & -0.081 \\ -0.288 & 0.825 & -0.322 & -0.215 & 0.058 & -0.026 & -0.003 & -0.029 \\ -0.218 & -0.322 & 0.86 & -0.321 & -0.005 & -0.022 & -0.085 & 0.112 \\ -0.16 & -0.215 & -0.321 & 0.696 & -0.1 & 0.019 & 0.083 & -0.002 \\ 0.048 & 0.058 & -0.005 & -0.1 & 0.831 & -0.267 & -0.29 & -0.274 \\ 0.029 & -0.026 & -0.022 & 0.019 & -0.267 & 0.284 & -0.295 & 0.279 \\ 0.005 & -0.003 & -0.085 & 0.083 & -0.29 & -0.295 & 0.893 & -0.308 \\ -0.081 & -0.029 & 0.112 & -0.002 & -0.274 & 0.279 & -0.308 & 0.303 \end{bmatrix},$$

$$\widehat{\Sigma}^*_A = \begin{bmatrix} 0.666 & -0.288 & -0.218 & -0.16 & 0.048 & 0.029 & 0.005 & -0.081 \\ -0.288 & 0.825 & -0.322 & -0.215 & 0.058 & -0.026 & -0.003 & -0.029 \\ -0.218 & -0.322 & 0.86 & -0.321 & -0.005 & -0.022 & -0.085 & 0.112 \\ -0.16 & -0.215 & -0.321 & 0.696 & -0.1 & 0.019 & 0.083 & -0.002 \\ 0.048 & 0.058 & -0.005 & -0.1 & 0.831 & -0.267 & -0.29 & -0.274 \\ 0.029 & -0.026 & -0.022 & 0.019 & -0.267 & 0.284 & -0.295 & 0.279 \\ 0.005 & -0.003 & -0.085 & 0.083 & -0.29 & -0.295 & 0.893 & -0.308 \\ -0.081 & -0.029 & 0.112 & -0.002 & -0.274 & 0.279 & -0.308 & 0.303 \end{bmatrix}.$$

Example 1: CWMLE

$$\widehat{\mu}_A = \begin{bmatrix} -0.035 & 4.942 \\ 40.059 & -0.029 \\ -0.04 & -4.945 \\ -39.985 & 0.031 \end{bmatrix},$$

$$\widehat{\Sigma}_A = \begin{bmatrix} 0.666 & -0.288 & -0.218 & -0.16 & 0.048 & 0.029 & 0.005 & -0.081 \\ -0.288 & 0.825 & -0.322 & -0.215 & 0.058 & -0.026 & -0.003 & -0.029 \\ -0.218 & -0.322 & 0.86 & -0.321 & -0.005 & -0.022 & -0.085 & 0.112 \\ -0.16 & -0.215 & -0.321 & 0.696 & -0.1 & 0.019 & 0.083 & -0.002 \\ 0.048 & 0.058 & -0.005 & -0.1 & 0.831 & -0.267 & -0.29 & -0.274 \\ 0.029 & -0.026 & -0.022 & 0.019 & -0.267 & 0.284 & -0.295 & 0.279 \\ 0.005 & -0.003 & -0.085 & 0.083 & -0.29 & -0.295 & 0.893 & -0.308 \\ -0.081 & -0.029 & 0.112 & -0.002 & -0.274 & 0.279 & -0.308 & 0.303 \end{bmatrix},$$

$$\widehat{\Sigma}^*_A = \begin{bmatrix} 0.666 & -0.288 & -0.218 & -0.16 & 0.048 & 0.029 & 0.005 & -0.081 \\ -0.288 & 0.825 & -0.322 & -0.215 & 0.058 & -0.026 & -0.003 & -0.029 \\ -0.218 & -0.322 & 0.86 & -0.321 & -0.005 & -0.022 & -0.085 & 0.112 \\ -0.16 & -0.215 & -0.321 & 0.696 & -0.1 & 0.019 & 0.083 & -0.002 \\ 0.048 & 0.058 & -0.005 & -0.1 & 0.831 & -0.267 & -0.29 & -0.274 \\ 0.029 & -0.026 & -0.022 & 0.019 & -0.267 & 0.284 & -0.295 & 0.279 \\ 0.005 & -0.003 & -0.085 & 0.083 & -0.29 & -0.295 & 0.893 & -0.308 \\ -0.081 & -0.029 & 0.112 & -0.002 & -0.274 & 0.279 & -0.308 & 0.303 \end{bmatrix}.$$

Example 2: Isotropic GPA

$$\widehat{\mu}_B = \begin{bmatrix} -0.005 & 4.976 \\ 40.219 & -0.129 \\ 0 & -4.976 \\ -40.213 & 0.13 \end{bmatrix},$$

$$\widehat{\Sigma}_B = \begin{bmatrix} 0.182 & -0.036 & -0.184 & 0.039 & 0.006 & 0.023 & -0.006 & -0.023 \\ -0.036 & 9.659 & 0.021 & -9.644 & 0.043 & -0.038 & -0.036 & 0.031 \\ -0.184 & 0.021 & 0.193 & -0.03 & -0.006 & -0.023 & 0.005 & 0.024 \\ 0.039 & -9.644 & -0.03 & 9.635 & -0.043 & 0.038 & 0.036 & -0.032 \\ 0.006 & 0.043 & -0.006 & -0.043 & 0.007 & -0.001 & -0.005 & -0.002 \\ 0.023 & -0.038 & -0.023 & 0.038 & -0.001 & 0.005 & -0.002 & -0.001 \\ -0.006 & -0.036 & 0.005 & 0.036 & -0.005 & -0.002 & 0.008 & -0.001 \\ -0.023 & 0.031 & 0.024 & -0.032 & -0.002 & -0.001 & -0.001 & 0.004 \end{bmatrix},$$

$$\widehat{\Sigma}^*_B = \begin{bmatrix} 0.182 & -0.036 & -0.184 & 0.039 & 0.006 & 0.023 & -0.006 & -0.023 \\ -0.036 & 9.659 & 0.021 & -9.644 & 0.043 & -0.038 & -0.036 & 0.031 \\ -0.184 & 0.021 & 0.193 & -0.03 & -0.006 & -0.023 & 0.005 & 0.024 \\ 0.039 & -9.644 & -0.03 & 9.635 & -0.043 & 0.039 & 0.036 & -0.032 \\ 0.006 & 0.043 & -0.006 & -0.043 & 0.007 & -0.001 & -0.005 & -0.002 \\ 0.023 & -0.038 & -0.023 & 0.039 & -0.001 & 0.005 & -0.002 & -0.001 \\ -0.006 & -0.036 & 0.005 & 0.036 & -0.005 & -0.002 & 0.008 & -0.001 \\ -0.023 & 0.031 & 0.024 & -0.032 & -0.002 & -0.001 & -0.001 & 0.005 \end{bmatrix}.$$

Example 2: CWMLE

$$\widehat{\mu}_B = \begin{bmatrix} -0.02 & 4.969 \\ 40.178 & -0.005 \\ 0.014 & -4.97 \\ -40.172 & 0.005 \end{bmatrix},$$

$$\widehat{\Sigma}_B = \begin{bmatrix} 0.248 & -0.084 & -0.251 & 0.087 & 0.004 & -0.45 & -0.004 & 0.45 \\ -0.084 & 9.672 & 0.07 & -9.658 & 0.036 & 0.39 & -0.028 & -0.398 \\ -0.251 & 0.07 & 0.26 & -0.079 & -0.005 & 0.451 & 0.003 & -0.45 \\ 0.087 & -9.658 & -0.079 & 9.649 & -0.036 & -0.391 & 0.029 & 0.398 \\ 0.004 & 0.036 & -0.005 & -0.036 & 0.004 & -0.004 & -0.002 & 0.002 \\ -0.45 & 0.39 & 0.451 & -0.391 & -0.004 & 3.36 & -0.006 & -3.349 \\ -0.004 & -0.028 & 0.003 & 0.029 & -0.002 & -0.006 & 0.006 & 0.003 \\ 0.45 & -0.398 & -0.45 & 0.398 & 0.002 & -3.349 & 0.003 & 3.344 \end{bmatrix},$$

$$\widehat{\Sigma}^*_B = \begin{bmatrix} 0.182 & -0.035 & -0.185 & 0.038 & 0.004 & 0.023 & -0.004 & -0.023 \\ -0.035 & 9.672 & 0.021 & -9.658 & 0.036 & -0.009 & -0.029 & 0.001 \\ -0.185 & 0.021 & 0.194 & -0.03 & -0.004 & -0.023 & 0.004 & 0.024 \\ 0.038 & -9.658 & -0.03 & 9.649 & -0.036 & 0.009 & 0.029 & -0.002 \\ 0.004 & 0.036 & -0.004 & -0.036 & 0.004 & -0.001 & -0.002 & -0.002 \\ 0.023 & -0.009 & -0.023 & 0.009 & -0.001 & 0.004 & -0.002 & -0.001 \\ -0.004 & -0.029 & 0.004 & 0.029 & -0.002 & -0.002 & 0.006 & -0.001 \\ -0.023 & 0.001 & 0.024 & -0.002 & -0.002 & -0.001 & -0.001 & 0.004 \end{bmatrix}.$$

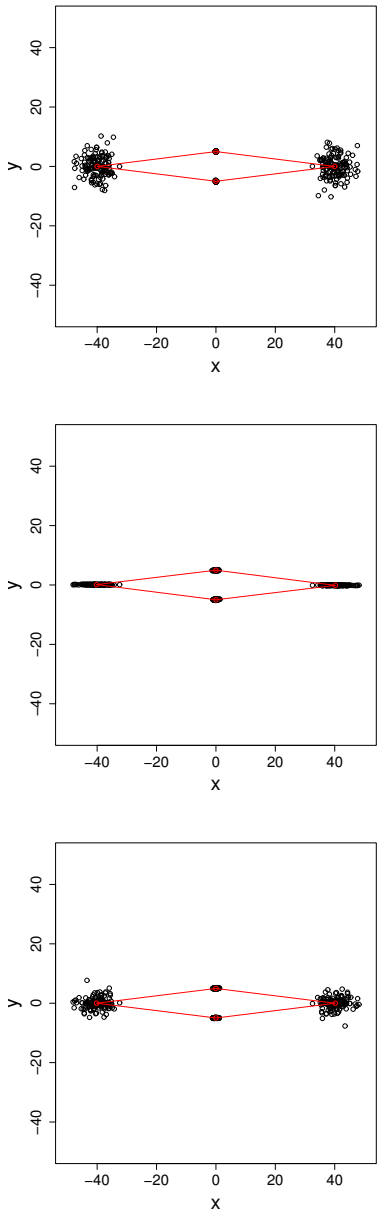

Figure 3.4: The data for example 2 (top), following the isotropic partial GPA algorithm (middle) and partial CWMLE algorithm (bottom). The mean shape is shown with a solid line.

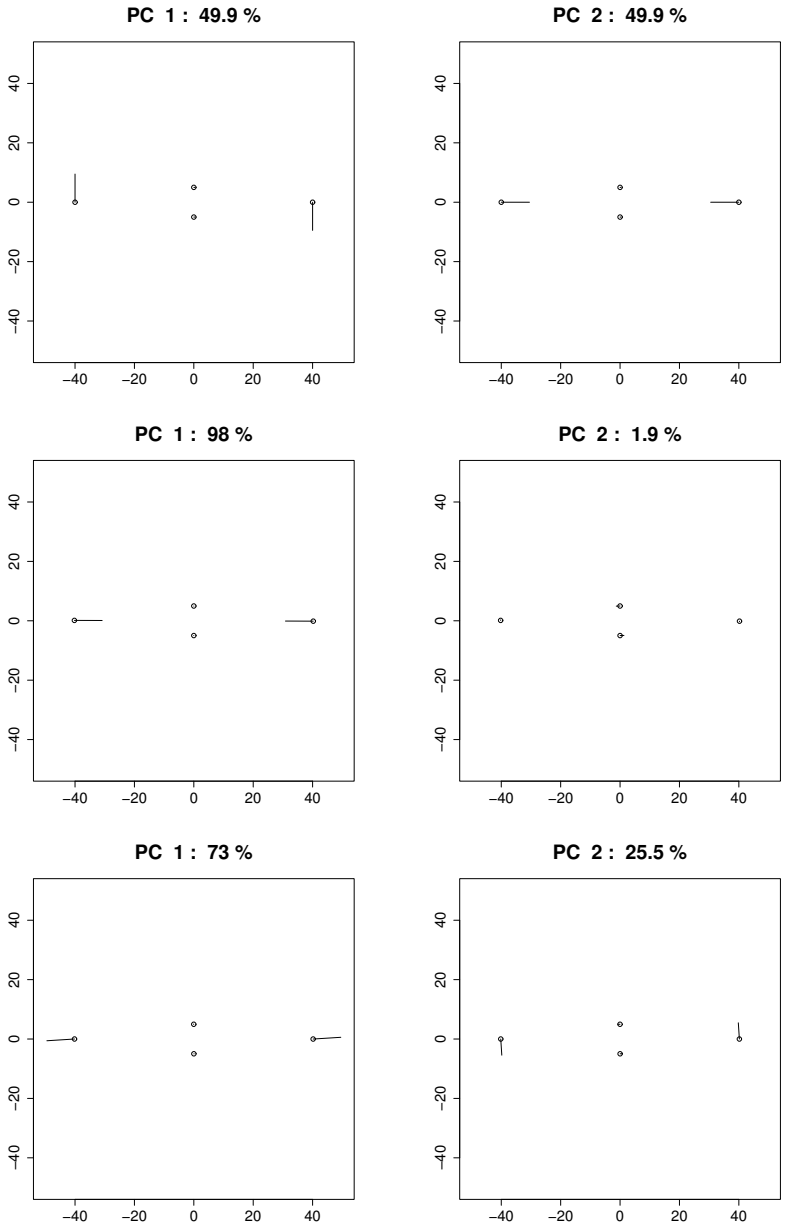

Figure 3.5: The first two principal components of Σ_B (top) and for the maximum likelihood estimate, $\widehat{\Sigma}_B$, following isotropic partial GPA (middle) and partial CWMLE (bottom). The mean shape plus three times the principal component vectors are shown.

Example 3: Isotropic GPA

$$\widehat{\mu}_C = \begin{bmatrix} -0.003 & 4.976 \\ 40.151 & -0.126 \\ 0.003 & -4.977 \\ -40.15 & 0.127 \end{bmatrix},$$

$$\widehat{\Sigma}_C = \begin{bmatrix} 0.177 & -0.013 & -0.177 & 0.013 & 0.005 & 0.022 & -0.005 & -0.022 \\ -0.013 & 0.055 & 0.013 & -0.055 & -0.002 & -0.003 & 0.004 & 0.001 \\ -0.177 & 0.013 & 0.177 & -0.012 & -0.005 & -0.022 & 0.005 & 0.022 \\ 0.013 & -0.055 & -0.012 & 0.054 & 0.002 & 0.003 & -0.004 & -0.001 \\ 0.005 & -0.002 & -0.005 & 0.002 & 0.007 & -0.001 & -0.005 & -0.002 \\ 0.022 & -0.003 & -0.022 & 0.003 & -0.001 & 0.004 & -0.002 & -0.001 \\ -0.005 & 0.004 & 0.005 & -0.004 & -0.005 & -0.002 & 0.008 & -0.001 \\ -0.022 & 0.001 & 0.022 & -0.001 & -0.002 & -0.001 & -0.001 & 0.004 \end{bmatrix},$$

$$\widehat{\Sigma}^*_C = \begin{bmatrix} 0.177 & -0.013 & -0.177 & 0.013 & 0.005 & 0.022 & -0.005 & -0.022 \\ -0.013 & 0.055 & 0.013 & -0.055 & -0.002 & -0.002 & 0.004 & 0.001 \\ -0.177 & 0.013 & 0.177 & -0.012 & -0.005 & -0.022 & 0.005 & 0.022 \\ 0.013 & -0.055 & -0.012 & 0.054 & 0.002 & 0.003 & -0.004 & -0.001 \\ 0.005 & -0.002 & -0.005 & 0.002 & 0.007 & -0.001 & -0.005 & -0.002 \\ 0.022 & -0.002 & -0.022 & 0.003 & -0.001 & 0.004 & -0.002 & -0.001 \\ -0.005 & 0.004 & 0.005 & -0.004 & -0.005 & -0.002 & 0.008 & -0.001 \\ -0.022 & 0.001 & 0.022 & -0.001 & -0.002 & -0.001 & -0.001 & 0.004 \end{bmatrix}.$$

Example 3: CWMLE

$$\widehat{\mu}_C = \begin{bmatrix} -0.016 & 4.987 \\ 40.099 & -0.028 \\ 0.016 & -4.989 \\ -40.099 & 0.03 \end{bmatrix},$$

$$\widehat{\Sigma}_C = \begin{bmatrix} 0.062 & -0.007 & -0.062 & 0.007 & 0.002 & 0.233 & -0.001 & -0.234 \\ -0.007 & 0.027 & 0.007 & -0.027 & 0.001 & -0.027 & 0 & 0.026 \\ -0.062 & 0.007 & 0.063 & -0.007 & -0.001 & -0.234 & 0.001 & 0.234 \\ 0.007 & -0.027 & -0.007 & 0.027 & -0.001 & 0.027 & 0 & -0.026 \\ 0.002 & 0.001 & -0.001 & -0.001 & 0.007 & 0.021 & -0.004 & -0.024 \\ 0.233 & -0.027 & -0.234 & 0.027 & 0.021 & 4.159 & -0.029 & -4.151 \\ -0.001 & 0 & 0.001 & 0 & -0.004 & -0.029 & 0.007 & 0.026 \\ -0.234 & 0.026 & 0.234 & -0.026 & -0.024 & -4.151 & 0.026 & 4.149 \end{bmatrix},$$

$$\widehat{\Sigma}^*_C = \begin{bmatrix} 0.179 & -0.011 & -0.179 & 0.011 & 0.005 & 0.023 & -0.005 & -0.022 \\ -0.011 & 0.027 & 0.011 & -0.027 & 0.001 & -0.002 & 0 & 0.001 \\ -0.179 & 0.011 & 0.179 & -0.011 & -0.005 & -0.023 & 0.005 & 0.022 \\ 0.011 & -0.027 & -0.011 & 0.027 & -0.001 & 0.002 & 0 & -0.001 \\ 0.005 & 0.001 & -0.005 & -0.001 & 0.007 & -0.001 & -0.004 & -0.002 \\ 0.023 & -0.002 & -0.023 & 0.002 & -0.001 & 0.004 & -0.002 & -0.001 \\ -0.005 & 0 & 0.005 & 0 & -0.004 & -0.002 & 0.007 & -0.001 \\ -0.022 & 0.001 & 0.022 & -0.001 & -0.002 & -0.001 & -0.001 & 0.004 \end{bmatrix}.$$

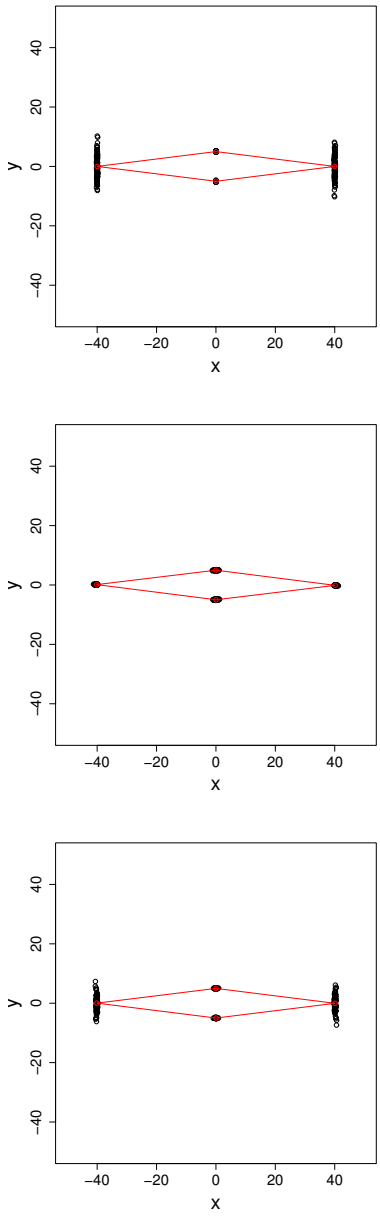

Figure 3.6: The data for example 3 (top), following the isotropic partial GPA algorithm (middle) and partial CWMLE algorithm (bottom). The mean shape is shown with a solid line.

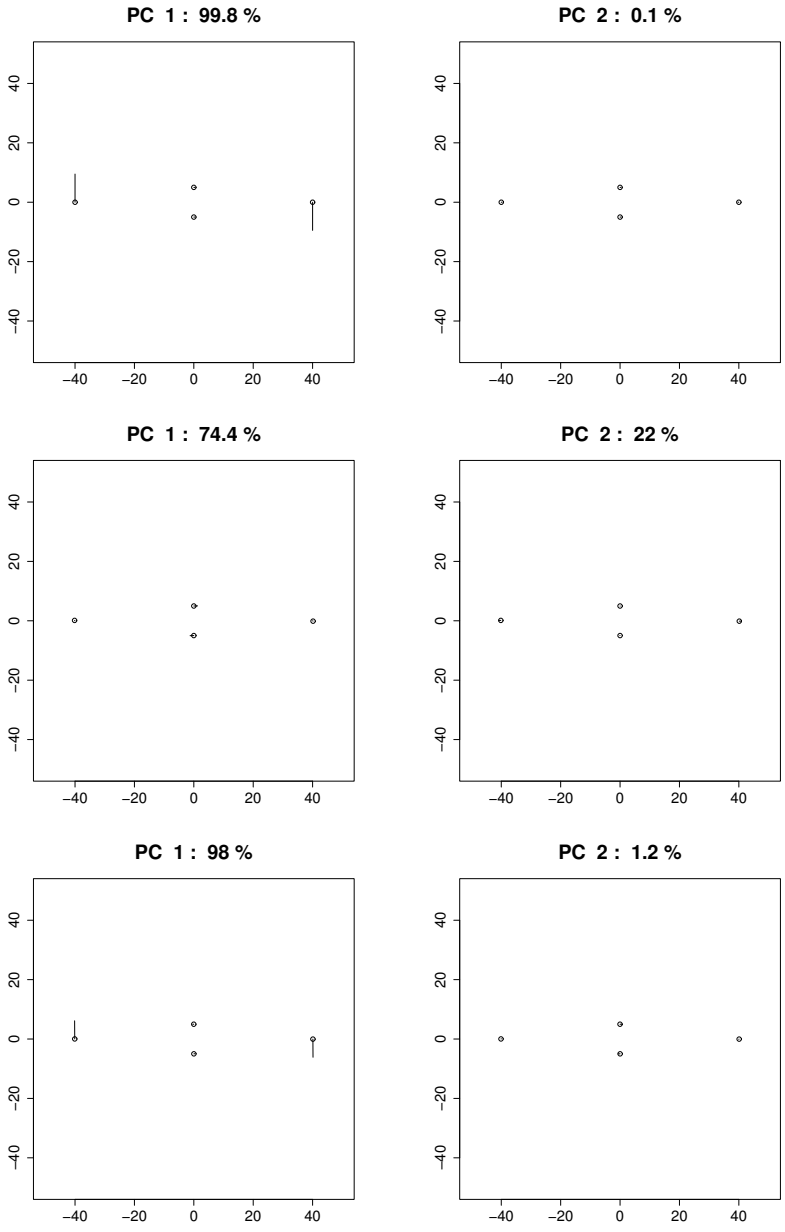

Figure 3.7: The first two principal components of Σ_C (top) and for the maximum likelihood estimate, $\widehat{\Sigma}_C$, following isotropic partial GPA (middle) and partial CWMLE (bottom). The mean shape plus three times the principal component vectors are shown.

3.3.2 Repeated simulations

To assess the difference between isotropic GPA and CWMLE, $N = 1000$ Monte Carlo samples of each example shown in Section 3.3.1 were generated. Following registration, the data were rotated to minimise the Euclidean distance between $\widehat{\mu}$, the estimated mean shape, and μ, the true mean shape of the model. The estimates of the covariance matrix, $\widehat{\Sigma}$, and the projected covariance matrix, $\widehat{\Sigma}^*$, were adjusted accordingly. This step removes the arbitrary rotation of the mean shape. The bias of $\widehat{\mu}$ and the root mean square error of $\widehat{\mu}$, $\widehat{\Sigma}$ and $\widehat{\Sigma}^*$ were calculated using,

$$\text{Bias}(\widehat{A}) = \frac{1}{N}\sum_{i=1}^{N}\widehat{A} - A,$$

$$\text{RMSE}(\widehat{A}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\|\widehat{A} - A\|^2}.$$

The results for each of the examples are given in Tables 3.1, 3.2 and 3.3.

	Isotropic GPA	CWMLE
$\text{Bias}(\widehat{\mu}_A)$	$\begin{bmatrix} -0.004 & 0 \\ 0.004 & -0.002 \\ 0.006 & 0.004 \\ -0.006 & -0.001 \end{bmatrix}$	$\begin{bmatrix} -0.004 & 0 \\ 0.004 & -0.002 \\ 0.006 & 0.004 \\ -0.006 & -0.001 \end{bmatrix}$
$\text{RMSE}(\widehat{\mu}_A)$	0.431	0.431
$\text{RMSE}(\widehat{\Sigma}_A)$	1.341	1.341
$\text{RMSE}(\widehat{\Sigma}^*_A)$	0.69	0.69

Table 3.1: The bias and root mean square error of the parameter estimates for example 1.

The simulation study shows that CWMLE can reduce the bias and root mean square error of the mean shape and covariance matrix estimates compared to isotropic GPA. This is particularly clear in the examples where the

	Isotropic GPA	CWMLE
Bias($\widehat{\mu}_B$)	$\begin{bmatrix} 0 & -0.015 \\ 0.119 & 0 \\ 0 & 0.016 \\ -0.119 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & -0.033 \\ 0.017 & 0 \\ 0 & 0.033 \\ -0.018 & 0 \end{bmatrix}$
RMSE($\widehat{\mu}_B$)	0.588	0.596
RMSE($\widehat{\Sigma}_B$)	4.489	3.849
RMSE($\widehat{\Sigma}_B^*$)	1.412	1.416

Table 3.2: The bias and root mean square error of the parameter estimates for example 2.

	Isotropic GPA	CWMLE
Bias($\widehat{\mu}_C$)	$\begin{bmatrix} 0.016 & -0.011 \\ 0.133 & 0 \\ -0.016 & 0.014 \\ -0.132 & -0.004 \end{bmatrix}$	$\begin{bmatrix} 0.016 & -0.003 \\ 0.05 & 0 \\ -0.016 & 0.006 \\ -0.05 & -0.004 \end{bmatrix}$
RMSE($\widehat{\mu}_C$)	0.428	0.304
RMSE($\widehat{\Sigma}_C$)	4.472	2.968
RMSE($\widehat{\Sigma}_C^*$)	0.285	0.217

Table 3.3: The bias and root mean square error of the parameter estimates for example 3.

true covariance matrix is far from isotropic. In particular, the output for example 3 shows a universal reduction in errors of both the mean shape and covariance matrix. Example 3, where much of the variability is in the direction of a rotation, is a classic example where isotropic Procrustes produces an inconsistent estimate of shape variability, see Lele (1993).

Conversely, example 1 where the true covariance matrix is isotropic, shows that CWMLE gives nearly identical results to isotropic Procrustes, when the rotation eigenvalue is small. Interestingly, because there is little rotation in the original data set, increasing the rotation eigenvalue has little effect, showing that the choice of constraint is somewhat academic in the isotropic case. This gives the analyst freedom to use CWMLE instead of isotropic Procrustes with the assurance that the resulting registration will be no worse than isotropic registration.

In all three examples the single eigenvalue constraint has been set *a priori*. The effect of reducing the eigenvalue will produce a registration that tends towards the isotropic registration. In examples 2 and 3, we see that CWMLE with this choice of eigenvalue greatly reduces the bias in the estimate of the mean shape. The choice of eigenvalue, however, is a little conservative, and increasing the value will reduce the error in the estimates of μ and Σ. This can be seen in the principal components of example 2, where the CWMLE algorithm produces a split in the two orthogonal directions roughly half way between those produced by isotropic GPA and the initial data set. Isotropic GPA induces large rotations for some of the shapes in examples 2 and 3, which causes the comparatively large bias in the x co-ordinate for landmarks 2 and 4. Increasing the eigenvalue beyond the optimal value will induce even larger rotations in the CWMLE algorithm, so the choice of eigenvalue is critical.

In summary, generalising Procrustes methods to minimise a Mahalanobis norm, rather than a Euclidean norm, dramatically reduces the error of Procrustes methods in covariance matrix estimation and this is also true of the mean shape for extreme covariance structures. However, Lele and McCul-

loch (2002) highlights that the isotropic Procrustes estimator of the shape covariance matrix is inconsistent for non-isotropic variability, as discussed in full by Kent and Mardia (1997), and future work is required to discover whether or not the CWMLE estimator is consistent.

3.4 A Bayesian approach to CW Procrustes

3.4.1 The model

An alternative to classical maximum likelihood estimation of the unknown parameters, is to estimate their posterior distributions using Bayesian inference and Markov Chain Monte Carlo (MCMC) simulation. For a detailed description of MCMC methods, see Gilks *et al.* (1996). For brevity, we will use the notation $\{X_i\}$ to imply the collection of shapes, X_i, $i = 1, ..., n$, and similar notation for the transformation parameters. We will assume in this section that scaling is included, although results for partial Procrustes analysis can be inferred by removing the β_i parameters throughout.

We assume that after differences in size, orientation and location have been removed from each configuration, the shapes follow a multivariate normal model with mean, μ, and covariance matrix, Σ. The shape differences are removed by applying the transformations, β_i, Γ_i and γ_i to X_i for $i = 1, ..., n$, where Γ_i is the $m \times m$ rotation matrix that is a function of the rotation angles, θ_{ij}, $j = 1, ..., m(m-1)/2$. Therefore,

$$\text{vec}(\beta_i X_i \Gamma_i + 1_k \gamma_i^T) \sim N_{km}(\text{vec}(\mu), \Sigma).$$

Likelihood. Let $\widetilde{X}_i = \beta_i X_i \Gamma_i + 1_k \gamma_i^T$ for $i = 1, ..., n$, then the likelihood of the multivariate normal model is,

$$L\left(\{X_i\} | \{\beta_i, \theta_{ij}, \gamma_i\}, \mu, \Sigma\right)$$

$$
\begin{aligned}
&= \prod_{i=1}^{n} L(X_i|\beta_i, \Gamma_i, \gamma_i, \mu, \Sigma), \\
&= |2\pi\Sigma|^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \mathrm{vec}(\widetilde{X}_i - \mu)^T \Sigma^{-1} \mathrm{vec}(\widetilde{X}_i - \mu) \right\}.
\end{aligned}
$$

Prior distributions. In Section 3.2 we included *a priori* information in the classical maximum likelihood framework through introducing constraints on Σ. With a Bayesian approach, we include *a priori* information by specifying multivariate normal and Wishart prior distributions for the mean shape and inverse covariance matrix respectively,

$$
\mathrm{vec}(\mu) \sim N_{km}(\mathrm{vec}(\mu_0), \Sigma_{\mu 0}), \qquad \Sigma^{-1} \sim W_{km}(\Sigma_0^{-1}, km), \tag{3.3}
$$

where μ_0, $\Sigma_{\mu 0}$ and Σ_0 are fixed quantities based on prior beliefs. The degrees of freedom for the Wishart distribution is minimised to make it as least informative as possible, see Johnson and Kotz (1972).

In addition, we specify uniform prior distributions for the n scaling parameters, $nm(m-1)/2$ rotation angles and n translation vectors of length m, to allow for arbitrary starting sizes, orientations and locations,

$$
\beta_i \sim U(0, L), \quad \theta_{ij} \sim U_{m(m-1)/2}(-\pi, \pi), \quad \gamma_i \sim U_m(-R/2, R/2),
$$

$i = 1, ..., n$, where R and L are large positive constants such that all the shapes are initially contained within a box of length R, and L is an upper bound on any possible scaling. The angles θ and $\theta + 2\pi$ are equivalent, so angles outside the range $(-\pi, \pi)$ are wrapped to be within this range. If necessary, an initial manual scaling and rotation of each configuration will remove any need to sample, in an MCMC algorithm, parameter values lying near the boundary of the prior distributions' support.

Posterior distribution. By Bayes' theorem, the joint posterior distri-

bution for the model parameters is,

$$\begin{aligned}
&\pi(\{\beta_i, \theta_{ij}, \gamma_i\}, \mu, \Sigma^{-1} | \{X_i\}) \\
&\quad \propto \quad P(\{\beta_i\})P(\{\theta_{ij}\})P(\{\gamma_i\})P(\mu)P(\Sigma^{-1})L(\{X_i\}|\{\beta_i, \theta_{ij}, \gamma_i\}, \mu, \Sigma^{-1}), \\
&\quad \propto \quad P(\mu)P(\Sigma^{-1})L(\{X_i\}|\{\beta_i, \theta_{ij}, \gamma_i\}, \mu, \Sigma^{-1}),
\end{aligned}$$

where π, P and L denote the posterior distribution, prior distributions and likelihood respectively. The prior distributions for the transformation parameters can be included within the proportionality constant because they are uniform.

3.4.2 Conditional posterior distributions

A conventional MCMC algorithm for the data is outlined in Algorithm 3.4.1.

Algorithm 3.4.1 ***A conventional MCMC algorithm for shape data***

1. *Produce initial samples of* β_i, θ_{ij}, γ_i, μ *and* Σ *from prior distributions for* $i = 1, ..., n$ *and* $j = 1, ..., m(m-1)/2$.

2. *Sample* $\{\beta_i\}$ *from their conditional posterior distributions (Gibbs steps).*

3. *Update* $\{\theta_{ij}\}$ *by Metropolis-Hastings steps.*

4. *Sample* $\{\gamma_i\}$ *from their conditional posterior distributions (Gibbs steps).*

5. *Sample* μ *from its conditional posterior distribution (Gibbs step).*

6. *Sample* Σ *from its conditional posterior distribution (Gibbs step).*

7. *Repeat steps (2)-(6).*

This algorithm updates the value of each parameter once in each iteration. Most of the parameters can be updated by sampling from their conditional posterior distribution. In the rest of this subsection, we calculate each parameter's conditional posterior distribution and explain why we update the rotation angles with a Metropolis-Hastings step.

Updating the ith scaling parameter. Let $\{\beta_i'\}$ be $\{\beta_i\}$ excluding the ith scaling parameter, then the conditional posterior distribution of the ith scaling parameter, $\beta_i > 0$, is,

$$\begin{aligned}
&\pi(\beta_i|\{X_i\},\{\beta_i'\},\{\theta_{ij}\},\{\gamma_i\},\mu,\Sigma^{-1}) \\
&\quad\propto\ P(\beta_i)L(\{X_i\}|\{\beta_i\},\{\theta_{ij}\},\{\gamma_i\},\mu,\Sigma^{-1}), \\
&\quad\propto\ \exp\left\{-\frac{1}{2}\mathrm{vec}(\mu-\beta_i X_i\Gamma_i-1_k\gamma_i^T)^T\Sigma^{-1}\mathrm{vec}(\mu-\beta_i X_i\Gamma_i-1_k\gamma_i^T)\right\}, \\
&\quad\propto\ \exp\left\{-\frac{1}{2}[\beta_i^2\mathrm{vec}(X_i\Gamma_i)^T\Sigma^{-1}\mathrm{vec}(X_i\Gamma_i)\right. \\
&\qquad\left.-2\beta_i\mathrm{vec}(X_i\Gamma_i)^T\Sigma^{-1}\mathrm{vec}(\mu-1_k\gamma_i^T)]\right\}, \\
&\quad\propto\ \exp\left\{-\frac{1}{2\sigma_{i\beta}^2}(\beta_i-\beta_{ip})^2\right\},
\end{aligned}$$

where $\sigma_{i\beta}^2 = \left[\mathrm{vec}(X_i\Gamma_i)^T\Sigma^{-1}\mathrm{vec}(X_i\Gamma_i)\right]^{-1}$ and $\beta_{ip} = \sigma_{i\beta}^2\mathrm{vec}(X_i\Gamma_i)^T\Sigma^{-1}\mathrm{vec}(\mu - 1_k\gamma_i^T)$. Given the prior distribution assigns a zero probability for negative values of β_i, the conditional posterior distribution has the form of a truncated normal distribution,

$$\pi(\beta_i|\{X_i\},\{\beta_i'\},\{\theta_{ij}\},\{\gamma_i\},\mu,\Sigma^{-1}) \sim N(\beta_{ip},\sigma_{i\beta}^2), \quad \beta_i > 0.$$

In practice, we simulate from the truncated normal using the method of Gelfand *et al.* (1992) by sampling from the unconstrained full conditional distribution, the related untruncated normal, and only retain a proposed value if it falls within the constraint region. If a proposed value is negative then it is automatically rejected and the ith scaling parameter is not changed on that iteration.

Updating the ijth rotation angle. Let θ_{ij} be the jth rotation angle of the ith shape, $i = 1,...,n$, $j = 1,...,m(m-1)/2$, and let $\{\theta_{ij}'\}$ be $\{\theta_{ij}\}$

excluding for the ijth angle, θ_{ij}, then the posterior distribution of θ_{ij} is,

$$\begin{aligned}
&\pi(\theta_{ij}|\{X_i\},\{\beta_i\},\{\theta'_{ij}\},\{\gamma_i\},\mu,\Sigma^{-1}) \\
&\quad\propto\ P(\theta_{ij})L(\{X_i\}|\{\beta_i\},\{\theta_{ij}\},\{\gamma_i\},\mu,\Sigma^{-1}), \\
&\quad\propto\ P(\theta_{ij})L(X_i|\beta_i,\theta_{i1},...,\theta_{i,m(m-1)/2},\gamma_i,\mu,\Sigma^{-1}), \\
&\quad\propto\ \exp\left\{-\frac{1}{2}\mathrm{vec}(\mu-\beta_iX_i\Gamma_i-1_k\gamma_i^T)^T\Sigma^{-1}\mathrm{vec}(\mu-\beta_iX_i\Gamma_i-1_k\gamma_i^T)\right\},
\end{aligned}$$

where Γ_i is the $m\times m$ rotation matrix based on the angles $\theta_{i1},...,\theta_{i,m(m-1)/2}$.

Due to the complicated form of this expression with respect to the individual angles, the rotation parameters are updated with a Metropolis-Hastings step, where the proposed value, $\theta^\star_{ij}$, is sampled from the distribution $N(\theta_{ij},\sigma^2_{ij})$, where σ_{ij} is a positive constant. Note that if the proposed value is outside the range $(-\pi,\pi)$ it is wrapped to be within this range. This random walk is symmetric, so $\theta^\star_{ij}$ is accepted with probability $\alpha_\theta=\min(1,r_\theta)$, where,

$$r_\theta=\frac{\pi(\theta^\star_{ij}|X_i,\beta_i,\theta_i,\gamma_i,\mu,\Sigma)}{\pi(\theta_{ij}|X_i,\beta_i,\theta_i,\gamma_i,\mu,\Sigma)}.$$

Let $\Gamma^\star_i$ be a new rotation matrix based on the angles $\theta_{i1},...\theta^\star_{ij},...,\theta_{i,m(m-1)/2}$, then the acceptance ratio is,

$$\begin{aligned}
2\log r_\theta\ &=\ \mathrm{vec}(\mu-\beta_iX_i\Gamma_i-1_k\gamma_i^T)^T\Sigma^{-1}\mathrm{vec}(\mu-\beta_iX_i\Gamma_i-1_k\gamma_i^T) \\
&\quad -\mathrm{vec}(\mu-\beta_iX_i\Gamma^\star_i-1_k\gamma_i^T)^T\Sigma^{-1}\mathrm{vec}(\mu-\beta_iX_i\Gamma^\star_i-1_k\gamma_i^T).
\end{aligned}$$

Updating the ith translation vector. Let $\{\gamma'_i\}$ be $\{\gamma_i\}$ excluding the ith translation vector, then the conditional posterior distribution of the ith translation vector is,

$$\begin{aligned}
&\pi(\gamma_i|\{X_i\},\{\beta_i\},\{\theta_{ij}\},\{\gamma'_i\},\mu,\Sigma^{-1}) \\
&\quad\propto\ P(\gamma_i)L(\{X_i\}|\{\beta_i\},\{\theta_{ij}\},\{\gamma_i\},\mu,\Sigma^{-1}),
\end{aligned}$$

$$
\begin{aligned}
&\propto \exp\left\{-\frac{1}{2}\mathrm{vec}(\mu-\beta_i X_i\Gamma_i-1_k\gamma_i^T)^T\Sigma^{-1}\mathrm{vec}(\mu-\beta_i X_i\Gamma_i-1_k\gamma_i^T)\right\},\\
&\propto \exp\Big\{-\frac{1}{2}[\gamma_i^T(I_m\otimes 1_k)^T\Sigma^{-1}(I_m\otimes 1_k)\gamma_i\\
&\quad -2\gamma_i^T(I_m\otimes 1_k)^T\Sigma^{-1}\mathrm{vec}(\mu-\beta_i X_i\Gamma_i)]\Big\},\\
&\propto \exp\left\{-\frac{1}{2}\left[(\gamma_i-\gamma_{ip})^T\Sigma_\gamma^{-1}(\gamma_i-\gamma_{ip})\right]\right\},
\end{aligned}
$$

where $\Sigma_\gamma^{-1}=(I_m\otimes 1_k)^T\Sigma^{-1}(I_m\otimes 1_k)$ and $\gamma_{ip}=\Sigma_\gamma(I_m\otimes 1_k)^T\Sigma^{-1}\mathrm{vec}(\mu-\beta_i X_i\Gamma_i)$. Therefore, we update the ith translation vector by sampling from the conditional posterior distribution,

$$\pi(\gamma_i|\{X_i\},\{\beta_i\},\{\theta_{ij}\},\{\gamma_i\},\mu,\Sigma^{-1})\sim N_m(\gamma_{ip},\Sigma_\gamma).$$

Updating the mean shape. Let,

$$
\begin{aligned}
\mathrm{vec}(\mu_p) &= (\Sigma_{\mu 0}^{-1}+n\Sigma^{-1})^{-1}\\
&\quad\times\left(\Sigma_{\mu 0}^{-1}\mathrm{vec}(\mu_0)+\sum_{i=1}^{n}\Sigma^{-1}\mathrm{vec}(\beta_i X_i\Gamma_i+1_k\gamma_i^T)\right), \quad (3.4)
\end{aligned}
$$

then the conditional posterior distribution of the mean shape is,

$$
\begin{aligned}
&\pi(\mu|\{X_i\},\{\beta_i\},\{\theta_{ij}\},\{\gamma_i\},\Sigma^{-1})\propto P(\mu)L(\{X_i\}|\{\beta_i\},\{\theta_{ij}\},\{\gamma_i\},\mu,\Sigma^{-1}),\\
&\propto \exp\left\{-\frac{1}{2}\mathrm{vec}(\mu-\mu_0)^T\Sigma_{\mu 0}^{-1}\mathrm{vec}(\mu-\mu_0)\right\}\times\\
&\quad \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\mathrm{vec}(\mu-\beta_i X_i\Gamma_i-1_k\gamma_i^T)^T\Sigma^{-1}\mathrm{vec}(\mu-\beta_i X_i\Gamma_i-1_k\gamma_i^T)\right\},\\
&\propto \exp\left\{-\frac{1}{2}\mathrm{vec}(\mu)^T(\Sigma_{\mu 0}^{-1}+n\Sigma^{-1})\mathrm{vec}(\mu)\right\}\\
&\quad \times\exp\left\{\mathrm{vec}(\mu)^T\left(\Sigma_{\mu 0}^{-1}\mathrm{vec}(\mu_0)+\sum_{i=1}^{n}\Sigma^{-1}\mathrm{vec}(\beta_i X_i\Gamma_i+1_k\gamma_i^T)\right)\right\},
\end{aligned}
$$

$$\propto \quad \exp\left\{-\frac{1}{2}\text{vec}(\mu-\mu_p)^T(\Sigma_{\mu 0}^{-1}+n\Sigma^{-1})\text{vec}(\mu-\mu_p)\right\}.$$

This has the form of the multivariate normal distribution,

$$\pi(\text{vec}(\mu)|\{X_i,\beta_i,\theta_{ij},\gamma_i\},\Sigma^{-1}) \sim N_{km}(\text{vec}(\mu_p),(\Sigma_{\mu 0}^{-1}+n\Sigma^{-1})^{-1}). \quad (3.5)$$

We remove the invariance of size, orientation and translation by transforming the distribution such that the mean is the isotropic Procrustes fit of μ_p on μ_0. Let $\widehat{\beta}_\mu$, $\widehat{\Gamma}_\mu$ and $\widehat{\gamma}_\mu$ be the isotropic Procrustes estimators of the similarity transformation mapping μ_p to μ_0, then the transformed distribution has mean, $\text{vec}(\widehat{\beta}_\mu\mu_p\widehat{\Gamma}_\mu + 1_k\widehat{\gamma}_\mu^T) = \widehat{\beta}_\mu(\widehat{\Gamma}_\mu\otimes I_k)\text{vec}(\mu_p) + (\widehat{\gamma}_\mu\otimes 1_k)$. Applying this same linear transformation to the covariance matrix, we update the mean shape by sampling from,

$$\pi(\text{vec}(\mu)|\{X_i,\beta_i,\theta_{ij},\gamma_i\},\Sigma^{-1}) \sim N_{km}\left(\text{vec}(\widehat{\beta}_\mu\mu_p\widehat{\Gamma}_\mu + 1_k\widehat{\gamma}_\mu^T),\Sigma_{\mu p}\right),$$

where,

$$\Sigma_{\mu p} = \widehat{\beta}_\mu^2(\widehat{\Gamma}_\mu\otimes I_k)(\Sigma_{\mu 0}^{-1}+n\Sigma^{-1})^{-1}(\widehat{\Gamma}_\mu\otimes I_k)^T.$$

The distribution has the same shape before and after the transformation, but the resized kernel has been rotated and translated across the parameter space.

Updating the inverse covariance matrix. The conditional posterior distribution of the inverse covariance matrix is,

$$\begin{aligned}\pi(\Sigma^{-1}|\{X_i\},\{\beta_i\},\theta_{ij}\},\{\gamma_i\},\mu) &\propto P(\Sigma^{-1})L(\{X_i\}|\{\beta_i\},\theta_{ij}\},\{\gamma_i\},\mu,\Sigma^{-1}),\\ \propto \quad & |\Sigma^{-1}|^{-\frac{1}{2}}|2\pi\Sigma|^{-\frac{n}{2}}\exp\left\{-\frac{1}{2}\text{tr}(\Sigma_0\Sigma^{-1})\right\}\\ & \times\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\text{vec}(\beta_iX_i\Gamma_i+1_k\gamma_i^T-\mu)^T\Sigma^{-1}\text{vec}(\beta_iX_i\Gamma_i+1_k\gamma_i^T-\mu)\right\},\end{aligned}$$

$$\propto \quad |\Sigma^{-1}|^{(n-1)/2} \exp\left\{-\frac{1}{2}\text{tr}\left[(\Sigma_0 + T)\Sigma^{-1}\right]\right\},$$

where,

$$T = \sum_{i=1}^{n} \text{vec}(\beta_i X_i \Gamma_i + 1_k \gamma_i^T - \mu)\text{vec}(\beta_i X_i \Gamma_i + 1_k \gamma_i^T - \mu)^T. \tag{3.6}$$

Therefore, the inverse covariance matrix is updated by sampling from the conditional posterior Wishart distribution,

$$\pi(\Sigma^{-1}|\{X_i\}, \{\beta_i\}, \{\theta_{ij}\}, \{\gamma_i\}, \mu) \sim W_{km}\left((\Sigma_0 + T)^{-1}, km + n\right). \tag{3.7}$$

3.4.3 Hybrid MCMC algorithm

Given a mean shape and covariance matrix, Chapter 2 gives the maximum likelihood estimates of the translation parameters. Therefore, an alternative to the conventional algorithm is to use an MCMC style algorithm to update μ and Σ but use CW OPA to estimate the transformation parameters at each step. This hybrid algorithm is specified in Algorithm 3.4.2, and differs from a conventional MCMC algorithm in two ways. Firstly, the shapes are registered during each step so the data (the shapes) are not static. Secondly, a generated value for the mean shape is identical to a scaled, rotated and translated version of itself, requiring the mean to be registered to a template, μ_0, to remove this invariance. In truth, these differences prevent us from demonstrating that the resulting Markov chain samples from the stationary distributions in the correct proportions as the required properties of MCMC listed in Section 1.7 are not satisfied. This, however, is of less consequence if we are merely interested in providing a point estimate of μ and Σ that maximises the posterior likelihood of the model, the maximum *a posteriori* (MAP) estimate.

Algorithm 3.4.2 ***A hybrid MCMC algorithm for shape data***

1. *Produce initial samples μ and Σ based on prior distributions, and let $t = 0$.*

2. *Register shapes using CW OPA to μ_t given Σ_t.*

3. *Obtain conditional posterior distribution of μ, given Σ_t and the data.*

4. *Generate value for μ_{t+1} from the conditional distribution in step (3).*

5. *Register shapes using CW OPA to μ_{t+1} given Σ_t.*

6. *Obtain conditional posterior distribution of Σ, given μ_{t+1} and the data.*

7. *Generate value for Σ_{t+1} from the conditional distribution in step (6).*

8. *Increment t.*

9. *Repeat steps (2)-(8).*

New values for μ and Σ could be generated by simply sampling from their posterior distributions, given in Section 3.4.2. However this Gibbs-step approach, does not consider how well the shapes could be registered using the new parameters, given the shapes are optimally matched using the current estimates. Instead, with a Metropolis-Hastings step, transformation parameters can be calculated by registering the shapes using both the current and proposed μ and Σ, and the whole model can be evaluated under both scenarios, and the parameters updated in one block. The proposed mean and covariance matrix could be sampled from a random walk, but the parameters' posterior distributions are more informative of suitable values. Therefore, we take the posterior distribution to be the sampling distribution.

Updating the mean shape. Given a mean shape, μ, and a proposed new mean shape, $\mu^\star$, sampled from the proposal distribution $q(\mu^\star|\mu)$, this new mean shape is accepted with probability $\alpha_\mu(\mu, \mu^\star) = \min(1, r_\mu)$, where,

$$r_\mu = \left(\frac{\pi(\mu^\star, \{\widehat{\beta}_i^\star\}, \{\widehat{\Gamma}_i^\star\}, \{\widehat{\gamma}_i^\star\}|\{X_i\}, \Sigma^{-1}) q(\mu|\mu^\star)}{\pi(\mu, \{\widehat{\beta}_i\}, \{\widehat{\Gamma}_i\}, \{\widehat{\gamma}_i\}|\{X_i\}, \Sigma^{-1}) q(\mu^\star|\mu)} \right),$$

and X_i are the initial shapes and $\{\widehat{\beta}_i\}$,$\{\widehat{\Gamma}_i\}$,$\{\widehat{\gamma}_i\}$ and $\{\widehat{\beta}_i^\star\}$, $\{\widehat{\Gamma}_i^\star\}$,$\{\widehat{\gamma}_i^\star\}$, are the transformation parameters to register the shapes to μ and $\mu^\star$ respectively, both using Σ. If the proposal distribution is $q(\mu^\star|\mu) \sim N_{km}(\mu, \Sigma_q)$, with Σ_q fixed, then it generates a random-walk proposal with $q(\mu^\star|\mu) = q(\mu|\mu^\star)$ cancelling, so the ratio is,

$$
\begin{aligned}
2\log r_\mu = &\\
&\mathrm{vec}(\mu-\mu_0)^T\Sigma_{\mu 0}^{-1}\mathrm{vec}(\mu-\mu_0) - \mathrm{vec}(\mu^\star-\mu_0)^T\Sigma_{\mu 0}^{-1}\mathrm{vec}(\mu^\star-\mu_0)\\
&+\sum_{i=1}^{n}\mathrm{vec}\left(\mu-\widehat{\beta}_i X_i\widehat{\Gamma}_i - 1_k\widehat{\gamma}_i^T\right)^T\Sigma^{-1}\mathrm{vec}\left(\mu-\beta_i X_i\widehat{\Gamma}_i - 1_k\widehat{\gamma}_i^T\right)\\
&-\sum_{i=1}^{n}\mathrm{vec}\left(\mu^\star-\widehat{\beta}_i^\star X_i\widehat{\Gamma}_i^\star - 1_k\widehat{\gamma}_i^{\star T}\right)^T\Sigma^{-1}\mathrm{vec}\left(\mu^\star-\widehat{\beta}_i^\star X_i\widehat{\Gamma}_i^\star - 1_k\widehat{\gamma}_i^{\star T}\right).
\end{aligned}
$$

Letting the posterior distribution be the proposal distribution gives,

$$
q(\mu^\star|\mu) \sim N_{km}\left(\mu_p(\{X_i,\widehat{\beta}_i,\widehat{\Gamma}_i,\widehat{\gamma}_i\}),(\Sigma_{\mu 0}^{-1}+n\Sigma^{-1})^{-1}\right),
$$

where μ_p is defined in Equation (3.4). The mean is denoted $\mu_p(\{X_i,\widehat{\beta}_i,\widehat{\Gamma}_i,\widehat{\gamma}_i\})$ as a reminder that μ_p is a function of the initial shapes and the current estimates of the rigid-body transformation parameters. The proposed value, $\mu^\star$, is sampled from this distribution and matched to μ_0 using isotropic OPA. This proposal is no longer symmetric, as $q(\mu|\mu^\star) \sim N_{km}(\mu_p(\{X_i,\widehat{\beta}_i^\star,\widehat{\Gamma}_i^\star,\widehat{\gamma}_i^\star\}),(\Sigma_{\mu 0}^{-1}+n\Sigma^{-1})^{-1})$, and so the probability of going from $\mu_p(\{X_i,\widehat{\beta}_i,\widehat{\Gamma}_i,\widehat{\gamma}_i\})$ to $\mu^\star$ is different to the probability of going from $\mu_p(\{X_i,\widehat{\beta}_i^\star,\widehat{\Gamma}_i^\star,\widehat{\gamma}_i^\star\})$ to μ. Therefore, the proposal distribution must be included in the acceptance ratio,

$$
\begin{aligned}
2\log r_\mu = &\\
&\mathrm{vec}(\mu-\mu_0)^T\Sigma_{\mu 0}^{-1}\mathrm{vec}(\mu-\mu_0) - \mathrm{vec}(\mu^\star-\mu_0)^T\Sigma_{\mu 0}^{-1}\mathrm{vec}(\mu^\star-\mu_0)\\
&+\sum_{i=1}^{n}\mathrm{vec}\left(\mu-\widehat{\beta}_i X_i\widehat{\Gamma}_i - 1_k\widehat{\gamma}_i^T\right)^T\Sigma^{-1}\mathrm{vec}\left(\mu-\widehat{\beta}_i X_i\widehat{\Gamma}_i - 1_k\widehat{\gamma}_i^T\right)
\end{aligned}
$$

$$
\begin{aligned}
&+\text{vec}\left(\mu^\star - \mu_p(\{X_i, \widehat{\beta}_i, \widehat{\Gamma}_i, \widehat{\gamma}_i\})\right)^T (\Sigma_{\mu 0}^{-1} + n\Sigma^{-1}) \\
&\quad \times \text{vec}\left(\mu^\star - \mu_p(\{X_i, \widehat{\beta}_i, \widehat{\Gamma}_i, \widehat{\gamma}_i\})\right) \\
&- \sum_{i=1}^{n} \text{vec}\left(\mu^\star - \widehat{\beta}_i^\star X_i \widehat{\Gamma}_i^\star - 1_k \widehat{\gamma}_i^{\star T}\right)^T \Sigma^{-1} \text{vec}\left(\mu^\star - \widehat{\beta}_i^\star X_i \widehat{\Gamma}_i^\star - 1_k \widehat{\gamma}_i^{\star T}\right) \\
&- \text{vec}\left(\mu - \mu_p(\{X_i^\star, \widehat{\beta}_i^\star, \widehat{\Gamma}_i^\star, \widehat{\gamma}_i^\star\})\right)^T (\Sigma_{\mu 0}^{-1} + n\Sigma^{-1}) \\
&\quad \times \text{vec}\left(\mu - \mu_p(\{X_i^\star, \widehat{\beta}_i^\star, \widehat{\Gamma}_i^\star, \widehat{\gamma}_i^\star\})\right).
\end{aligned}
$$

Updating the inverse covariance matrix. In a similar fashion, the covariance matrix can be updated using a proposed value, $\Sigma^\star$, which allows for the shapes to be matched using the proposed value as well as the current value, and a more genuine comparison made. The proposed matrix is accepted with probability $\alpha_\Sigma(\Sigma^{-1}, \Sigma^{\star -1}) = \min(1, r_\Sigma)$, where,

$$
r_\Sigma = \left(\frac{\pi(\Sigma^{\star -1}, \{\widehat{\beta}_i^\star\}, \{\widehat{\Gamma}_i^\star\}, \{\widehat{\gamma}_i^\star\} | \{X_i\}, \mu) q(\Sigma^{-1} | \Sigma^{\star -1})}{\pi(\Sigma^{-1}, \{\widehat{\beta}_i\}, \{\widehat{\Gamma}_i\}, \{\widehat{\gamma}_i\} | \{X_i\}, \mu) q(\Sigma^{\star -1} | \Sigma^{-1})} \right).
$$

If the proposal distribution is taken to be the posterior distribution, then,

$$
q(\Sigma^{\star -1} | \Sigma^{-1}) \sim W_{km}\left([\Sigma_p(\{X_i, \widehat{\beta}_i, \widehat{\Gamma}_i, \widehat{\gamma}_i\})]^{-1}, km + n \right),
$$

where $\Sigma_p = \Sigma_0 + T$ and is a function of the initial shapes and the current estimates of the transformation parameters, as implied by Equation (3.6). Likewise, $q(\Sigma^{-1} | \Sigma^{\star -1}) \sim W_{km}\left([\Sigma_p(\{X_i, \widehat{\beta}_i^\star, \widehat{\Gamma}_i^\star, \widehat{\gamma}_i^\star\})]^{-1}, km + n \right)$. Therefore, the acceptance ratio is calculated as,

$$
\begin{aligned}
&2 \log r_\Sigma = \\
&\quad (km + n) \left(\log \left| \Sigma_p(\{X_i^\star, \widehat{\beta}_i^\star, \widehat{\Gamma}_i^\star, \widehat{\gamma}_i^\star\}) \right| - \log \left| \Sigma_p(\{X_i, \widehat{\beta}_i, \widehat{\Gamma}_i, \widehat{\gamma}_i\}) \right| \right) \\
&\quad + \text{tr}(\Sigma_0 \Sigma^{-1}) + \sum_{i=1}^{n} (\widehat{\beta}_i X_i \widehat{\Gamma}_i + 1_k \widehat{\gamma}_i^T - \mu)^T \Sigma^{-1} (\widehat{\beta}_i X_i \widehat{\Gamma}_i - 1_k \widehat{\gamma}_i^T - \mu)
\end{aligned}
$$

$$-\text{tr}(\Sigma_0\Sigma^{\star -1}) - \sum_{i=1}^{n}(\widehat{\beta}_i^{\star} X_i \widehat{\Gamma}_i^{\star} - 1_k \widehat{\gamma}_i^{\star^T} - \mu)^T \Sigma^{\star -1}(\widehat{\beta}_i^{\star} X_i \widehat{\Gamma}_i^{\star} - 1_k \widehat{\gamma}_i^{\star^T} - \mu)$$
$$+\text{tr}\left(\Sigma_p(\{X_i, \widehat{\beta}_i, \widehat{\Gamma}_i, \widehat{\gamma}_i\})\Sigma^{\star -1}\right) - \text{tr}\left(\Sigma_p(\{X_i^{\star}, \widehat{\beta}_i^{\star}, \widehat{\Gamma}_i^{\star}, \widehat{\gamma}_i^{\star}\})\Sigma^{-1}\right).$$

3.5 Comparison of the two MCMC algorithms

We compare the two MCMC algorithms, specified in Algorithms 3.4.2 and 3.4.1, by applying them to simulated data. The first, conventional, algorithm uses MCMC to update all the model parameters. The second, hybrid, algorithm uses CW OPA to estimate the transformation parameters and Metropolis-Hastings steps to update the mean and covariance matrix.

3.5.1 Mean and maximum likelihood estimates

The algorithms are applied to a data set consisting of $n = 50$ shapes simulated from the multivariate normal distribution $\text{vec}(X) \sim N_{km}(\text{vec}(\mu_C), \Sigma_C)$, where μ_C and Σ_C are given in example 3 of the simulation study in Section 3.3. No scaling was applied to the shapes in the MCMC algorithms to be consistent with the simulation study. The prior distributions for the mean shape and covariance matrix are given by Equation (3.3), with the parameters,

$$\begin{aligned}
\mu_0 &= \begin{bmatrix} 0 & 5 \\ 40 & 0 \\ 0 & -5 \\ -40 & 0 \end{bmatrix}, \\
\Sigma_{\mu 0} &= \text{diag}(0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1), \\
\Sigma_0^{-1} &= \text{diag}(12500, 12.5, 12500, 12.5, 12.5, 0.0125, 12.5, 0.0125).
\end{aligned} \tag{3.8}$$

The prior distribution for the mean shape is centred on the true value, with low variability, because the mean shape can be estimated using isotropic Procrustes with a reasonable amount of certainty. The prior distribution for the covariance matrix assigns an amount of variability to each co-ordinate which approximately reflects the true distribution, but without the correlations.

Each algorithm ran for $N = 10,000$ iterations, and during each iteration every parameter was updated and recorded and the log-likelihood calculated. For each of the algorithms, two estimates of the model parameters, μ, Σ and Σ^* are given, where Σ^* is the projected covariance matrix given by Equation (3.2). The first estimates, $\bar{\mu}$, $\bar{\Sigma}$, $\bar{\Sigma}^*$, are the means of the parameters over the second half of the iterations. The first half is discarded as a burn-in period. The second estimates, $\widehat{\mu}$, $\widehat{\Sigma}$, $\widehat{\Sigma}^*$, are the parameter values at the iteration with the highest likelihood for the model, the maximum likelihood (ML) estimate. Note that we can expect the ML estimate to be close to the maximum *a posteriori* (MAP) estimate due to the comparatively uninformative priors used. Results for each algorithm are given below and summarised in Table 3.4. Note that entries are rounded to three decimal places. The shapes registered to $\bar{\mu}$, using $\bar{\Sigma}$, for both algorithms, are plotted in Figure 3.8. MCMC traces for some of the model parameters are plotted in Figure 3.9, conventional algorithm, and Figure 3.10, hybrid algorithm. The results are discussed in Section 3.5.3.

	Conventional	Hybrid
$\|\bar{\mu} - \mu_C\|^2$	0.004	0.004
$\|\widehat{\mu} - \mu_C\|^2$	0.164	0.027
$\|\bar{\Sigma} - \Sigma_C\|^2$	35.529	42.984
$\|\widehat{\Sigma} - \Sigma_C\|^2$	14.887	19.979
$\|\bar{\Sigma}^* - \Sigma_C^*\|^2$	0.703	0.699
$\|\widehat{\Sigma}^* - \Sigma_C^*\|^2$	0.983	0.502

Table 3.4: The distance of the mean and ML estimates from the true parameters for both algorithms.

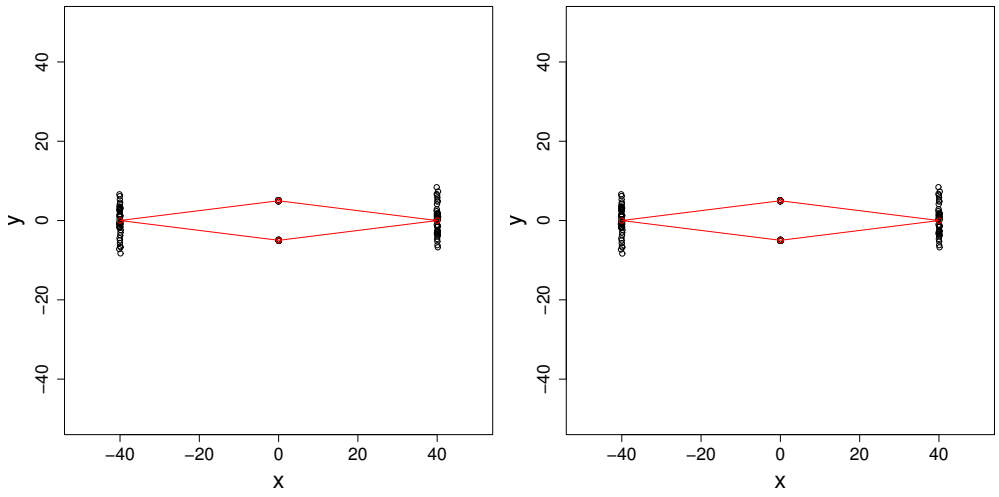

Figure 3.8: The shapes registered to $\bar{\mu}$ using $\bar{\Sigma}$, using the mean estimates of the transformation parameters following the conventional algorithm (left) and CW OPA following the hybrid algorithm (right).

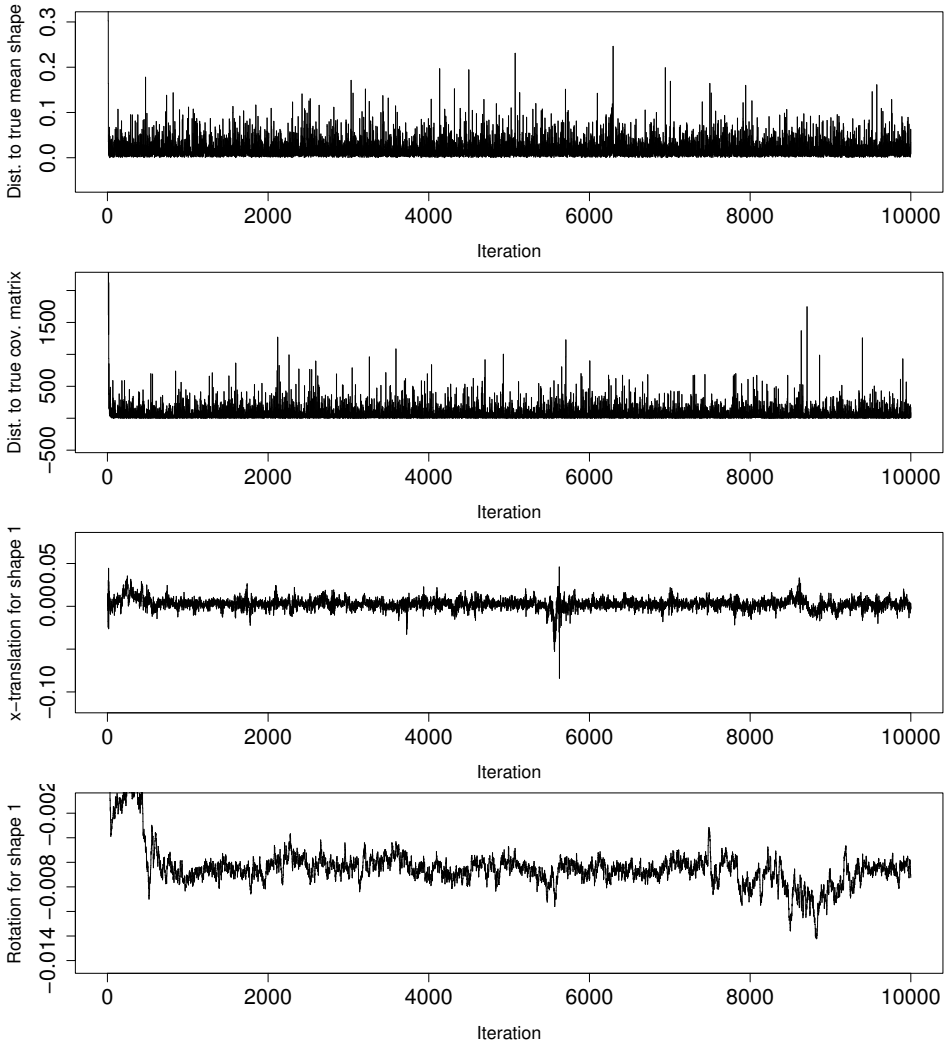

Figure 3.9: For the conventional algorithm, the squared Euclidean norm distance of the mean shape (top) and covariance matrix (second from top) from the true model parameters, and the x translation parameter (next to bottom) and rotation parameter (bottom) for one of the shapes, at each iteration.

Conventional algorithm: mean estimates

$$\bar{\mu} = \begin{bmatrix} -0.033 & 4.989 \\ 40.027 & -0.003 \\ 0.033 & -4.988 \\ -40.027 & 0.007 \end{bmatrix},$$

$$\bar{\Sigma} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0.004 & 0 & -0.004 \\ 0 & 0.011 & 0 & -0.009 & 0.002 & -0.069 & 0.002 & 0.07 \\ 0 & 0 & 0 & 0 & 0 & -0.006 & 0 & 0.006 \\ 0 & -0.009 & 0 & 0.01 & -0.002 & 0.069 & -0.001 & -0.069 \\ 0 & 0.002 & 0 & -0.002 & 0.034 & 0.074 & 0.024 & -0.019 \\ 0.004 & -0.069 & -0.006 & 0.069 & 0.074 & 13.924 & 0.045 & -11.98 \\ 0 & 0.002 & 0 & -0.001 & 0.024 & 0.045 & 0.034 & 0.012 \\ -0.004 & 0.07 & 0.006 & -0.069 & -0.019 & -11.98 & 0.012 & 13.497 \end{bmatrix},$$

$$\bar{\Sigma}^* = \begin{bmatrix} 0.195 & -0.009 & -0.194 & 0.009 & -0.002 & 0.029 & -0.008 & -0.02 \\ -0.009 & 0.01 & 0.008 & -0.01 & 0.001 & -0.002 & 0.001 & 0 \\ -0.194 & 0.008 & 0.195 & -0.009 & 0.002 & -0.029 & 0.008 & 0.02 \\ 0.009 & -0.01 & -0.009 & 0.01 & -0.001 & 0.002 & -0.001 & 0 \\ -0.002 & 0.001 & 0.002 & -0.001 & 0.215 & -0.21 & 0.205 & -0.21 \\ 0.029 & -0.002 & -0.029 & 0.002 & -0.21 & 0.214 & -0.21 & 0.207 \\ -0.008 & 0.001 & 0.008 & -0.001 & 0.205 & -0.21 & 0.214 & -0.208 \\ -0.02 & 0 & 0.02 & 0 & -0.21 & 0.207 & -0.208 & 0.212 \end{bmatrix}.$$

Conventional algorithm: ML estimates

$$\widehat{\mu} = \begin{bmatrix} -0.034 & 4.996 \\ 40.029 & -0.33 \\ 0.034 & -4.986 \\ -40.033 & 0.224 \end{bmatrix},$$

$$\widehat{\Sigma} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0.004 & 0 & -0.004 \\ 0 & 0.01 & 0 & -0.009 & 0.001 & 0.005 & 0.001 & -0.022 \\ 0 & 0 & 0 & 0 & 0 & -0.011 & 0 & 0.01 \\ 0 & -0.009 & 0 & 0.011 & -0.001 & 0.02 & 0.001 & 0.01 \\ 0 & 0.001 & 0 & -0.001 & 0.011 & 0.063 & 0.001 & -0.08 \\ 0.004 & 0.005 & -0.011 & 0.02 & 0.063 & 13.582 & 0.036 & -10.473 \\ 0 & 0.001 & 0 & 0.001 & 0.001 & 0.036 & 0.005 & -0.037 \\ -0.004 & -0.022 & 0.01 & 0.01 & -0.08 & -10.473 & -0.037 & 11.257 \end{bmatrix},$$

$$\widehat{\Sigma}^* = \begin{bmatrix} 0.173 & -0.01 & -0.174 & 0.008 & -0.028 & 0.053 & -0.035 & 0.01 \\ -0.01 & 0.01 & 0.01 & -0.01 & 0.007 & -0.009 & 0.007 & -0.006 \\ -0.174 & 0.01 & 0.176 & -0.008 & 0.03 & -0.055 & 0.036 & -0.011 \\ 0.008 & -0.01 & -0.008 & 0.011 & -0.009 & 0.009 & -0.008 & 0.007 \\ -0.028 & 0.007 & 0.03 & -0.009 & 0.255 & -0.253 & 0.243 & -0.246 \\ 0.053 & -0.009 & -0.055 & 0.009 & -0.253 & 0.258 & -0.249 & 0.244 \\ -0.035 & 0.007 & 0.036 & -0.008 & 0.243 & -0.249 & 0.245 & -0.239 \\ 0.01 & -0.006 & -0.011 & 0.007 & -0.246 & 0.244 & -0.239 & 0.241 \end{bmatrix}.$$

Hybrid algorithm: mean estimates

$$\bar{\mu} = \begin{bmatrix} -0.033 & 4.988 \\ 40.027 & -0.004 \\ 0.033 & -4.988 \\ -40.027 & 0.004 \end{bmatrix},$$

$$\bar{\Sigma} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.012 & 0 & -0.011 & 0.001 & -0.076 & 0 & 0.073 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -0.011 & 0 & 0.012 & -0.001 & 0.075 & 0 & -0.073 \\ 0 & 0.001 & 0 & -0.001 & 0.006 & 0.004 & -0.003 & -0.004 \\ 0 & -0.076 & 0 & 0.075 & 0.004 & 14.094 & -0.024 & -12.332 \\ 0 & 0 & 0 & 0 & -0.003 & -0.024 & 0.007 & 0.026 \\ 0 & 0.073 & 0 & -0.073 & -0.004 & -12.332 & 0.026 & 13.911 \end{bmatrix},$$

$$\bar{\Sigma}^* = \begin{bmatrix} 0.199 & -0.01 & -0.198 & 0.009 & 0 & 0.028 & -0.007 & -0.021 \\ -0.01 & 0.012 & 0.009 & -0.011 & 0.001 & -0.002 & 0.001 & 0 \\ -0.198 & 0.009 & 0.198 & -0.01 & 0 & -0.028 & 0.007 & 0.021 \\ 0.009 & -0.011 & -0.01 & 0.012 & -0.001 & 0.002 & -0.001 & 0 \\ 0 & 0.001 & 0 & -0.001 & 0.214 & -0.209 & 0.204 & -0.209 \\ 0.028 & -0.002 & -0.028 & 0.002 & -0.209 & 0.213 & -0.21 & 0.206 \\ -0.007 & 0.001 & 0.007 & -0.001 & 0.204 & -0.21 & 0.214 & -0.208 \\ -0.021 & 0 & 0.021 & 0 & -0.209 & 0.206 & -0.208 & 0.211 \end{bmatrix}.$$

Hybrid algorithm: ML estimates

$$\widehat{\mu} = \begin{bmatrix} -0.08 & 5.046 \\ 40.043 & -0.06 \\ 0.08 & -4.945 \\ -40.043 & -0.04 \end{bmatrix},$$

$$\widehat{\Sigma} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0.001 & 0 & -0.001 \\ 0 & 0.012 & 0 & -0.01 & 0.004 & -0.157 & 0 & 0.174 \\ 0 & 0 & 0 & 0 & 0 & -0.003 & 0 & 0.002 \\ 0 & -0.01 & 0 & 0.01 & -0.004 & 0.153 & 0 & -0.159 \\ 0 & 0.004 & 0 & -0.004 & 0.006 & -0.151 & -0.003 & 0.136 \\ 0.001 & -0.157 & -0.003 & 0.153 & -0.151 & 12.664 & -0.056 & -11.382 \\ 0 & 0 & 0 & 0 & -0.003 & -0.056 & 0.005 & 0.086 \\ -0.001 & 0.174 & 0.002 & -0.159 & 0.136 & -11.382 & 0.086 & 12.958 \end{bmatrix},$$

$$\widehat{\Sigma}^* = \begin{bmatrix} 0.187 & -0.023 & -0.183 & 0.018 & -0.003 & 0.025 & 0 & -0.022 \\ -0.023 & 0.012 & 0.022 & -0.01 & -0.001 & 0 & -0.004 & 0.005 \\ -0.183 & 0.022 & 0.179 & -0.018 & 0.004 & -0.025 & 0.001 & 0.02 \\ 0.018 & -0.01 & -0.018 & 0.01 & 0 & 0.001 & 0.003 & -0.004 \\ -0.003 & -0.001 & 0.004 & 0 & 0.193 & -0.183 & 0.173 & -0.182 \\ 0.025 & 0 & -0.025 & 0.001 & -0.183 & 0.181 & -0.171 & 0.174 \\ 0 & -0.004 & 0.001 & 0.003 & 0.173 & -0.171 & 0.17 & -0.171 \\ -0.022 & 0.005 & 0.02 & -0.004 & -0.182 & 0.174 & -0.171 & 0.179 \end{bmatrix}.$$

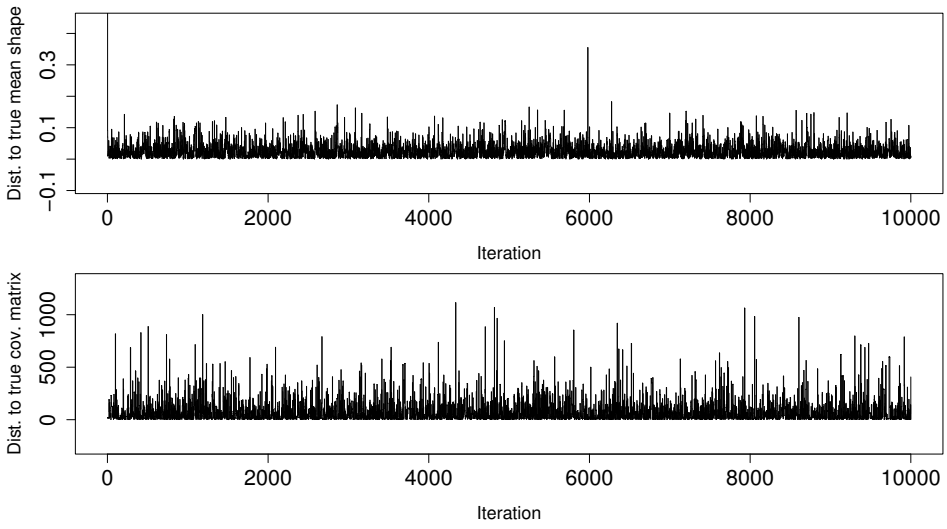

Figure 3.10: The squared Euclidean norm distance of the mean shape (top) and covariance matrix (bottom) from the true model parameters at each iteration for the hybrid algorithm.

3.5.2 Prior distribution sensitivity

The prior distributions chosen for this simulation have expected values of the mean shape and covariance matrix close to the true values. Practically, it is realistic to specify a prior for the mean shape close to the true value following analysis of isotropic GPA, for example. However, specifying a prior distribution for the covariance matrix is less straightforward, so we now conduct a sensitivity analysis on Σ_0, to see its effect in the hybrid algorithm. Given $\Sigma^{-1} \sim W_{km}(\Sigma_0^{-1}, km)$, different values of Σ_0 are chosen to shrink the expected value of Σ towards the identity matrix. 7 different values were selected such that,

$$\begin{aligned}
\text{Prior 1: } E(\Sigma) &= \text{diag}(0.00001, 0.01, 0.00001, 0.01, 0.01, 10, 0.01, 10), \\
\text{Prior 2: } E(\Sigma) &= \text{diag}(0.01, 0.02, 0.01, 0.02, 0.02, 10, 0.02, 10), \\
\text{Prior 3: } E(\Sigma) &= \text{diag}(0.1, 0.1, 0.1, 0.1, 0.1, 9.7, 0.1, 9.7),
\end{aligned}$$

$$\begin{aligned}
\text{Prior 4: } E(\Sigma) &= \text{diag}(0.25, 0.25, 0.25, 0.25, 0.25, 9.25, 0.25, 9.25), \\
\text{Prior 5: } E(\Sigma) &= \text{diag}(0.5, 0.5, 0.5, 0.5, 0.5, 8.5, 0.5, 8.5), \\
\text{Prior 6: } E(\Sigma) &= \text{diag}(1, 1, 1, 1, 1, 7, 1, 7), \\
\text{Prior 7: } E(\Sigma) &= \text{diag}(2.5, 2.5, 2.5, 2.5, 2.5, 2.5, 2.5, 2.5).
\end{aligned}$$

The 50 shapes simulated from $\text{vec}(X) \sim N_{km}(\text{vec}(\mu_C), \Sigma_C)$, were initially registered using isotropic GPA, and then the hybrid MCMC algorithm applied with the 7 different priors, each for $N = 5000$ iterations. The distances of the estimated mean shape and covariance matrix from the true parameters are given in Table 3.5. The shapes registered to $\bar{\mu}$, using $\bar{\Sigma}$, for each of the simulations are plotted in Figure 3.11.

Prior	1	2	3	4	5	6	7
$\|\bar{\mu} - \mu_C\|^2$	0.00	0.22	0.01	0.03	0.05	0.06	0.05
$\|\widehat{\mu} - \mu_C\|^2$	0.01	0.01	0.03	0.04	0.05	0.09	0.09
$\|\bar{\Sigma} - \Sigma_C\|^2$	43.25	49.48	3.44	102.56	284.07	344.35	384.95
$\|\widehat{\Sigma} - \Sigma_C\|^2$	38.61	20.42	8.48	160.39	337.80	353.16	383.07
$\|\bar{\Sigma}^* - \Sigma_C^*\|^2$	0.70	0.42	0.68	0.65	0.63	0.62	1.00
$\|\widehat{\Sigma}^* - \Sigma_C^*\|^2$	0.27	0.38	0.30	0.13	0.44	0.48	0.67

Table 3.5: The squared Euclidean distance of the mean and ML estimates from the true model parameters for different prior distributions.

3.5.3 Discussion

Based on one simulation for both algorithms, we see less error in the estimates of the mean shape and projected covariance matrix for the hybrid algorithm and more error in the estimation of the true covariance matrix. The maximum likelihood estimates are obviously dependent on the algorithm sampling a covariance matrix close to the true value, so the performance of each algorithm should not be judged on this statistic. Based on the mean estimates, both algorithms have similar errors and take similar

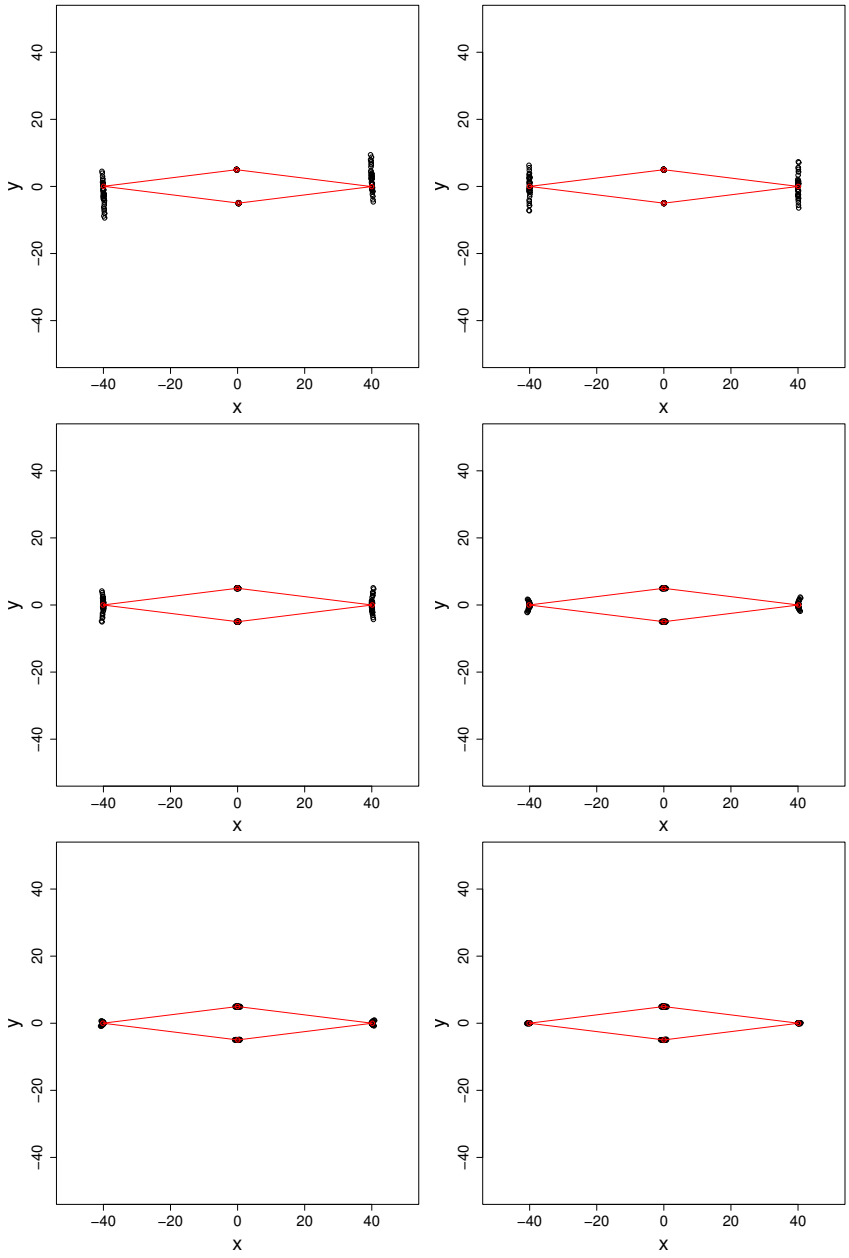

Figure 3.11: The shapes registered to the mean estimate of the mean shape using the mean estimate of the covariance matrix, following the use of priors 2 (top left), 3 (top right), 4 (middle left), 5 (middle right), 6 (bottom left) and 7 (bottom right).

lengths of time to run for one iteration. However, the hybrid algorithm converges quicker when large rotation angles are involved as the random walk of the rotation angle shown in Figure 3.9 takes 1000 iterations to converge. The mean shape and covariance matrix mix well for both algorithms.

The MCMC algorithms are obviously dependent on the choice of prior distributions. The errors in the estimates of the mean shape and projected covariance matrix do not increase substantially as the prior on Σ becomes less accurate. However, the error in the estimate of the true covariance matrix is sensitive to the prior, making the MCMC methods just as dependent on *a priori* information as the maximum likelihood methods. As the prior tends towards isotropy, the registration of the shapes tends towards the output of isotropic GPA, as expected. However, the prior does not need to be as extreme, see prior 3 in Table 3.5, as the true covariance model for the posterior distribution to have an expected value close to the true value.

3.6 Bayesian method applied to missing data

Previously, it was assumed that all the shapes in the data set have the same number of landmarks. However, in some brain imaging applications, a portion of the landmarks may be missing in each image due to poor image quality or diseased tissue, such as brain lesions. In this section, we apply our Bayesian methods to the case where the number of landmarks in each shape differs. Data sets with this characteristic are commonly found in biological morphometry, where parts of structures may be missing, or cheminformatics, where molecules with differing number of atoms are considered.

3.6.1 Model and algorithm

Suppose there exists a set of shapes, $X_1, ..., X_n$, where the number of landmarks is allowed to vary, but there still exists an underlying model with a mean shape, μ $(k \times m)$, such that the jth landmark of shape X_i corre-

sponds to a landmark in μ for all $i = 1, ..., n$, $j = 1, ..., k_i$, where k_i is the number of landmarks present in the ith shape. Denote the missing landmarks of the ith shape, Y_i, and the composite shape of both actual and estimated landmarks, Z_i. Therefore, the jth landmark of Z_i corresponds to the jth landmark of μ. The full shapes after Procrustes registration, Z_i, will continue to be modelled with the distribution,

$$\text{vec}(Z_i) \sim N_{km}(\text{vec}(\mu), \Sigma).$$

The prior distribution for the missing data in each shape is a multivariate normal distribution,

$$\text{vec}(Y_i) \sim N_{(k-k_i)m}(\text{vec}(\mu_{0i}), \Sigma_{0i}),$$

where μ_{0i} and Σ_{0i} are composed of the submatrices of μ_0 and Σ_0, as defined in Equation (3.3), that correspond to the $k - k_i$ missing landmarks.

The full likelihood of the model is,

$$L(\{X_i\}|\{Y_i\}, \{\beta_i\}, \{\theta_{ij}\}, \{\gamma_i\}, \mu, \Sigma) = |2\pi\Sigma|^{-\frac{n}{2}} \times \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} \text{vec}(\beta_i Z_i \Gamma_i + 1_k \gamma_i^T - \mu)^T \Sigma^{-1} \text{vec}(\beta_i Z_i \Gamma_i + 1_k \gamma_i^T - \mu)\right\}.$$

The optimum transformation parameters can be determined, either by use of CW OPA or MCMC. Further, the estimates can either be based only on actual landmarks, or on the combined set of actual and estimated landmarks. The posterior distributions for the mean shape and covariance matrix continue to be defined as in Equations (3.5) and (3.7) (with X's replaced by Z's and being conditional on the estimated data). The model parameters and the missing data can be estimated using the hybrid algorithm, Algorithm 3.4.2, with the additional step of sampling each missing landmark from its posterior distribution at every iteration.

Updating the missing data. The missing data for each shape is

updated by sampling from its posterior distribution, a Gibbs step. For notational simplicity, re-order the rows (and columns) of Z_i, μ and Σ such that the observed and missing data are stacked separately,

$$Z_i = \begin{bmatrix} X_i \\ Y_i \end{bmatrix}, \qquad \mu = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \qquad \Sigma^{-1} = \begin{bmatrix} \Omega_X & \Omega_{XY} \\ \Omega_{XY}^T & \Omega_Y \end{bmatrix},$$

and for notational brevity let $\widetilde{A}_i = \beta_i A_i \Gamma_i + 1_k \gamma_i^T$, for $A = X, Y, Z$.

Let $\{Y_i'\}$ be the missing data from all shapes excluding the ith shape, then the conditional posterior distribution of the ith shape is,

$$\begin{aligned}
\pi(Y_i|\{X_i\},\{Y_i'\},\{\beta_i\},\{\theta_{ij}\},\{\gamma_i\},\mu,\Sigma) &\propto P(Y_i)L(X_i|Y_i,\beta_i,\theta_i,\gamma_i,\mu,\Sigma), \\
&\propto \exp\left\{-\frac{1}{2}\mathrm{vec}(\widetilde{Y}_i-\mu_{0i})^T\Sigma_{0i}^{-1}\mathrm{vec}(\widetilde{Y}_i-\mu_{0i})\right\} \\
&\quad \times \exp\left\{-\frac{1}{2}\mathrm{vec}(\widetilde{Z}_i-\mu)^T\Sigma^{-1}\mathrm{vec}(\widetilde{Z}_i-\mu)\right\}, \\
&\propto \exp\left\{-\frac{1}{2}\Big[\mathrm{vec}(\widetilde{Y}_i)^T(\Sigma_{0i}^{-1}+\Omega_Y)\mathrm{vec}(\widetilde{Y}_i)\right. \\
&\quad \left.-2\mathrm{vec}(\widetilde{Y}_i)^T\left(\Sigma_{0i}^{-1}\mathrm{vec}(\mu_{0i})+\Omega_Y\mathrm{vec}(\mu_Y)+\Omega_{XY}^T\mathrm{vec}(\mu_X-\widetilde{X}_i)\right)\Big]\right\}, \\
&\propto \exp\left\{-\frac{1}{2}\mathrm{vec}(\widetilde{Y}_i-\mu_G)^T(\Sigma_0^{-1}+\Omega_Y)\mathrm{vec}(\widetilde{Y}_i-\mu_G)\right\},
\end{aligned}$$

where, $\mu_G = (\Sigma_0^{-1}+\Omega_Y)^{-1}\left[\Sigma_{0i}^{-1}\mathrm{vec}(\mu_{0i})+\Omega_Y\mathrm{vec}(\mu_Y)+\Omega_{XY}^T\mathrm{vec}(\mu_X-\widetilde{X}_i)\right]$. So,

$$\pi(Y_i|\{X_i\},\{Y_i'\},\{\beta_i\},\{\theta_{ij}\},\{\gamma_i\},\mu,\Sigma) \sim N_{(k-k_i)m}(\mu_G,(\Sigma_{0i}^{-1}+\Omega_Y)^{-1}).$$

3.6.2 Simulated data

We simulate $n = 50$ shapes from the multivariate normal distribution $\mathrm{vec}(X) \sim N_{km}(\mathrm{vec}(\mu_C),\Sigma_C)$, where μ_C and Σ_C are given in example 3 of the simulation study in Section 3.3. The prior distributions for the mean

shape and covariance matrix are given by Equation (3.3), with the parameters given in Equation (3.8). At random, 7 shapes were selected to have landmark two removed and 5 shapes had landmark four removed. The hybrid algorithm, Algorithm 3.4.2, was applied with the additional step of sampling each missing landmark from its posterior distribution at every iteration. Only the remaining landmarks were used to estimate the transformation parameters to minimise the impact of any incorrectly estimated data. The algorithm was run for $N = 10,000$ iterations. The mean and ML estimates of the mean shape and covariance matrix are given below, with a summary in Table 3.6. The first half of the iterations were discarded as the burn-in period. The shapes, with the missing landmarks in their mean position, registered to $\bar{\mu}$, using $\bar{\Sigma}$ are plotted in Figure 3.12.

	Hybrid
$\|\bar{\mu} - \mu_C\|^2$	0.005
$\|\widehat{\mu} - \mu_C\|^2$	0.033
$\|\bar{\Sigma} - \Sigma_C\|^2$	29.657
$\|\widehat{\Sigma} - \Sigma_C\|^2$	18.851
$\|\bar{\Sigma}^* - \Sigma_C^*\|^2$	1.197
$\|\widehat{\Sigma}^* - \Sigma_C^*\|^2$	0.278

Table 3.6: The squared Euclidean distances of the mean and ML estimates from the true model parameters.

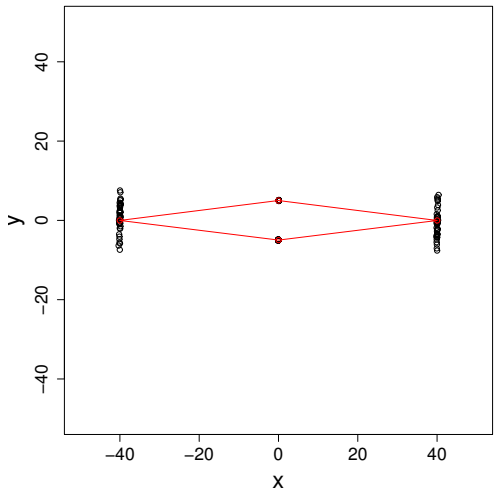

Figure 3.12: The shapes, with the missing landmarks in their mean position, registered to the mean estimate of the mean shape using the mean estimate of the covariance matrix, using CW OPA.

Missing landmarks algorithm: mean estimates

$$\bar{\mu} = \begin{bmatrix} -0.03 & 5.008 \\ 40.027 & -0.024 \\ 0.03 & -4.967 \\ -40.027 & -0.017 \end{bmatrix},$$

$$\bar{\Sigma} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.011 & 0 & -0.009 & 0 & -0.05 & 0 & 0.054 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -0.009 & 0 & 0.012 & 0 & 0.059 & 0 & -0.066 \\ 0 & 0 & 0 & 0 & 0.006 & 0.013 & -0.004 & -0.015 \\ 0 & -0.05 & 0 & 0.059 & 0.013 & 13.459 & -0.019 & -11.393 \\ 0 & 0 & 0 & 0 & -0.004 & -0.019 & 0.006 & 0.024 \\ 0 & 0.054 & 0 & -0.066 & -0.015 & -11.393 & 0.024 & 13.711 \end{bmatrix},$$

$$\bar{\Sigma}^* = \begin{bmatrix} 0.189 & -0.008 & -0.188 & 0.006 & 0.007 & 0.02 & 0 & -0.027 \\ -0.008 & 0.011 & 0.007 & -0.01 & -0.001 & 0 & -0.001 & 0.002 \\ -0.188 & 0.007 & 0.187 & -0.007 & -0.007 & -0.02 & -0.001 & 0.027 \\ 0.006 & -0.01 & -0.007 & 0.011 & 0.001 & -0.001 & 0.001 & -0.002 \\ 0.007 & -0.001 & -0.007 & 0.001 & 0.281 & -0.274 & 0.269 & -0.276 \\ 0.02 & 0 & -0.02 & -0.001 & -0.274 & 0.276 & -0.273 & 0.271 \\ 0 & -0.001 & -0.001 & 0.001 & 0.269 & -0.273 & 0.277 & -0.273 \\ -0.027 & 0.002 & 0.027 & -0.002 & -0.276 & 0.271 & -0.273 & 0.278 \end{bmatrix}.$$

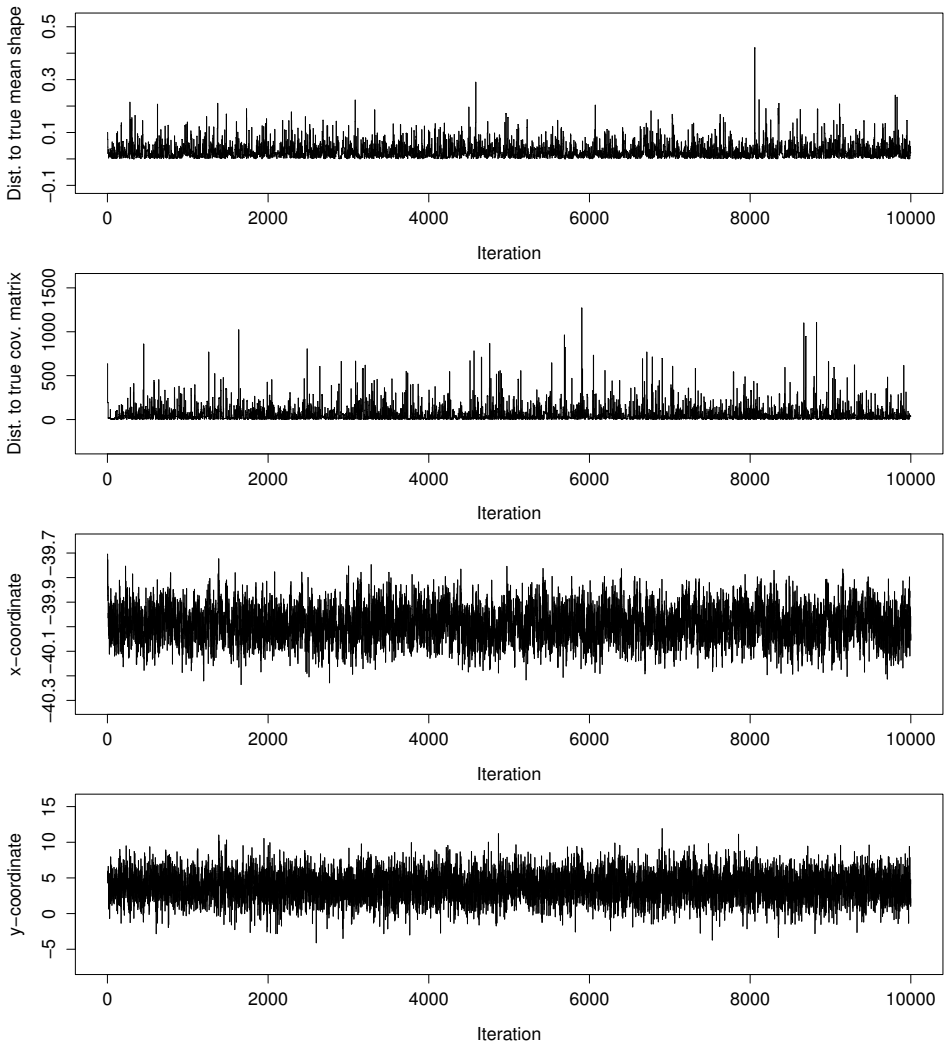

Figure 3.13: The squared Euclidean norm distance of the mean shape (top) and covariance matrix (second from top) from the true model, and the x (next to bottom) and y (bottom) co-ordinates for a missing landmark, at each iteration.

Missing landmarks algorithm: ML estimates

$$\widehat{\mu} = \begin{bmatrix} -0.121 & 5 \\ 40.044 & -0.025 \\ 0.119 & -4.981 \\ -40.042 & 0.005 \end{bmatrix},$$

$$\widehat{\Sigma} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0.004 & 0 & -0.004 \\ 0 & 0.012 & 0 & -0.012 & 0 & -0.25 & 0.005 & 0.254 \\ 0 & 0 & 0 & 0 & 0 & 0.004 & 0 & -0.004 \\ 0 & -0.012 & 0 & 0.016 & 0 & 0.283 & -0.006 & -0.279 \\ 0 & 0 & 0 & 0 & 0.006 & -0.017 & -0.003 & 0.016 \\ 0.004 & -0.25 & 0.004 & 0.283 & -0.017 & 12.731 & -0.142 & -11.548 \\ 0 & 0.005 & 0 & -0.006 & -0.003 & -0.142 & 0.006 & 0.145 \\ -0.004 & 0.254 & -0.004 & -0.279 & 0.016 & -11.548 & 0.145 & 12.436 \end{bmatrix},$$

$$\widehat{\Sigma}^* = \begin{bmatrix} 0.181 & -0.033 & -0.182 & 0.033 & 0.003 & 0.032 & -0.021 & -0.014 \\ -0.033 & 0.013 & 0.033 & -0.013 & -0.002 & -0.005 & 0.004 & 0.003 \\ -0.182 & 0.033 & 0.183 & -0.034 & -0.002 & -0.033 & 0.022 & 0.013 \\ 0.033 & -0.013 & -0.034 & 0.014 & 0.001 & 0.006 & -0.005 & -0.002 \\ 0.003 & -0.002 & -0.002 & 0.001 & 0.135 & -0.129 & 0.124 & -0.13 \\ 0.032 & -0.005 & -0.033 & 0.006 & -0.129 & 0.135 & -0.132 & 0.127 \\ -0.021 & 0.004 & 0.022 & -0.005 & 0.124 & -0.132 & 0.134 & -0.126 \\ -0.014 & 0.003 & 0.013 & -0.002 & -0.13 & 0.127 & -0.126 & 0.13 \end{bmatrix}.$$

The mixing of the missing landmarks is good, with not much movement in the x direction and considerably more movement in the y direction as expected. In the example plotted in Figure 3.13, the y movement is quite variable as dictated by the covariance matrix, but negatively correlated with landmark 2 of the same shape, which had a negative y value. The missing landmark algorithm did not mix well when shapes were missing landmarks 1 or 3 because these landmarks have much less variability and are, therefore, more important to the registration. However, most data sets would not have landmarks with such comparatively high and low values of variability, so this shouldn't be a problem, especially when data sets with more landmarks are analysed.

3.7 Discussion

In this chapter we have shown how combining *a priori* information and the CW OPA estimates of Chapter 2, it is possible to produce estimates of the covariance matrix of Procrustes residuals following registration. In the maximum likelihood algorithm, a variety of parameterisations of the covariance matrix were suggested and appropriate constraints given for the most general case. Constraining the variability in the directions of scaling, rotation and translation allows for the pre-isotropic registration covariance matrix to be reconstructed, while filtering out the remaining transformations. The method is dependent on the choice of eigenvalues for the constrained directions. These should be specified *a priori*, but could be varied to give the analyst a choice of optimum registrations.

We have also considered a Bayesian MCMC approach to shape data, as an alternative to the frequentist maximum likelihood method. We have shown that the posterior distributions of the transformation parameters, μ and Σ are standard distributions from which we can easily sample, using Gibbs steps, for the multivariate Gaussian model. The exception is the rotation angle which can be updated using a Metropolis-Hastings step. For a given mean shape and covariance matrix, the maximum likelihood estimates of the transformation parameters are known, so a hybrid algorithm has also been suggested that considers the transformation parameters as artifacts of the data given the other model parameters.

The hybrid algorithm has also been shown to be robust when adapted to cope with missing data. The mean shape and covariance matrix continue to mix well, and the estimates of their true values are comparable to their estimates using the full data set. The ability to cope with missing landmarks is an advantage of the MCMC methods. In future work, missing landmarks could be incorporated into the maximum likelihood methods using an EM algorithm, with the expected locations of the missing landmarks given by the expectation step and the model parameters estimated with the

maximisation step.

Both the CWMLE algorithm and the MCMC algorithms require *a priori* knowledge of the covariance structure. The method of CWMLE gives a framework for this knowledge to be incorporated into Procrustes methodology. MCMC methods are more robust when the *a priori* knowledge is less accurate, but computationally they take much longer. However, different constraints on Σ in the maximum likelihood method may give different registrations which highlight different features of shape variability. Both methodologies give improved estimates of the covariance matrix, compared with isotropic GPA, for data sets with non-isotropic covariance structures, which achieves the aim of covariance weighted Procrustes analysis.

Chapter 4

Surface shape and symmetry analysis

4.1 Introduction

Landmark based methods, such as Procrustes analysis, summarise the shape of an object with a few co-ordinates. However, it is becoming increasingly important, with the advent of greater computational power, to treat shapes as continuous curves and surfaces. Statistical shape analysis of surfaces ensures all the data regarding an object's shape is included within the analysis but raises questions regarding the sensible labelling of an object's features to ensure a valid correspondence between individuals is maintained. Correspondence can be either in terms of biological homology or in a geometrical sense. There have been a growing number of examples in the literature where shapes of curves and surfaces are investigated (e.g. van Essen *et al.*, 1998, Bookstein *et al.*, 1999, Kent *et al.*, 2000, Fischl *et al.*, 2001, Hobolth *et al.*, 2002, Klassen *et al.*, 2004). In this chapter we develop maximum likelihood and Bayesian based approaches to the statistical analysis of surfaces. We shall investigate methods for surface shape analysis by concentrating on an important application in neuroscience which motivated our work.

4.2 The application

The data are 68 magnetic resonance images (MRI) of the human brain that were collected by Sean Flynn, University of British Columbia. The volunteers were composed of healthy controls and patients clinically diagnosed with schizophrenia. Our aim is to develop and apply a statistical analysis of the cortical surfaces to test for correlations between volunteer group and surface shape. The analysis of shape is more likely to yield significant differences than global indices such as hemispheric volumes (Csernansky *et al.*, 1998). We are particularly interested in large-scale shape differences, such as asymmetry. The cortical surface of the brain tends to exhibit asymmetry, and in particular the right frontal region is larger than the left, and the left occipital region is larger than the right. This particular asymmetry is often called 'brain torque' and the torque differs with handedness and gender (Kertesz *et al.*, 1990). In particular, females tend to exhibit less torque on average (Barrick *et al.*, 2005). It has been suggested that schizophrenia patients tend to have less torque (Bilder *et al.*, 1994, Mackay *et al.*, 2003).

We use the following Euclidean co-ordinate system throughout: x-axis: posterior $\rightarrow$ anterior (back to front), y-axis: inferior $\rightarrow$ superior (bottom to top), z-axis: right $\rightarrow$ left (n.b. 'left' = patient's left). The sagittal plane is the $x-y$ plane, the coronal plane is the $y-z$ plane, and the axial plane is the $x-z$ plane. This is an unconventional labelling of the x, y, z axes, but of course the choice of axis labels is arbitrary. Each volunteer's image consists of $256 \times 256 \times 256$ voxels (three-dimensional pixels) and each voxel has an intensity value. Voxels with high intensity are commonly shown as white on a grey-scale, with low intensity values as black.

All volunteers were aged under 50 and in total there were 29 male controls, 25 male schizophrenia patients, 9 female controls and 5 female schizophrenia patients. The mean ages are: male controls (36.6), male patients (33.2), female controls (33.9), female patients (33.4). All the subjects were right-handed (writing hand) except one male patient and one male

control. We denote the images as, Y_j, $j = 1, ..., n = 68$ (each 256^3 vectors), and the covariate vector, x_j, consisting of patient group (control=1/ schizophrenia=2), age and gender (male=1/female=2).

4.3 The model

In this section we will specify a model, and a likelihood, for the distribution of voxel values in each image. There are two major confounds in MRI analysis, the identification and removal of non-brain voxels and the registration of images to a template. We shall consider the removal of non-brain voxels from the data set to be a pre-processing step. However, we shall include the registration parameters in the model and estimate their values using maximum likelihood and Bayesian approaches.

4.3.1 Cortical surface segmentation

There are many sophisticated image analysis programs available to assist with common tasks in image analysis. We use the brain extraction tool (BET) of Smith (2002), which is available in the FSL software package (Smith *et al.*, 2004), to extract the cortical surface boundary, which is the boundary between the grey matter and the CSF. The tool fits a balloon-like template through an energy minimisation scheme. The algorithm is optimised for each brain through the control of a tuning parameter. The resulting image is the same size as the original but with zero voxel values for those outside the cortical surface boundary. Note that any error in estimating the cortical surface is considerably less than the variability between different brains and so it is a reasonable practical approach to treat the cortical surface boundary as part of the data (i.e. as known).

4.3.2 Model parameters

Each brain has been approximately orientated in the scanner, but we wish to remove all differences in translation and rotation to make comparisons between different individuals. The registration of each brain is obtained by estimating rigid-body transformations, with translation represented by $\xi = (\xi_x, \xi_y, \xi_z)^T \in \mathbb{R}^3$ (new location of the origin) and rotation matrix $\Gamma(\theta_p, \theta_r, \theta_y) \in SO(3)$ about this point which is a function of the three Eulerian angles: θ_p (pitch angle about x axis), θ_r (roll angle about y axis), θ_y (yaw angle about z axis). Let $\phi_j = (\xi_{xj}, \xi_{yj}, \xi_{zj}, \theta_{pj}, \theta_{rj}, \theta_{yj})^T$ be the vector of the six rigid-body registration parameters and C_j be the cortical surface, for scan $j = 1, ..., n$.

We register each brain into Talairach space (Talairach and Tournoux, 1988), a three dimensional frame of reference, based on the Cartesian co-ordinate system, for locating internal and external features of the brain. This procedure involves locating the inter-hemispherical join, which we call the midplane, that separates the left and right halves of the brain, and locating two commissure landmarks, the anterior commissure (AC) and the posterior commissure (PC). The AC and PC are areas of white matter in the midplane that link the left and right hemispheres. The final registration involves rotating the image so that the midplane is vertical and the line joining the AC and PC is horizontal. In this new co-ordinate system, we take the origin to be the AC and the location of the PC is given by an additional parameter, $\xi_c > 0$, specifying the distance between them. For any registration the new registered image is defined on the same voxel grid as the old image using a suitable wrap around the edges of the image.

4.3.3 The likelihood

We assume the scans are independent of each other, and we specify the voxel values of the jth individual, Y_j, to have the distribution, $Y_j|\{\phi_j, x_j\}$, where ϕ_j is the vector of 6 parameters and x_j is the covariate vector of patient

group, gender and age. The first part of the registration into Talairach space is based on the approximate symmetrical structure of the brain in the close vicinity of the midplane, and so a very important part of the model is the distribution of the voxel values in the midplane region. The remaining registration steps consist of finding the midplane, the AC and the PC, so we partition the voxel grid into corresponding distinct regions which depend on the parameters of interest and specify independent distributions for each region. The regions, defined with respect to the constants $\epsilon_M = 15$, $\epsilon_A = \epsilon_P = 10$ in millimetres/voxels, used in the model are:

1. the midplane region, $\mathcal{M}$, defined as the voxels within distance ϵ_M of the midplane such that the voxel and its mirror image about the midplane are both non-zero;

2. the AC region $\mathcal{A}$, defined as the voxels within distance ϵ_A in the x, y and z directions of the AC landmark;

3. the PC region $\mathcal{P}$, defined as the voxels within distance ϵ_P in the x, y and z directions of the PC landmark;

4. the region inside the cortical surface, $\mathcal{O}_1$, containing non-zero valued pixels that are not included in $\mathcal{M}$, $\mathcal{A}$ or $\mathcal{P}$;

5. the region outside the cortical surface, $\mathcal{O}_2$, containing zero-value pixels which do not contribute to the likelihood.

Both voxels are required to be non-zero in the midplane region in order to examine the symmetry of the voxel-values. A schematic diagram of these regions is shown is Figure 4.1. Therefore, omitting the j subscripts, the density for one individual consists of parts for the midplane, the AC and the PC, plus one part for non-zero voxels not near these three features,

$$\begin{aligned} f(Y|\phi, x) &= f_1(Y \cap \mathcal{M}|(\theta_p, \theta_r, \xi_z), x) f_2(Y \cap \mathcal{A}|(\theta_p, \theta_r, \xi_x, \xi_y, \xi_z), x) \\ &\quad \times f_3(Y \cap \mathcal{P}|\phi, x) f_4(Y \cap \mathcal{O}_1|\phi, x). \end{aligned}$$

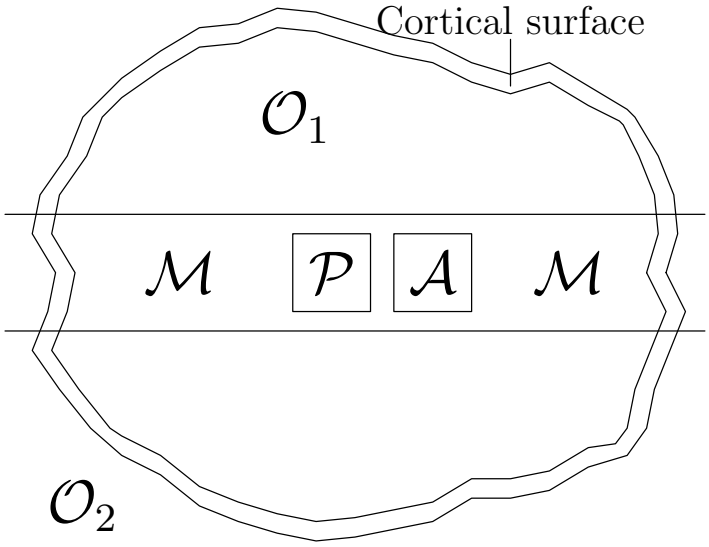

Figure 4.1: A schematic diagram showing the brain regions.

We shall carry out inference using the likelihood function,

$$L(Y_1, ..., Y_n|\phi_1, ..., \phi_n, x_1, ..., x_n) = \prod_{j=1}^{n} f(Y_j|\phi_j, x_j).$$

Given the distribution is uniform for the voxels in $\mathcal{O}_1$, the contribution to the likelihood from these voxels is constant with respect to the model parameters.

For one individual, let Y_t denote the voxel value at location $t = (t_x, t_y, t_z) \in \mathcal{S} = \{0, ..., 255\}^3$. The first stage of registration involves a rigid-body transformation with translation in z to ξ_z and rotations by θ_p, θ_r. Following a transformation, each voxel t is sent to a new location $s = s(t, \theta) = (s_x, s_y, s_z)$ rounded to the nearest integer. The midplane is given by new co-ordinates, (s_x, s_y, ξ_z) and we take the midplane as lying in the half-integer position for ξ_z, lying half way between two planes of voxels.

Let s indicate voxels which are to the left of the midplane ($s_z > \xi_z$). Let $s' = (s_x, s_y, \xi_z - |s_z - \xi_z|)$ be the reflection of s about the midplane. The midplane region is $\mathcal{M} = \{s : |s_z - \xi_z| \leq \epsilon_M\}$, where the paired voxels, s and s', are both non-zero. We transform each pair of voxel values in $\mathcal{M}$ from $(Y_s, Y_{s'})$ to $U_s = (Y_s - Y_{s'})/2$, $V_s = (Y_s + Y_{s'})/2$, with Jacobian 1. We assume that all V_s are independent of all U_s, and the joint likelihood of all V_s is assumed constant with respect to the model parameters. For voxels in $\mathcal{M}$, we consider a model where the expected values of the grey levels Y_s are symmetrical about the midplane, $\mathbb{E}(Y_s) = \mathbb{E}(Y_{s'})$, so $\mathbb{E}(U_s) = 0$. For simplicity, we assume that the U_s are independent.

Exploratory data analysis is carried out by examining the histogram of U_s for a good choice of midplane for several MRI scans. A Laplace (double exponential) distribution fits the data well in $\mathcal{M}$, as seen in the example in Figure 4.2 middle plot. Therefore, a suitable model for U_s with $s(t, \theta) \in \mathcal{M}$ is the Laplace distribution with scale parameter λw_s,

$$f(u_s) = \frac{\lambda w_s}{2} \exp(-\lambda w_s |u_s|), \tag{4.1}$$

where $w_s = w_{s'}$ are predetermined weights. The weights were obtained by examining the sample variance of the voxels at parallel planes to a good midplane choice. The weight of the sth voxel at a perpendicular distance from the midplane, ξ_z, is taken as $w_s = \max\{(10.5 - |s_z - \xi_z|)/10, 0.5\}$. The Laplace distribution is often used in image analysis instead of a Gaussian distribution because it is more robust to outliers. In this application, the observed data has heavier tails and a sharper peak than a fitted Gaussian or Student's t distribution.

To model the commissure regions, $\mathcal{A}$ and $\mathcal{P}$, training data is obtained by manually locating the AC and PC on the midplane of $n_t = 7$ scans. After translating and rotating these images into their final registration the voxel intensities, Y_s, in the region $\mathcal{A} \cup \mathcal{P}$ are standardised for each scan to a common mean and variance, by means of a transformation of the form,

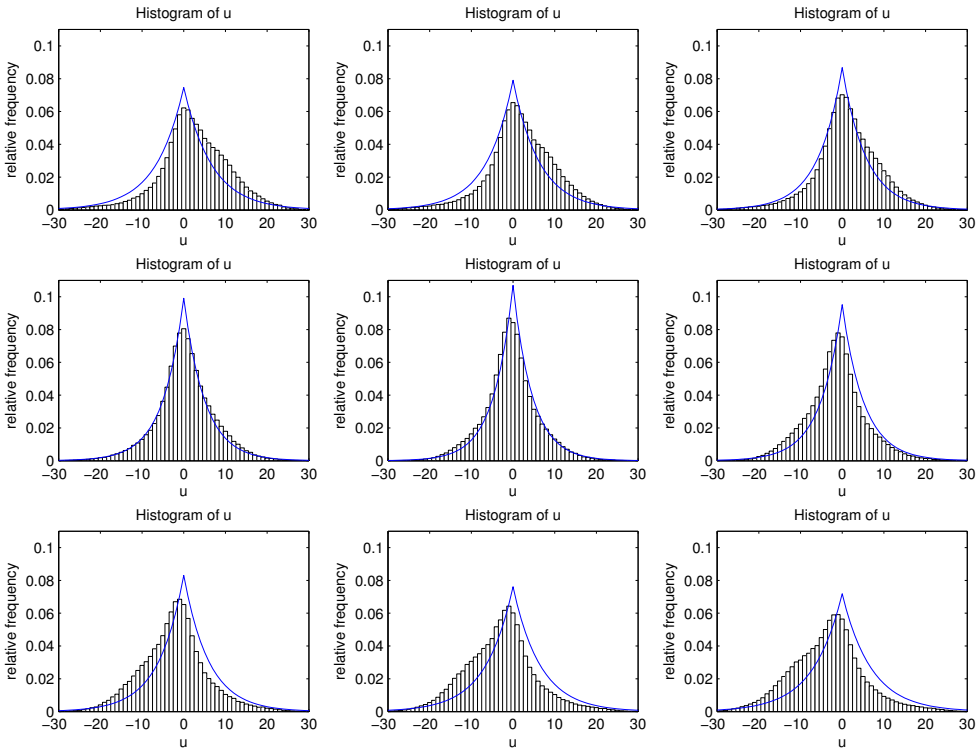

Figure 4.2: Histograms of u_s for several choices of midplane and fitted Laplace distributions. The middle plot with $\xi_z = 130.5$ corresponds to the best choice here.

$Y_s^* = \beta_j Y_s + \gamma_j$, $j = 1, ..., n_t$. For each voxel, s, in $\mathcal{A}$ and $\mathcal{P}$, we can calculate an estimate for the mean (μ_{As} or μ_{Ps}) and variance (σ^2_{As} or σ^2_{Ps}) of Y_s^* from the training data. The use of training data and models of voxel intensities is common in landmark identification (e.g. Izard *et al.*, 2005).

We model the voxel intensity at each voxel in $\mathcal{A}$ and $\mathcal{P}$ to be independent with the Gaussian distributions,

$$N(\mu_{As}, \sigma^2_{As}/w_{As}), \qquad N(\mu_{Ps}, \sigma^2_{Ps}/w_{Ps}),$$

respectively, where w_{As}, w_{Ps} are weights based on the distance of voxel s from the each commissure. We take the weights as the reciprocal of $1 + (c_x - s_x)^2 + (c_y - s_y)^2 + (c_z - s_z)^2$, where (c_x, c_y, c_z) is the location of the commissure and (s_x, s_y, s_z) is the location of the sth voxel. Likewise, for the jth individual, we linearly transform the voxel values in $\mathcal{A} \cup \mathcal{P}$ such that the mean (and variance) of the voxel values in the image and the template are identical. Rather than treating β_j, γ_j as additional parameters we substitute estimates, $\widehat{\beta}_j$, $\widehat{\gamma}_j$, based on the voxel values in $\mathcal{A} \cup \mathcal{P}$.

Therefore, ignoring all terms independent of the registration parameters and omitting the j subscript, the log-likelihood for the non-zero voxels in one individual is,

$$\log L(U, V|\phi, \lambda, \xi_c) = \sum_{s,s' \in \mathcal{M}} \left\{ \log\left(\frac{\lambda w_s}{2}\right) - \lambda w_s |u_s| \right\} \tag{4.2}$$
$$- \sum_{s \in \mathcal{A}} \frac{w_{As}}{2\sigma^2_{As}} \left(\widehat{\beta} y_s + \widehat{\gamma} - \mu_{As}\right)^2 - \sum_{s \in \mathcal{P}} \frac{w_{Ps}}{2\sigma^2_{Ps}} \left(\widehat{\beta} y_s + \widehat{\gamma} - \mu_{Ps}\right)^2.$$

In order to formulate this likelihood it has been necessary to make a number of modelling assumptions regarding the choice of distributions, training data, weights and tuning parameters such as ϵ_M, ϵ_A and ϵ_P. Such assumptions are commonly made in image analysis to simplify the model such that it is computationally feasible. In practice, these choices will influence the bias and variability in the estimates of the model parameters, such as the ro-

tations and translations, however, the values chosen seem reasonable based on some preliminary sensitivity analysis to tune the model to the acquired dataset.

4.4 Maximum likelihood registration

We consider registration of each brain to Talairach space using a four stage procedure:

1. The midplane is estimated by maximising over $\theta = (\theta_p, \theta_r, \xi_z)$ in the region $\mathcal{M}$ to give $\widehat{\theta}$;

2. Approximate estimates of (ξ_x, ξ_y, θ_y) are found by maximising the likelihood over $\mathcal{A} \cup \mathcal{P}$, given $\widehat{\theta}$;

3. The location of the AC is estimated by maximising over (ξ_x, ξ_y), given $\widehat{\theta}$, in the region $\mathcal{A}$ to give $\widehat{\xi}_x$, $\widehat{\xi}_y$;

4. The location of the PC is estimated by maximising over (θ_y, ξ_c), given $(\widehat{\theta}, \widehat{\xi}_x, \widehat{\xi}_y)$, in the region $\mathcal{P}$ to give $\widehat{\theta}_y$, $\widehat{\xi}_c$.

In each case, a simple grid search over the parameters is performed, with steps of 0.01 radians for the angles, and 1mm for the translations. Han and Park (2004) propose an alternative rigid-body registration based on a similar 4-step procedure, but do not specify a likelihood for the voxel values.

For the midplane registration, the log likelihood given the cortical surface, C, from BET is,

$$\log L(U|\theta, \lambda, C) = m \log \lambda + \left(\sum_{s \in \mathcal{M}: s_z > 0} \log \frac{w_s}{2} \right) - \lambda \left(\sum_{s \in \mathcal{M}: s_z > 0} w_s |u_s| \right),$$

where m is the number of voxel pairs in $\mathcal{M}$. Taking derivatives,

$$\frac{\partial \log L}{\partial \lambda} = \frac{m}{\lambda} - \sum_{s \in \mathcal{M}: s_z > 0} w_s |u_s|, \qquad \Rightarrow \qquad \widehat{\lambda}^{-1} = \frac{1}{m} \sum_{s \in \mathcal{M}: s_z > 0} w_s |u_s|.$$

Therefore, the maximum likelihood registration given C is obtained by maximising,

$$\log L(U|\theta, \widehat{\lambda}, C) = m(\log \widehat{\lambda} - 1) + \sum_{s \in \mathcal{M}_1: s_z > 0} \log \frac{w_s}{2},$$

over θ.

After the midplane has been estimated, the registration is completed by locating the AC and PC landmarks in the new midplane. Translating the AC to the origin and rotating the AC-PC line to horizontal fixes the final registration in Talairach space. In practice, locating the AC and PC is more accurate when the size of the regions $\mathcal{A}$ and $\mathcal{P}$ is small, only containing the voxels that include the commissures. Therefore, after estimating approximate locations, we let w_{As} and w_{Ps} tend to zero for those voxels outside the immediate neighbourhood of the AC and PC.

The above four step procedure is applied to each image in turn. Here we present the results for one individual. For stage 1, we find that the maximum likelihood estimators are approximately $\widehat{\theta}_p = -0.07$, $\widehat{\theta}_r = -0.06$, and $\widehat{\xi}_z = 130.5$, with $\widehat{\lambda}^{-1} = 3.3613$. In Figure 4.3, we see the image transformed from its original orientation to the maximum likelihood registration of the midplane. Note that after the transformation the crosshairs bisect the brain's two hemispheres. In Figure 4.2 we saw histograms of the voxels in $\mathcal{M}$ and the fitted Laplace densities for different choices of $\xi_z \in \{126.5, ..., 134.5\}$ with $\xi_z = 130.5$ fitting well in the middle plot, although $\xi_z = 129.5$ also looks reasonable.

Evaluating the log-likelihood to register our example image on the AC and PC, we find the maximum likelihood estimators are approximately $\widehat{\xi}_x =$

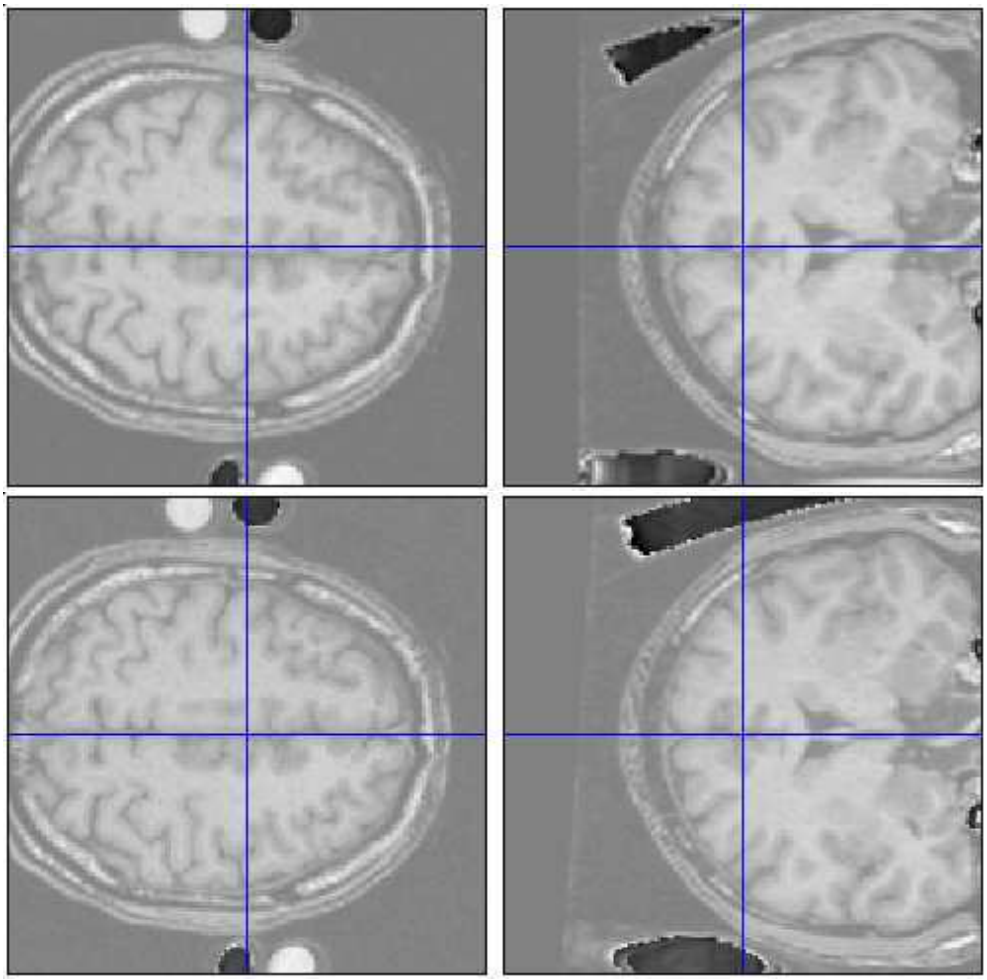

Figure 4.3: An image transformed from its original orientation (top row) to the maximum likelihood registration of the midplane (bottom row).

124.5 and $\widehat{\xi}_y = 169.5$. In Figure 4.4 we see the image translated from its midplane registration such that the origin coincides with the AC. In the final step we rotate the image about this origin such that the AC-PC line is horizontal. The maximum likelihood estimator was found to be $\widehat{\theta}_y = 0.29$. The final registration is seen in Figure 4.5.

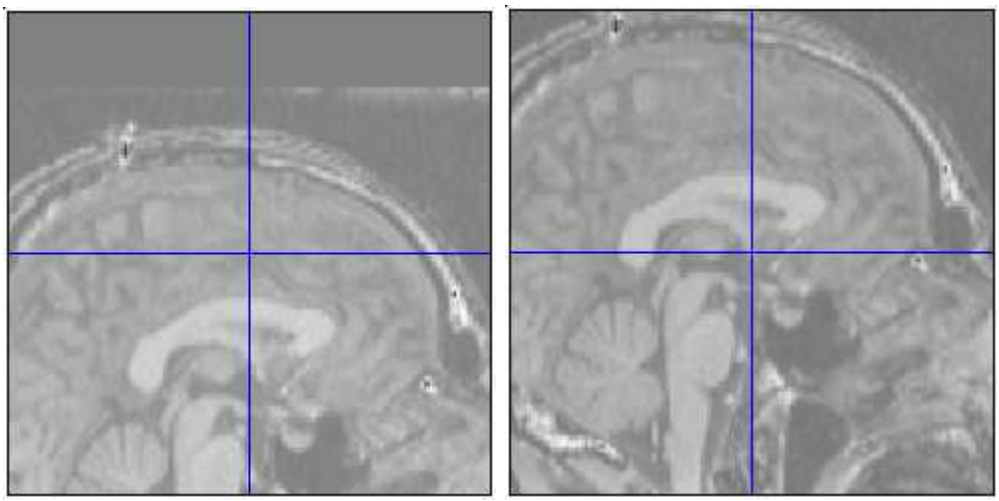

Figure 4.4: Following midplane registration (left), the image is transformed so the origin coincides with the AC (right).

4.5 Bayesian registration

Alternatively, we could consider a Bayesian procedure for the registration. For an introduction to Bayesian image analysis, see Hurn *et al.* (2003). We assume that the eight parameters, the registration parameters, $\phi = (\xi_x, \xi_y, \xi_z, \theta_p, \theta_r, \theta_y)$, the concentration parameter for voxels in $\mathcal{M}$, λ, and the inter-commissure distance, ξ_c, have independent prior distributions. The prior distributions for ϕ and ξ_c are taken to be uniform on a bounded region, and the prior distribution for the concentration parameter is $\lambda \sim \Gamma(\alpha_0, \beta_0)$. For each voxel, s, in the AC and PC regions, $\mathcal{A}$ and $\mathcal{P}$, we calculate an

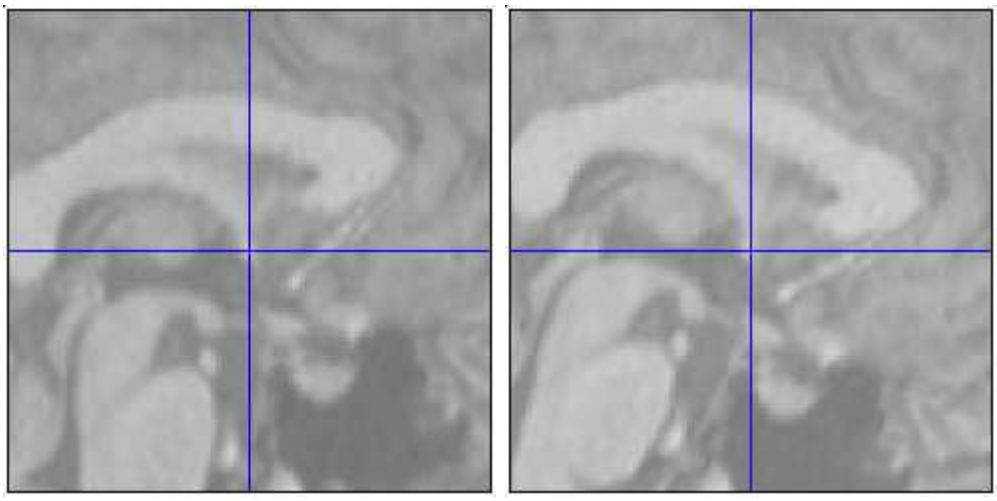

Figure 4.5: The image is rotated about the AC (left) such that the AC-PC line is horizontal (right), with the PC being the white area on the horizontal axis approximately three quarters of the way between the centre and the left edge of the image.

estimate for the mean (μ_{As} or μ_{Ps}) and variance (σ^2_{As} or σ^2_{Ps}) from the training data and, therefore, assume these parameters to be known.

Let m be the number of paired voxels in $\mathcal{M}$, then the conditional posterior density for λ is given by,

$$\begin{aligned}\pi(\lambda|Y,\phi,\xi_c,C) &\propto P(\lambda)L(Y|\phi,\lambda,\xi_c,C),\\ &\propto \lambda^{\alpha_0-1}\exp(-\beta_0\lambda)\lambda^m\exp\left\{-\lambda\sum_{s\in\mathcal{M}:s_z>0}w_s|u_s|\right\},\end{aligned}$$

where the likelihood is obtained from Equation (4.1). Therefore,

$$\pi(\lambda|Y,\phi,\xi_c,C)\sim\Gamma\left(m+\alpha_0,\beta_0+\sum_{s\in\mathcal{M}:s_z>0}w_s|u_s|\right). \tag{4.3}$$

Recall that we only sum over $s_z>0$ because u_s is a function of the difference in voxel values of pairs located in left and right hemispheres.

The log posterior density for the whole model is given by,

$$\begin{aligned}\log \pi(\phi, \lambda, \xi_c | Y, C) &= \log P(\phi, \lambda, \xi_c) + \log L(Y|\phi, \lambda, \xi_c, C) + k, \\ &= (\alpha_0 - 1)\log\lambda - \beta_0\lambda + \log L(Y|\phi, \lambda, \xi_c, C) + k,\end{aligned}$$

where k is a constant and $\log L(Y|\phi, \lambda, \xi_c, C)$ is given by Equation (4.2). The form of the posterior is very complicated but we can simulate from the posterior using a Markov chain Monte Carlo (MCMC) algorithm.

Algorithm 4.5.1 ***MCMC algorithm for brain registration***

1. *Start at $\phi = \phi_0$, $\xi_c = \xi_{c0}$, and simulate λ from its prior distribution.*

2. *Metropolis-Hastings (MH) step: Propose $\xi_z^* \sim N(\xi_z, \sigma_1^2)$. Let $\phi^* = (\xi_x, \xi_y, \xi_z^*, \theta_p, \theta_r, \theta_y)$. Accept $\phi = \phi^*$ with probability,*

$$p = \min\left(1, \frac{\pi(\phi^*, \lambda, \xi_c|Y, C)}{\pi(\phi, \lambda, \xi_c|Y, C)}\right). \tag{4.4}$$

3. *MH step with proposal $\theta_p^* \sim N(\theta_p, \sigma_2^2)$ and $\phi^* = (\xi_x, \xi_y, \xi_z, \theta_p^*, \theta_r, \theta_y)$. Accept $\phi = \phi^*$ with probability p, see Equation (4.4).*

4. *MH step with proposal $\theta_r^* \sim N(\theta_r, \sigma_3^2)$ and $\phi^* = (\xi_x, \xi_y, \xi_z, \theta_p, \theta_r^*, \theta_y)$. Accept $\phi = \phi^*$ with probability p, see Equation (4.4).*

5. *Gibbs step: Simulate λ from its conditional posterior distribution, given by Equation (4.3).*

6. *MH step with proposal $\xi_x^* \sim N(\xi_x, \sigma_4^2)$ and $\phi^* = (\xi_x^*, \xi_y, \xi_z, \theta_p, \theta_r, \theta_y)$. Accept $\phi = \phi^*$ with probability p, see Equation (4.4).*

7. *MH step with proposal $\xi_y^* \sim N(\xi_y, \sigma_5^2)$ and $\phi^* = (\xi_x, \xi_y^*, \xi_z, \theta_p, \theta_r, \theta_y)$. Accept $\phi = \phi^*$ with probability p, see Equation (4.4).*

8. *MH step with proposal $\theta_y^* \sim N(\theta_y, \sigma_6^2)$ and $\phi^* = (\xi_x, \xi_y, \xi_z, \theta_p, \theta_r, \theta_y^*)$. Accept $\phi = \phi^*$ with probability p, see Equation (4.4).*

9. *MH step with proposal* $\xi_c^* \sim N(\xi_c, \sigma_7^2)$*. Accept with probability,*

$$p = \min\left(1, \frac{\pi(\phi, \lambda, \xi_c^*|Y, C)}{\pi(\phi, \lambda, \xi_c|Y, C)}\right).$$

10. *Repeat steps 2-9 for a large number of times.*

Note that the above proposals are symmetric and so the proposal densities cancel in the Hastings ratios. Commonly, the variances for the sampling distributions would be specified *a priori*. However, we use an adapting stage (Browne and Draper, 2000) to adjust the sampling variances during the first few thousand iterations. The adapting procedure monitors the acceptance rates for each parameter in batches of 100 iterations. If the acceptance rates are in the interval (0.4,0.6) for three successive batches then the proposal density variances are fixed and the adapting stage ends. However, if this criterion is not satisfied, let r be the acceptance rate in the latest batch then the current proposal variance, σ_i, is modified to σ_i^* as follows.

$$\begin{aligned} &\text{If } r \geq 0.5, && \sigma_i^* = 2\sigma_i r. \\ &\text{If } r < 0.5, && \sigma_i^* = \frac{\sigma_i}{2(1-r)}. \end{aligned}$$

Since the Markov chain from Algorithm 4.5.1 is aperiodic, and irreducible, after a large number of iterations the chain converges to its stationary distribution. Therefore, we will have simulated an observation from the posterior distribution. In practice the changes in log-likelihood with different registrations are so large that the algorithm converges quickly to the neighbourhood of the maximum *a posteriori* (MAP) estimator, and the credibility intervals for the parameters are extremely narrow.

We apply the MCMC algorithm to the same image as in the previous section and the prior distributions are taken to be uniform distributions. Figure 4.6 shows the parameters over $20,000$ iterations. The starting values were taken to be the maximum likelihood estimators given in Section 4.4.

The MAP estimator is estimated as $\widehat{\xi}_x = 124.6, \widehat{\xi}_y = 169.4, \widehat{\xi}_z = 130.5, \widehat{\theta}_p = -0.071, \widehat{\theta}_r = -0.056, \widehat{\theta}_y = 0.277, \widehat{\xi}_c = 25.84, \widehat{\lambda}^{-1} = 2.8835$. Over the first 2000 iterations, we use the adapting stage to choose the variances for the sampling distributions, and we take the next 2000 iterations as the burn-in period. The marginal histograms of the parameters after the burn-in are given in Figure 4.7 as well as the log-likelihood (up to a constant). The posterior variability is small for all the registration parameters, especially those involved with midplane registration. As expected the (approximate) MAP estimator is close to the maximum likelihood estimate (MLE).

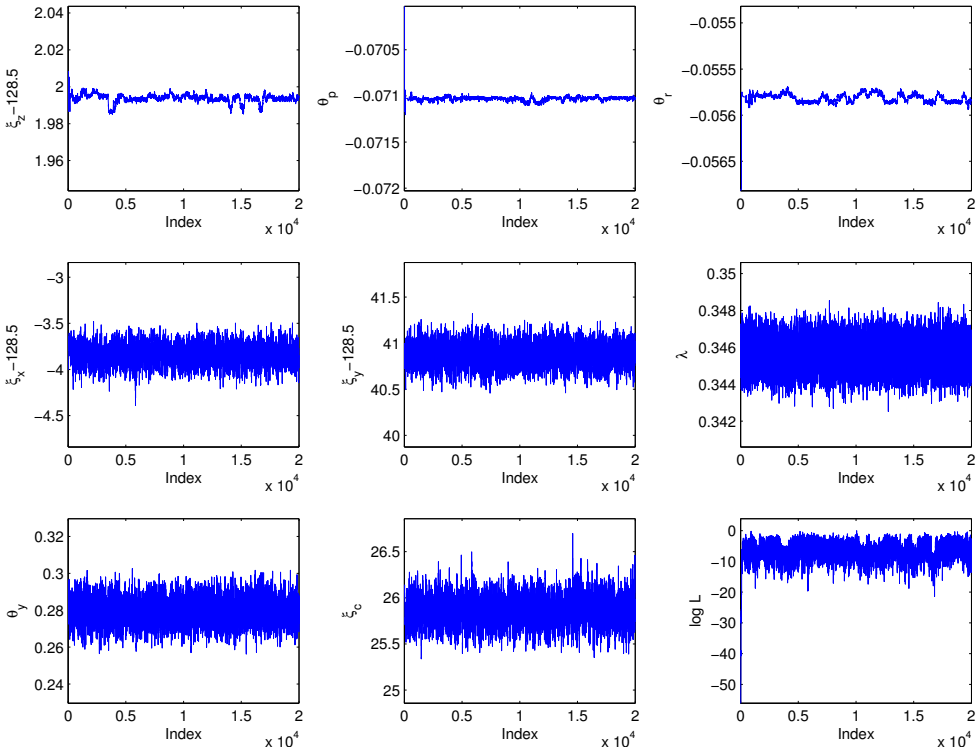

Figure 4.6: Plots of parameter values and the log likelihood from the MCMC algorithm over the first 20,000 iterations.

Similar results were obtained using a second MCMC simulation with starting values away from the maximum likelihood estimator. Figure 4.8

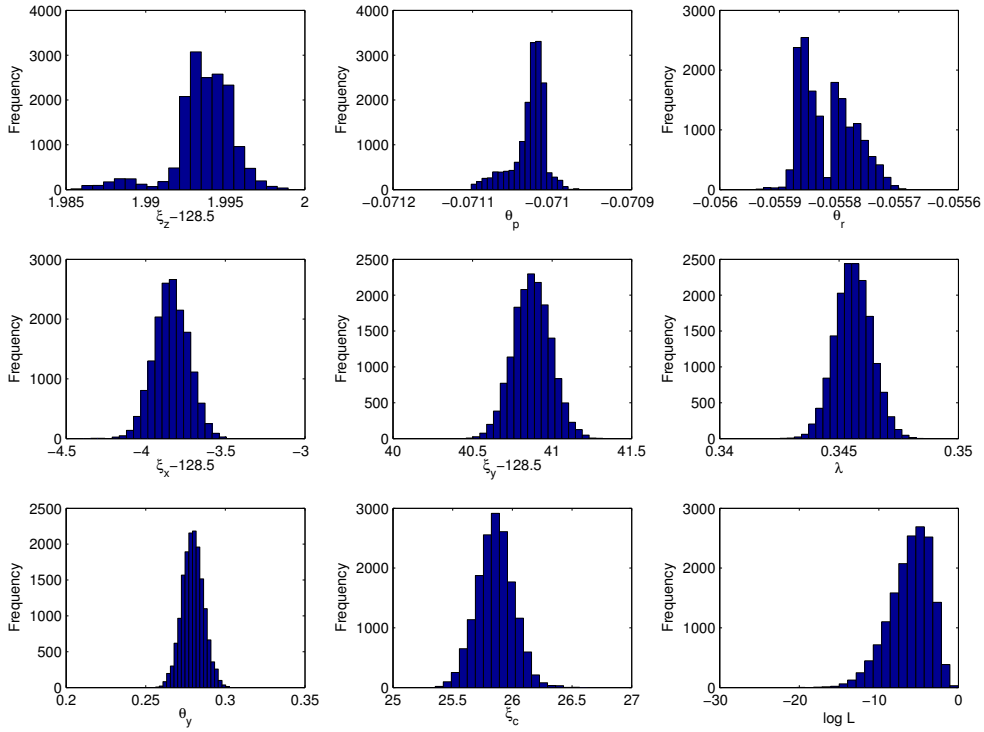

Figure 4.7: Histograms of the parameter values and the log likelihood from the MCMC run (after the burn-in period).

shows how each of the parameters converged to the vicinity of their posterior distributions during the first 500 iterations. It also illustrates how the adapting stage initially allows quite large jumps, before reducing the proposal variances to allow good mixing in the posterior distribution. Note that the variances of the posterior distributions for ξ_z, θ_p and θ_r are so small that the majority of jumps cannot be observed in Figure 4.8. The starting values of $\xi_x = 128.5, \xi_y = 158.5, \xi_z = 128.5, \theta_p = 0.0, \theta_r = 0.0, \theta_y = 0.2, \xi_c = 20.0$, are outside the range of human error for a manual registration, so the MCMC method is not reliant on the maximum likelihood registration.

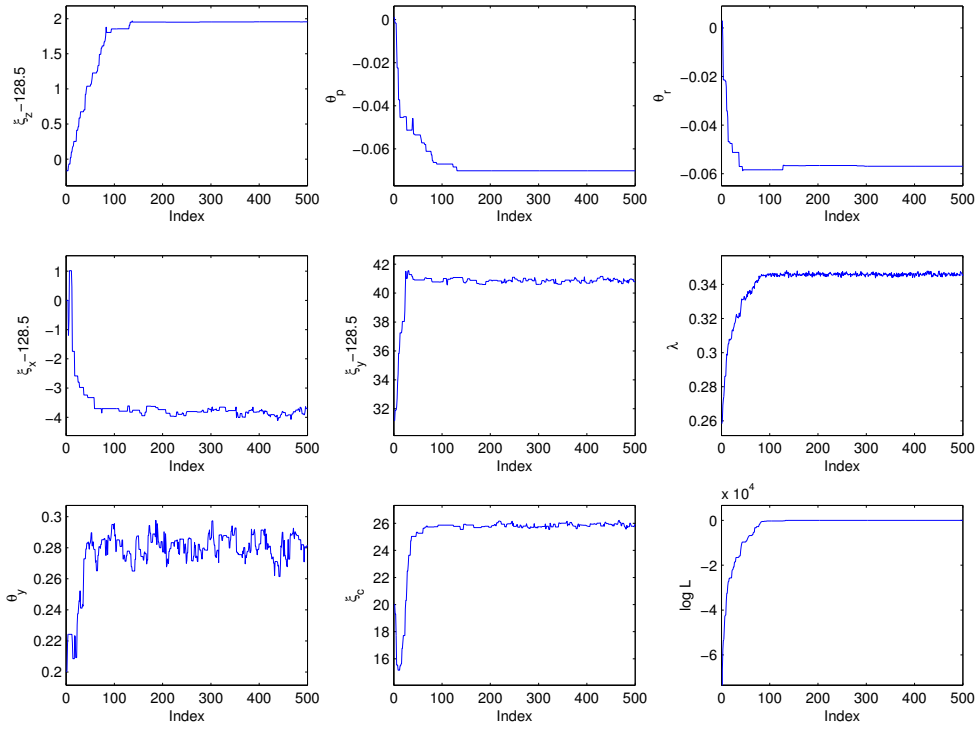

Figure 4.8: Plots of parameter values and the log likelihood from the MCMC algorithm over the first 500 iterations.

The MCMC algorithm has been applied to several of the images in the data set. The posterior distribution is always tightly concentrated around the MAP due to the large amount of information in each image. Also, the

MAP is always very close to the maximum likelihood estimator because of the weak prior information. Therefore, given the very small discrepancies between the two approaches, the maximum likelihood method of registration is applied to each image because it is considerably faster to compute.

4.6 Asymmetric shape analysis

In order to carry out shape analysis we must first provide a labelling of each brain. If anatomical landmarks on the cortical surface were easily identifiable, we could analyse asymmetry using Procrustes methods. Mardia *et al.* (2000) and Klingenberg *et al.* (2002) supplement their data sets with a reflected copy of each landmark configuration before implementing Procrustes superimposition. The total shape variability, and principal components, could then be allocated into symmetric and asymmetric components. However, the cortical surface is extremely intricate, with many folds and lobes, and highly variable between subjects so the use of anatomical landmarks is problematic. Theobald *et al.* (2004) investigates symmetry for star-shaped objects based on pseudo-landmarks given by the length of rays between the origin and the surface for a set of given angles. A similar analysis of our data set is presented in Brignell *et al.* (2006). We consider a labelling which gives rough correspondence between parts of the brain. This will enable us to examine the large-scale shape changes, such as brain torque, which we are interested in.

In order to measure asymmetry we consider a similar method described in outline by Collinson *et al.* (2003). Following registration by maximum likelihood to Talairach space, each scan is divided into $m = 100$ equally spaced axial slices. We estimate the volumes contained within the cortical surface boundaries in each slice above the horizontal plane containing the AC and PC, for the left hand side (V_{rj}^L) and right hand side (V_{rj}^R), for $r = 1, ..., m$, $j = 1, ..., n$, where $n = 68$ is the total number of subjects. We restrict the analysis to the cortical surface lying above the axial plane

in which the AC and PC lies, as the surface is most clearly defined in this region. The asymmetry function has components,

$$\eta_{rj} = (V_{rj}^R - V_{rj}^L)/T_j,$$

where T_j is the maximum slice volume in the jth scan. Therefore, the asymmetry function for scan j is $\eta_j = (\eta_{1j}, ..., \eta_{mj})^T$. Each function is smoothed using a Loess smoother with fraction $f = 0.05$. Smoothing is commonly carried out at a preliminary stage in functional data analysis (see Ramsay and Silverman, 2004).

In Figure 4.9 we see a plot of the mean smoothed asymmetry functions for each sub-group. The slices with low index (left end of each picture) are in the occipital region of the brain and the slices with high index (right end of each picture) are in the frontal region of the brain. In Figure 4.10 we display the smoothed asymmetry functions for the four sub-groups of data (male control, male patients, female control, female patients).

In Figure 4.11 we see the results of conducting a t-test on $\bar{\eta}_c - \bar{\eta}_s = 0$ at each slice, where $\bar{\eta}_c$ and $\bar{\eta}_s$ are the means of the control and patient group respectively. Controls appear to have significantly greater rightward asymmetry between slices 82 and 86. However, the effect is only just significant, and multiple tests are being carried out, so the evidence for an effect is not very strong.

While the underlying Gaussian distribution assumption behind the Student-t test seems reasonable here, we could consider performing a non-parametric test instead. An appropriate statistic, assuming the observations have been randomly and independently sampled from their respective populations, would be the Mann-Whitney U test obtained by ordering all $(n_c + n_s)$ observations of η at each slice, where n_c and n_s are the number of control and patient scans respectively. The Mann-Whitney U statistic

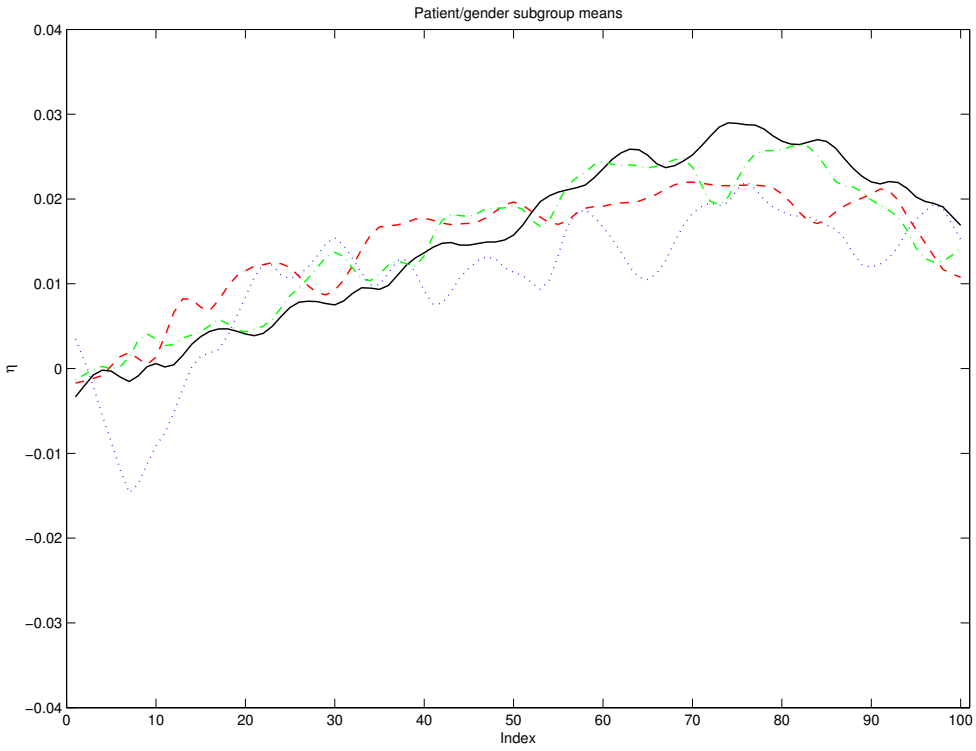

Figure 4.9: The mean smoothed asymmetry functions for the male controls (solid, black), male patients (dashed, red), female controls (dash/dot, green), female patients (dotted, blue).

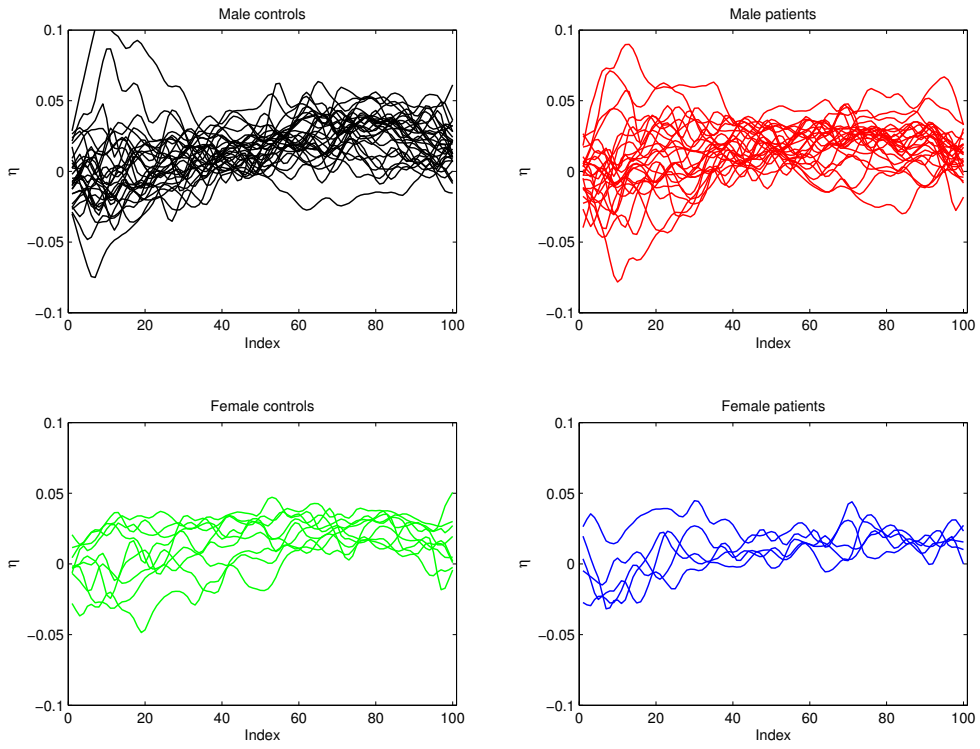

Figure 4.10: The smoothed asymmetry functions for the male controls, male patients, female controls and female patients.

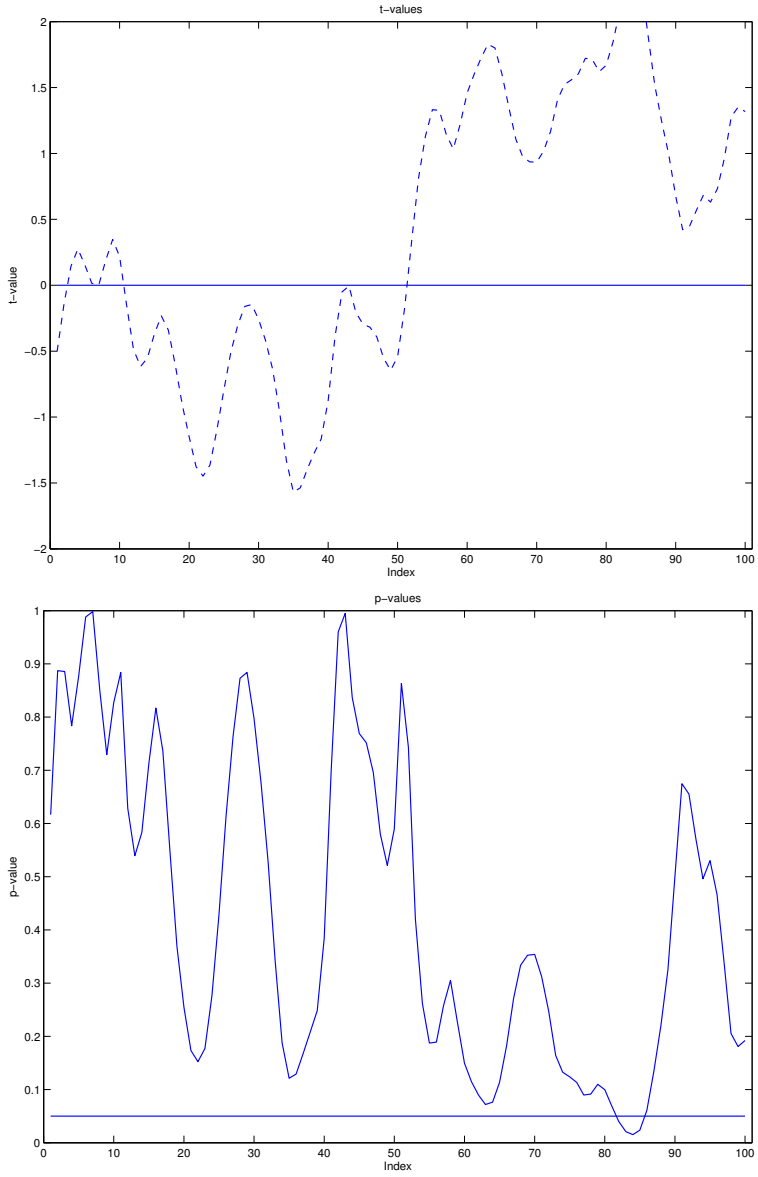

Figure 4.11: t-values (top) and p-values (bottom) for t-tests between control and patient groups at each slice. High t-values indicates greater rightward asymmetry in the control group. The value $p = 0.05$ is shown for reference.

would be the smaller of,

$$U_c = n_c n_s + \frac{n_c(n_c+1)}{2} - W_c,$$
$$U_s = n_c n_s + \frac{n_s(n_s+1)}{2} - W_s,$$

where W_c and W_s are the sums of ranks of the observations in the control and patient groups. If n_c and n_s are both large, as in this case, we could use the Mann-Whitney U test for large samples with the test statistic,

$$Z = \frac{U_c - (n_c n_s)/2)}{\sqrt{n_c n_s (n_c + n_s + 1)/12}}.$$

The use of non-parametric tests could be applicable to some applications of this symmetry analysis. Non-parametric tests, however, have less power than their parametric equivalents when the distributional assumption is valid, so we shall only consider parametric tests for this data set.

We carried out principal components analysis (PCA) on the pooled sample of $n = 68$ smoothed asymmetry functions. In Figure 4.12 we show the pooled mean and the loadings of the first 5 principal components (PCs). The loadings on the left of each picture are in the occipital region and those on the right are in the frontal region. From the plots of the PC loadings it seems clear from PC 1 that the main source in variability is in the occipital region. PC 2 shows a gradual increase in variability nearer the front. PC 3, however, highlights a more general twisting in the brain, and will best detect regions where the control group is more asymmetric. A plot of the PC score 3 with covariate information is given in Figure 4.13.

We consider fitting a linear regression model with response PC score 3 (see Arnold (1981) for definitions of concepts relating to linear models). The fitted parameters and statistical analysis are given in Table 4.1. We see that there are statistically significant differences in PC score 3 between patients and controls, but there is not a significant association with age or

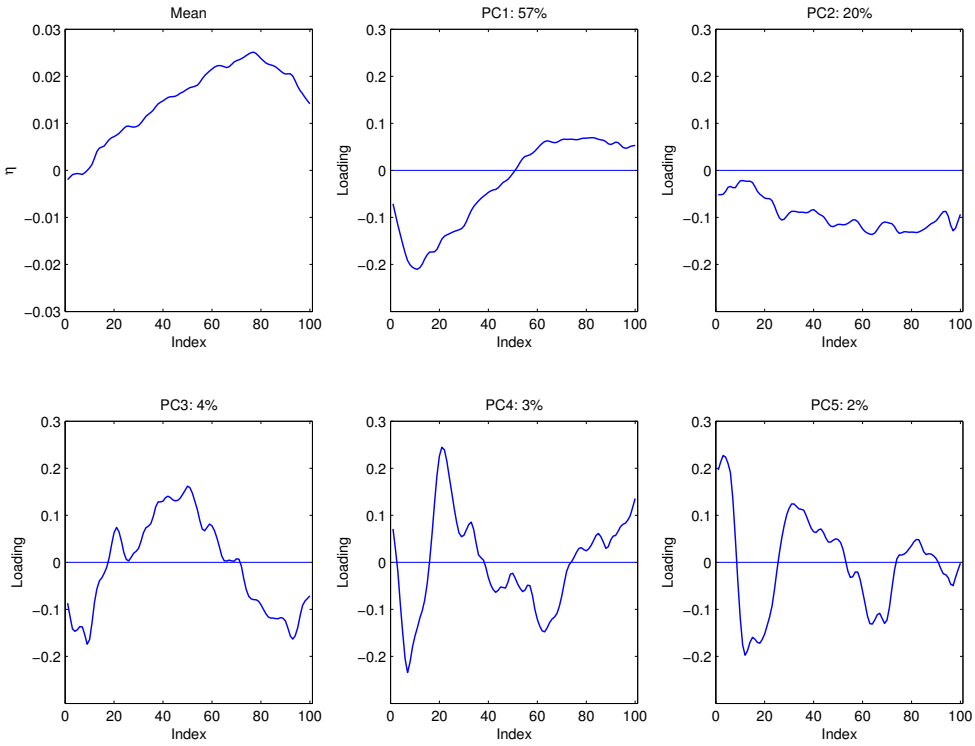

Figure 4.12: The mean (top left) and loadings for PCs 1-5.

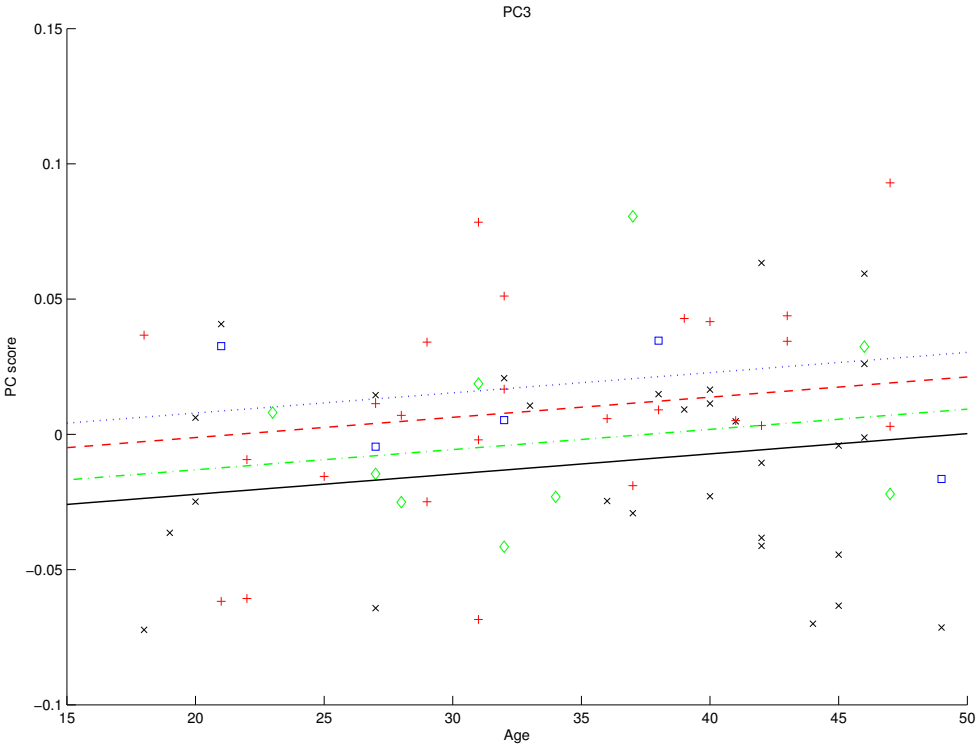

Figure 4.13: PC score 3 versus age. The colours represent the group and sex information: male controls (×, black), male patients (+, red), female controls (♢, green), female patients (□, blue).

sex. The results for the first twenty slices should be treated with care given the majority of the brain is below the AC-PC plane in this occipital region. The fitted PC 3 scores are positive for the patient group and negative for the control group. Applying the loading and fitted PC scores to the mean in the front 80 slices gives a more consistent rightward asymmetry in the patient group and a more extreme torque in the control group. This is reflected in the mean male asymmetry functions of Figure 4.9, where the control mean has greater rightward asymmetry in slices 55-100 and less rightward asymmetry in the remainder. The numbers of females in this study is small, so particular care should be taken with interpreting that torque is not associated with sex.

	Estimate	SE	t value	$Pr(>\|t\|)$
Intercept	-0.0672	0.0293	-2.2920	0.0252
group: Control (1) Patient (2)	0.0210	0.0093	2.2588	0.0273
Age	0.0007	0.0005	1.4417	0.1543
Sex: Male (1), Female (2)	0.0091	0.0113	0.8027	0.4251

Table 4.1: The fitted parameters and standard errors (SE) from a normal linear model with PC score 3 as the response.

4.7 Curved midplane analysis

In calculating the asymmetry function, we assumed that the join of the left and right brain hemispheres was the flat plane $\xi_z = \widehat{\xi}_z$, given by the maximum likelihood registration. In reality, inspection of the scans shows a tendency for the inter-hemispherical join to curve, especially at anterior and posterior extremities and, to a lesser extent, in superior regions. An alternative symmetry analysis, adjusting the slice volumes for this correction in the join's location would be preferable. For simplicity, we will translate the z-axis such that $\widehat{\xi}_z = 0$ for each scan.

To establish the location of the inter-hemispherical join, we apply the model for registering the entire midplane on a localised region of $\mathcal{M}$, and maximise the likelihood over ξ_z, keeping all other registration parameters fixed. The size of the localised region, $\mathcal{L}$, was chosen large enough to avoid detecting local symmetries not centred on the join but small enough to detect general movement of the join, keeping $\epsilon_M = 15$mm. The region, $\mathcal{L}$, was centred on locations at 5mm intervals in the $x-y$ plane and $\widehat{\xi}_z$ recorded, after a discrete grid search in unit steps, as the displacement from $\xi_z = 0$ at that location.

Let the maximum likelihood estimate of ξ_z for the jth individual at location (x, y) be z_{xyj}. Let $\mathcal{C}$ be the set of control scans and let $\mathcal{S}$ be the set of patient scans, then for each location we define,

$$\begin{aligned} \bar{z}_{xyc} &= \frac{1}{n_c} \sum_{j \in \mathcal{C}} z_{xyj}, \\ \bar{z}_{xys} &= \frac{1}{n_s} \sum_{j \in \mathcal{S}} z_{xyj}, \end{aligned}$$

to be the sample mean leftward displacements of the control and patient groups, respectively, where n_c and n_s are the number of scans in the control and patient groups. At each location, the displacements were analysed using t-tests of $\bar{z}_{xyc} - \bar{z}_{xys} = 0$. In the region between 25mm and 45mm anterior of the AC, and extending up to 20mm above the AC-PC axis, the join is significantly further to the left in the patient group, as seen in Figures 4.14 and 4.15. This might explain the reduced rightward asymmetry seen in male patients, compared to male controls, seen between slices 70 to 85 (approx) in Figure 4.9.

Suppose the region $\mathcal{L}$ was centred at the x, y co-ordinates, $t_1, ..., t_k$ in the midplane for a particular scan. Let T be the $(k \times 2)$ matrix of x, y co-ordinates and let Y, $(k \times 1)$, be the vector of corresponding maximum likelihood estimates of the midplane displacements, z_{xyj}. To incorporate a

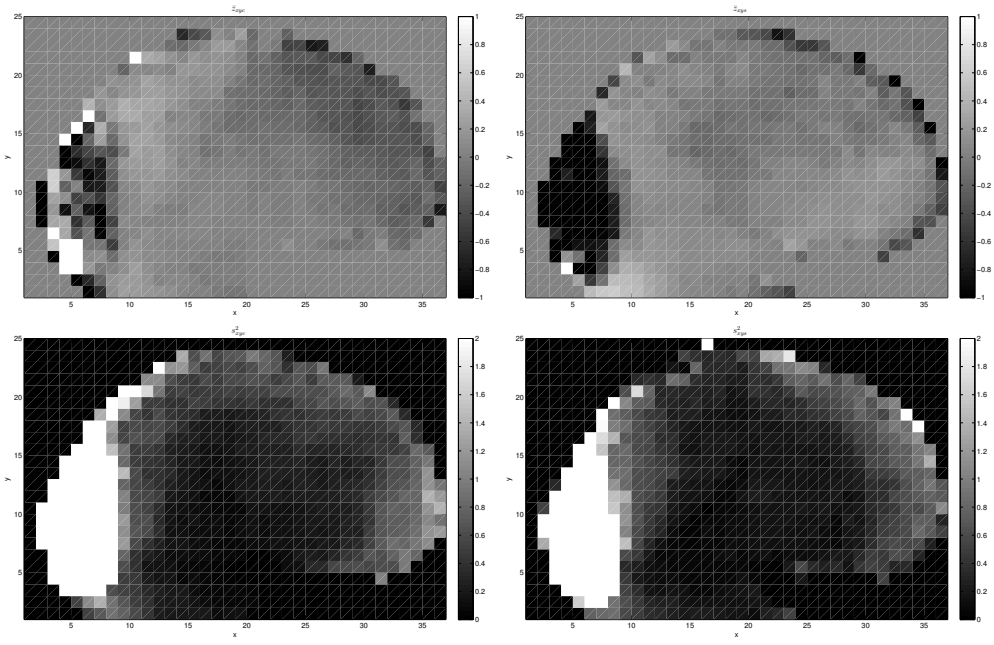

Figure 4.14: Top row: The mean of the inter-hemispherical join's displacement from the plane $\xi_z = 0$ at 5mm intervals in the $x-y$ plane for the control group (left) and the patient group (right), with darker (lighter) areas indicating a displacement to the 'right' ('left'). Bottom row: The variance of the inter-hemispherical join's displacement at each location for the control group (left) and patient group (right), with darker (lighter) areas indicating low (high) variance.

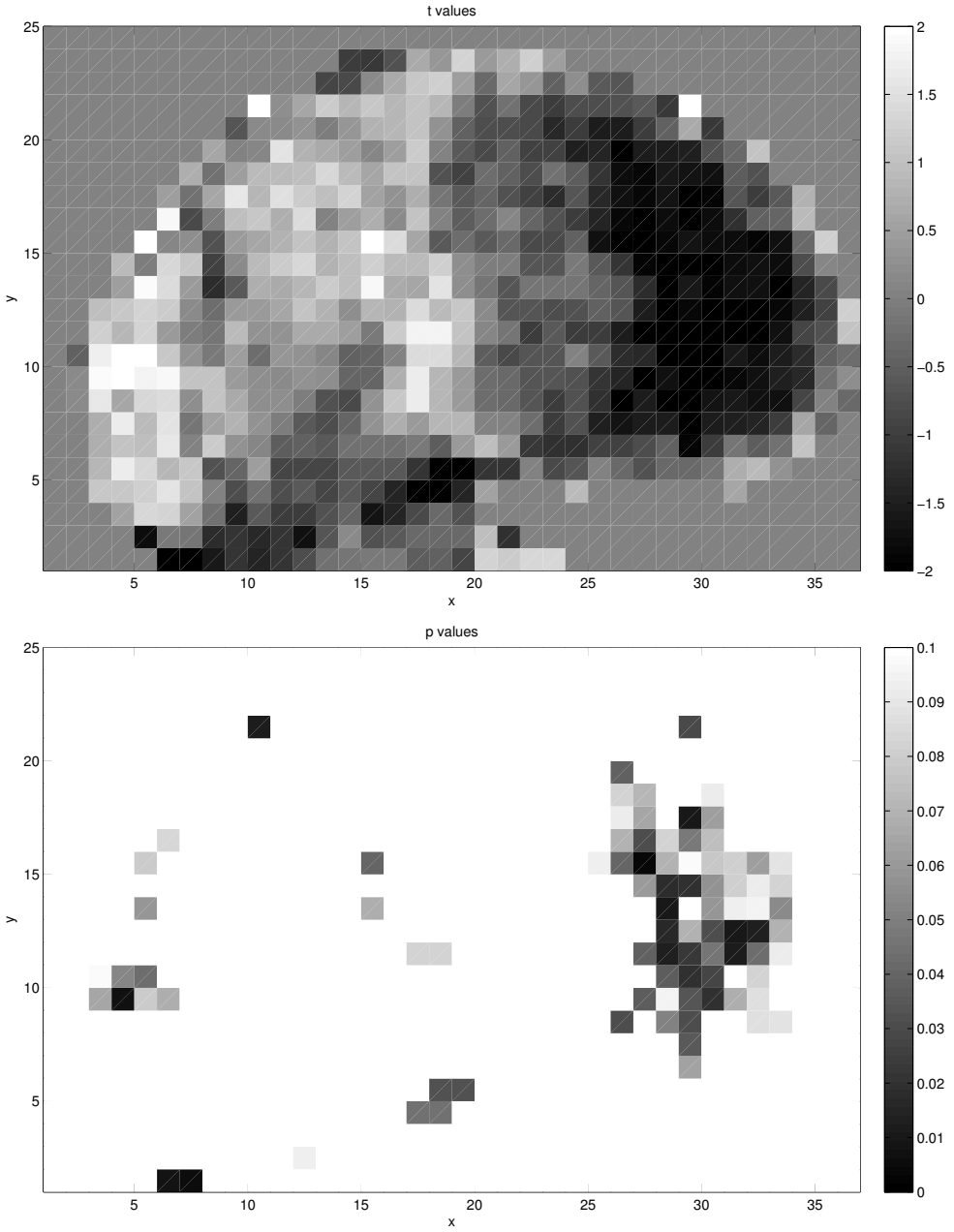

Figure 4.15: Top: t-values for the difference in the two groups, with darker (lighter) areas indicating the patient group further displaced to the left (right) than the control group. Bottom: p-values for each t-test thresholded at $p = 0.1$. Darker areas indicate higher significance.

curved inter-hemispherical join in the symmetry analysis, a curved midplane is fitted to the maximum likelihood estimates by means of a thin-plate spline for each scan to smooth the data. The fitted displacement at the x, y co-ordinate, t (2×1), is given by,

$$\Phi(t) = c + At + W^T s(t),$$

where c (1×1), A (1×2) and W $(k \times 1)$ are the solutions of,

$$\begin{bmatrix} S + \lambda I_k & 1_k & T \\ 1_k^T & 0 & 0 \\ T^T & 0 & 0 \end{bmatrix} \begin{bmatrix} W \\ c \\ A^T \end{bmatrix} = \begin{bmatrix} Y \\ 0 \\ 0 \end{bmatrix},$$

λ is a smoothing parameter, $s(t) = (\sigma(t - t_1), ..., \sigma(t - t_k))^T$, $(k \times 1)$, and $(S)_{ij} = \sigma(t_i - t_j)$ where,

$$\sigma(h) = \begin{cases} \|h^2\| \log(\|h\|), & \|h\| > 0, \\ 0, & \|h\| = 0. \end{cases}$$

The fitted thin-plate spline for one of the scans is shown in Figure 4.16. It clearly demonstrates that a curved midplane fits the inter-hemispherical join better than a flat midplane. The curved midplanes for eight of the female controls are shown in Figure 4.17. The variability between individuals is quite large, particularly towards the top and rear of the brain.

The symmetry analysis was repeated with a curved midplane, measuring the slice volume bounded by the cortical surface, the AC-PC plane, and the fitted inter-hemispherical join. The resulting mean smoothed asymmetry functions for each sub-group are shown in Figure 4.18. Although the principal components for this analysis were similar to those produced with a flat midplane, fitting a linear regression model with response PC score 3 only produced a weak significant difference between patients and controls. The fitted parameters and statistical analysis are given in Table 4.2. This

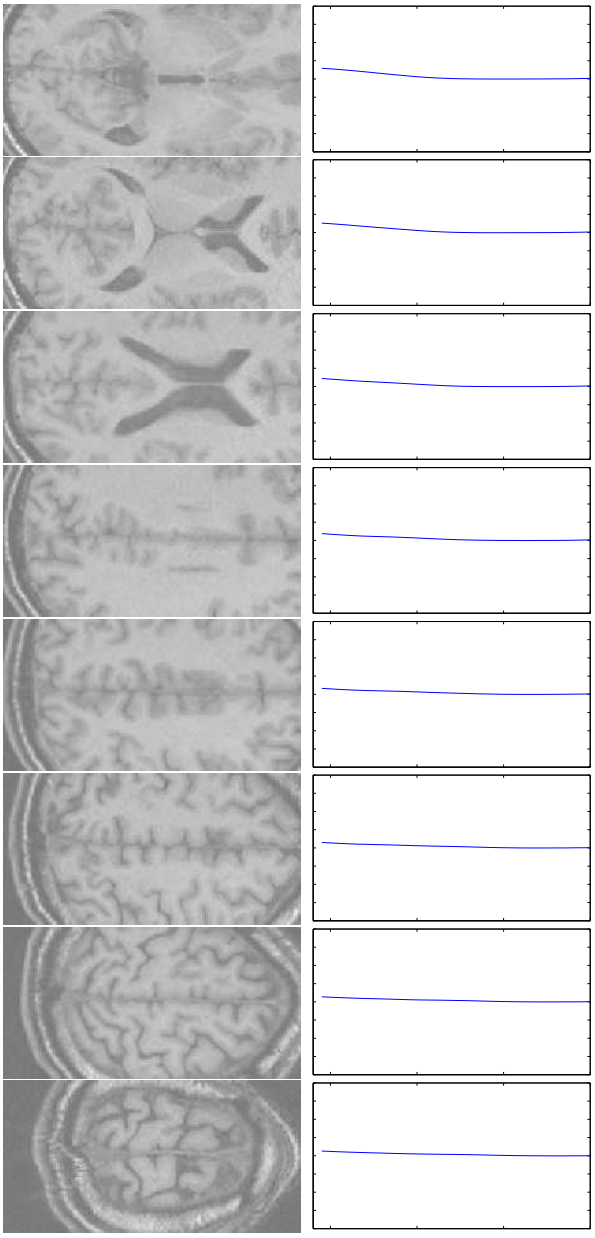

Figure 4.16: A comparison of the actual inter-hemispherical join and the fitted midplanes on axial slices at $y = 0, 10, 20, 30, 40, 50, 60, 70$ mm above the AC-PC line.

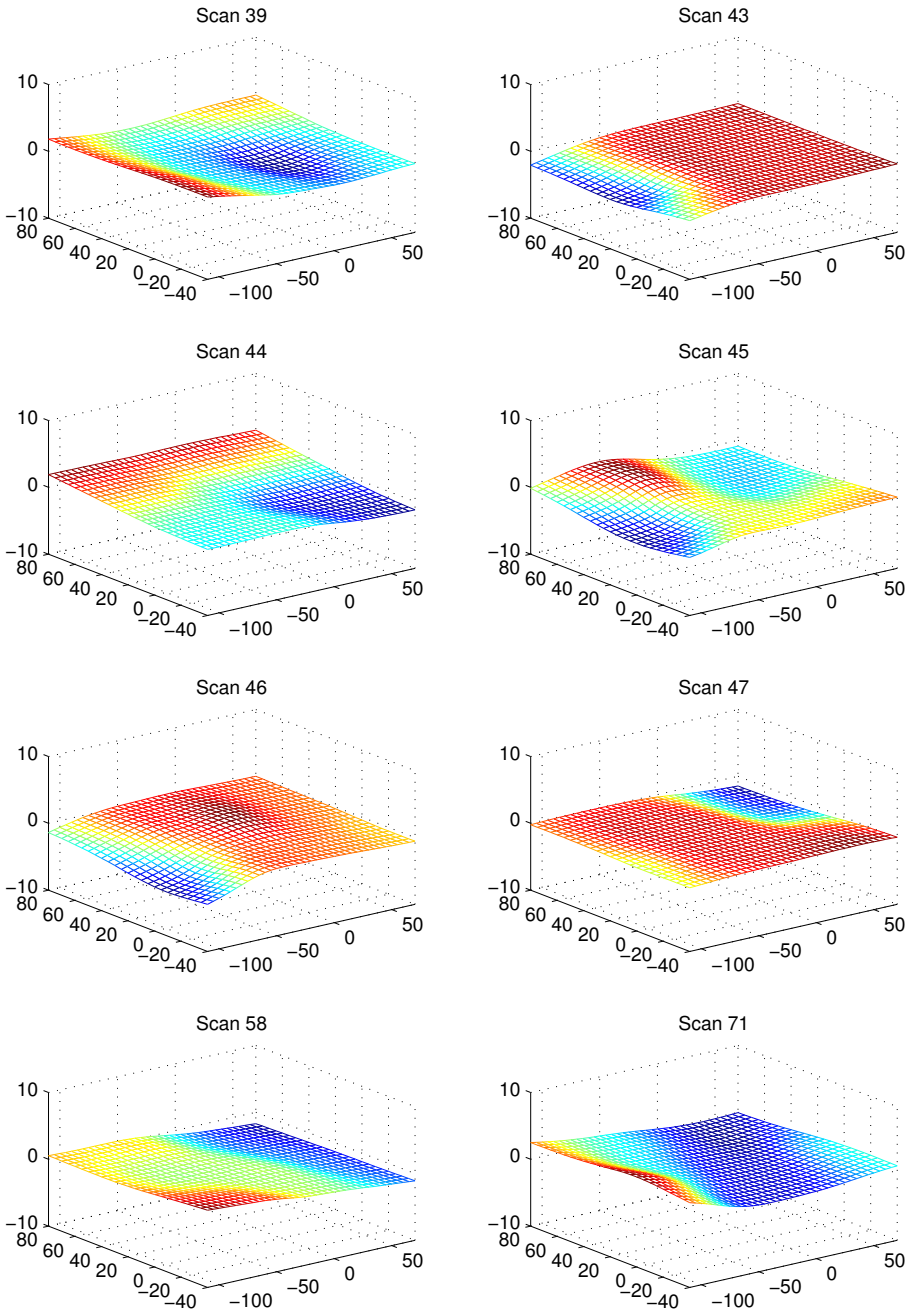

Figure 4.17: The fitted inter-hemispherical join for eight female controls. The vertical axis shows the leftward displacement in millimetres.

suggests that differences in brain torque between controls and patients can partly be explained by a curved midplane.

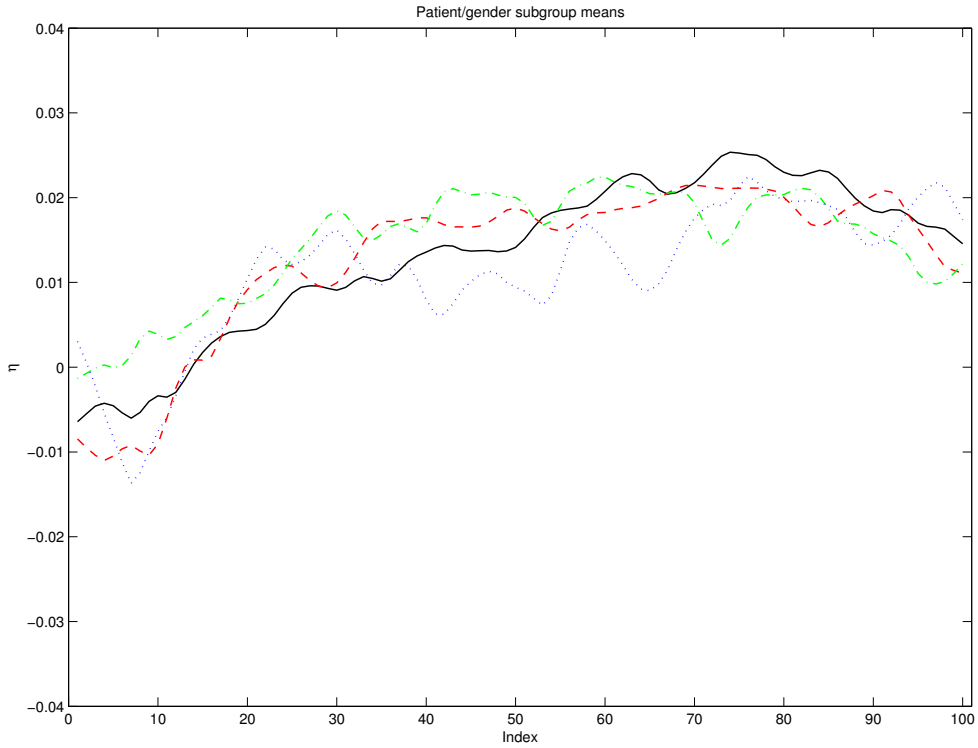

Figure 4.18: The mean smoothed asymmetry functions for the male controls (solid, black), male patients (dashed, red), female controls (dash/dot, green), female patients (dotted, blue).

4.8 Voxel-based morphometry

An alternative whole-brain technique for analysing brain shape is voxel-based morphometry (VBM). We apply this method to our data set of schizophrenia patients and healthy controls to compare the results with the asymmetry analysis presented in Section 4.6. VBM aims to characterise regional cerebral volume and tissue concentration differences in structural

	Estimate	SE	t value	$Pr(>\|t\|)$
Intercept	-0.0434	0.0358	-1.2124	0.2298
group: Control (1) Patient (2)	0.0198	0.0113	1.7494	0.0850
Age	0.0003	0.0006	0.5236	0.6023
Sex: Male (1), Female (2)	0.0027	0.0138	0.1958	0.8454

Table 4.2: The fitted parameters and standard errors (SE) from a normal linear model with PC score 3 as the response, using a curved midplane.

MRI scans (Ashburner and Friston, 2000). Neuroscientists routinely use VBM to identify differences in the composition of the brain between subjects. For example, Good *et al.* (2002) identifies brain regions with volume loss in patients with Alzheimer's disease and semantic dementia, with results similar to those obtained using traditional region-of-interest volume measurements. We implemented VBM using the Statistical Parametric Mapping 2 (SPM2) software, following the Good *et al.* (2001a,b) protocol for data preprocessing. Initially, each scan was manually translated and rotated such that the origin coincided with the anterior commissure (AC), with the axis aligned to pass through the posterior commissure (PC) and the inter-hemispherical boundary.

4.8.1 Data pre-processing

The data is preprocessed using the optimised VBM protocol. The protocol, as applied to our data set using SPM2, is outlined in Algorithm 4.8.1. The protocol is designed to extract the brain's grey matter from the rest of the image and register it to a study-specific grey matter template, rather than the generic whole-brain T1 template supplied within SPM2. The grey matter is of particular interest as it lies closest to the cortical surface. The registration is carried out using "normalisation" steps to transform the images from native (acquired) space to Talairach space (Talairach and Tournoux, 1988). Spatial normalisation aims to remove the majority of shape differ-

ences between an individual and the template. The registration includes a 12 parameter affine transformation (3 translations, 3 rotations, 3 scalings and 3 shears), see Ashburner *et al.* (1997) for details. Non-linear shape differences are modelled using a linear combination of smooth spatial basis functions (Ashburner and Friston, 1999). The normalisation uses prior knowledge of expected variability in brain shape in a Bayesian framework (Ashburner *et al.*, 1999, 2000). A mask is used so normalisation is based mainly on brain tissue.

Algorithm 4.8.1 ***Optimised VBM protocol***

Step 1. Creation of a separate grey matter template.

1a. Spatial Normalisation. *Each of the 68 scans are spatially normalised to the T1 template. The images are transformed from native (acquired) space to Talairach space by minimising the residual sum of squared differences in voxel values. The spatially normalised images are resliced with voxel size* $1 \times 1 \times 1mm^3$.

1b. Segmentation. *The normalised images are segmented into grey matter, white matter and cerebrospinal fluid (CSF). SPM2 uses a mixture of model cluster analysis and prior knowledge of tissue type locations in healthy brains, to identify voxel intensities and match particular tissue types.*

1c. Smoothing. *The normalised, segmented, grey matter images are smoothed using a 12mm full-width at half-maximum (FWHM) isotropic Gaussian kernel. The average of the 68 normalised, segmented, smoothed, grey matter images is calculated, which will be our study-specific template.*

Step 2. Estimation of normalisation parameters.

2a. Segmentation and extraction of brain image. *Each original image in native space is segmented into grey matter, white matter and CSF and all non-brain voxels removed.*

2b. Normalisation of grey matter images. *The extracted, segmented grey matter images are normalised to the grey matter template created in (1c), and the normalisation parameters recorded.*

Step 3. Creation of optimally normalised, segmented, smoothed images.

3a. Normalisation. *The normalisation parameters recorded in (2b) are reapplied to the whole image. This brings each image from native space into Talairach space. The spatially normalised images are resliced with voxel size* $1 \times 1 \times 1mm^3$.

3b. Segmentation and extraction of brain image. *Optimal segmentation of the optimally normalised images in Talairach space creates separate grey matter, white matter and CSF images. The brain extraction step is repeated to filter out any non-brain voxels.*

3c. Smoothing. *Each optimally normalised, segmented image is smoothed using a 12mm FWHM isotropic Gaussian kernel. These images are then used in the following statistical analysis.*

The aim of the VBM protocol is to accurately split (segment) the image into separate grey matter, white matter and CSF images, and to estimate the registration parameters necessary to transform the grey matter image to the template. However, these two aims confound each other because the normalisation step requires the image to already be segmented, but to segment the image in SPM2 requires the use of Bayesian priors that assume the image is already normalised. The optimised VBM protocol carries out an approximate segmentation before estimating the normalisation parameters (step 2), and then reverses the process, applying the estimated normalisation parameters before estimating the segmentation (step 3). Once in Talairach space the optimal segmentation is then available (Ashburner and Friston, 2000).

The optimised VBM protocol has two advantages over the simpler VBM protocol, which only implements step 1. Firstly, the normalisation parameters from (2b) are optimised compared to (1a), being based only on grey matter rather the whole image. Secondly, step 1 is used to create a study-specific template.

4.8.2 Statistical analysis of VBM data

The grey matter images from the optimised VBM protocol were analysed using SPM2. At each voxel, a linear model is constructed with the voxel values as the response, and the patient group as the covariate. The parameter values are estimated using ordinary least squares. The significance at each voxel is assessed using a t-test of the parameter value with r degrees of freedom, where r is the number of scans minus the rank of the design matrix. The standard error of the parameters are estimated using the residuals from the fitted model and the significance level is obtained using the theory of Gaussian random fields (Friston *et al.*, 1996) to correct for multiple comparisons. A significant difference in the voxels around the left superior temporal lobe was detected. These voxels are highlighted in Figure 4.19 with their t-statistic and can be compared to the location of the superior temporal lobe, labelled in Figure 4.20. The analysis shows controls have higher voxel values in this region, or appear lighter, than patients with schizophrenia. This implies the grey matter is significantly less concentrated, or has significantly less volume, in patients with schizophrenia.

The superior temporal lobe has been identified in other studies investigating schizophrenia as being significant. This region is generally responsible for language. Language development in humans has been aided by increasing hemispheric specialisation. Tim Crow (1997) has developed the theory that the lack of structural asymmetry seen in the brains of schizophrenia patients, is symptomatic of one hemisphere failing to have dominance for language. Therefore, it seems reasonable to expect to see differences in the superior temporal lobe. Further, Highley *et al.* (2001) cites other studies (Wible *et al.*, 1995; Highley *et al.*, 1999; McDonald *et al.*, 2000) reporting a reduction in the volume of the temporal lobe structures, many noting the reduction greatest on the left, as seen in our study.

Interestingly, Rajarethinam *et al.* (2004), have studied the superior temporal gyrus of the children of schizophrenia patients. This report found

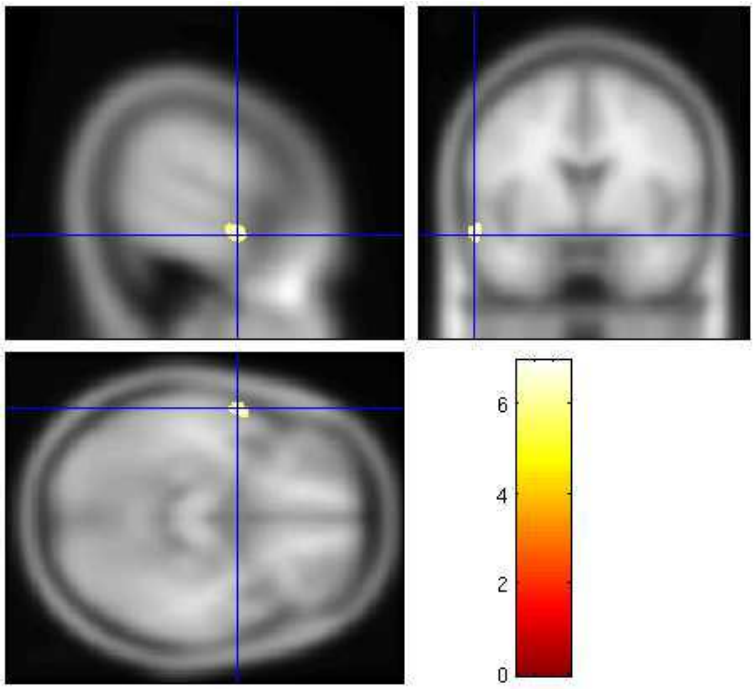

Figure 4.19: The t-statistic of the highlighted voxels in the superior temporal lobe show this region is significantly smaller in patients with schizophrenia.

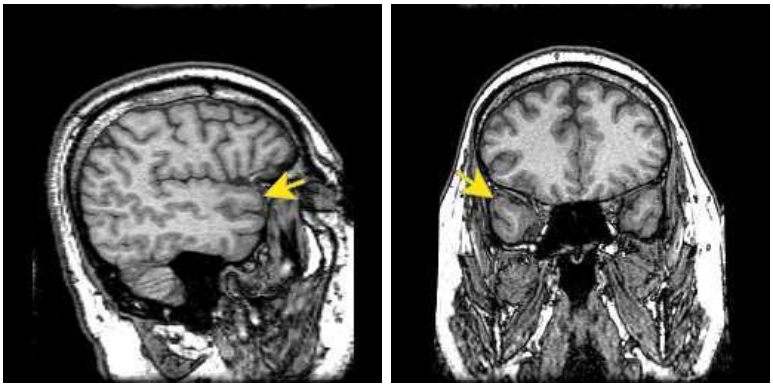

Figure 4.20: The location of the superior temporal lobe taken from the Whole Brain Atlas. Copyright © 1995-1999 Keith A. Johnson and J. Alex Becker.

them to be significantly smaller than controls, implying the genetic link of schizophrenia led to an increased risk of abnormality within the superior temporal gyrus. A review by McCarley *et al.* (1999) of studies on schizophrenia between 1987 and 1998 shows that of the 16 studies that considered the superior temporal gyrus, 13 discovered significant results. Like the VBM analysis above, 7 of the 16 studies considered grey matter only, and all of these studies showed significant results.

While VBM has produced results consistent with the literature, a debate remains regarding the validity of the technique. Bookstein (2001) argues that voxel-based statistical analysis is confounded by failures in local registration, so the technique is only valid away from edges. Ashburner and Friston (2001) disagree but acknowledge that significant differences detected by VBM may not necessarily result from differences in grey matter density, but could be caused by systematic differences in registration errors, motion artifacts, tissue classification or relative grey/white matter intensities.

In our experience, the optimised VBM algorithm itself is quite convoluted and it is undesirable that the normalisation and segmentation steps are mutually dependent. In addition, the normalisation step removes much of the shape difference through a series of non-linear warps. Ashburner *et al.* (1998) shows statistical analysis of the parameters corresponding to the non-linear warps, Deformation Based Morphometry (DBM), can detect differences in brain asymmetry between males and females but the results are difficult to interpret.

4.9 Discussion

The analysis of the MRI data has been pragmatic given the complexity of brain extraction, image registration and the geometry of the cortical surface. The approach taken has been sufficient to answer questions regarding our primary focus of large-scale differences such as symmetry. The use of BET to perform brain extraction is reasonable given the small uncertainty

in locating the cortical surface compared to the large variability in surface shape between individuals. Similarly, the small uncertainty in the registration justifies conducting the shape analysis conditional on the estimated cortical surface and registration. Alternative labelling functions are possible, for example, Fischl *et al.* (2001) uses a flattened cortical surface, but the labelling of the cortical surface used here is suitable for large-scale shape analysis given the lack of anatomical correspondence of brain features between individuals.

The results from the symmetry analysis show large-scale differences between controls and patients with schizophrenia in the frontal section of the brain, and less torque in patients. These differences are not apparent in the VBM analysis conducted using the SPM2 software, which only finds a difference in the superior temporal lobe. The difference in these findings can be explained when one considers the aims of each technique. The VBM technique only identifies small-scale tissue concentration differences because it filters out large-scale shape differences. The symmetry analysis only detects large-scale shape difference, such as symmetry, because it uses a rigid-body registration. Therefore, the two techniques complement each other.

An alternative analysis of this data set is presented in Brignell *et al.* (2006), which compares the lengths of radii between a central origin and the cortical surface for various angles. A smaller right frontal region in patients is observed in the data. Principal components analysis showed the brains to be "taller" in patients with schizophrenia. Independent components analysis showed this effect is associated with less torque in patients. Again, the method presented in this chapter complements these pseudo-landmark methods.

The introduction of a simple non-linear method for finding the inter-hemispherical join reduces the significance of our analysis, so care must be taken with interpreting the results. Clearly, torque is related to the twisting of the midplane as well as the cortical surface. An advantage of a

rigid-body registration over a non-linear normalisation is that the location of the midplane and other features can be easily examined, and the location of the midplane was shown to be significantly different between the two groups in this study. Therefore, we recommend shape analysis conditional on a rigid-body registration to examine large-scale shape differences such as those investigated in this study.

Chapter 5

Modelling haemodynamic response functions

5.1 Introduction

In response to a stimuli, the level of neuronal activation in particular parts of the brain can be studied through examining the change in blood flow to a particular region over a period of time. The neuronal activation is detected by changes in the blood oxygenation level dependent (BOLD) haemodynamic response signal that lead to an increase in MR signal intensity which correlates with the paradigm presented to the subject. In this study, a series of magnetic resonance images (MRI) was taken of the motor cortex. The motor cortex was chosen because many functional MRI (fMRI) studies often involve a physical response to a visual or audio stimulus, for example selecting an option from a keypad. It is therefore vital, for a wide range of applications, that the time course of the BOLD signal change associated with neuronal activity, the haemodynamic response function (HRF), in the motor cortex can be modelled.

We will fully consider existing techniques for examining fMRI data in Section 5.8, but most methods fall into two classes, either model-driven or

data-driven. Data-driven methods, such as independent component analysis, can identify areas of the MR image where the grey-level fluctuates at the same frequency as the stimuli. Independent components analysis can be implemented, for example, within the MELODIC (Multivariate Exploratory Linear Optimised Decomposition into Independent Components) software (Beckmann and Smith, 2004). Another leading piece of software, SPM2 (Statistical Parametric Mapping 2) introduced in Section 4.8, is model-driven and uses a linear model with a typical haemodynamic response basis function, shown in Figure 5.1, as a covariate. The general nature of the basis function derives from prior empirical evidence, but it assumes the exact form can be prescribed by an analytic function. For example, SPM2's basis function is the discretised summation of two curves taken from a gamma probability density function. We wish to relax this assumption and allow the data to specify the form of the response function while keeping within the structure of the model-driven framework.

In a typical study, a volunteer is asked to repeat a task several times because the BOLD signal change is of the order of 2.5% and the signal-to-noise ratio is low. Currently, most methods for analysing standard fMRI experiments assume that the response to each stimuli is dependent only on location and time since the last stimulus, and that each trial is independent. In truth, this is probably a false assumption. For example, behavioural studies have shown "warm up" (Eysenck and Frith, 1977) and "carry over" (Ward *et al.*, 2001) effects, where peak performance is not achieved when starting a new task from rest or when switching immediately from one task to another. The ability to test the assumption that each trial is independent, however, has only been available recently due to the advent of scanners with stronger magnetic fields, such as the ultra-high field of 7T. This increases the BOLD signal change and the signal to noise ratio.

Examining this single-trial variability at 7T is a topic of current research. Debener *et al.* (2005) have studied electroencephalogram data (recordings of the small electrical signals caused by activation) and noted that if a subject

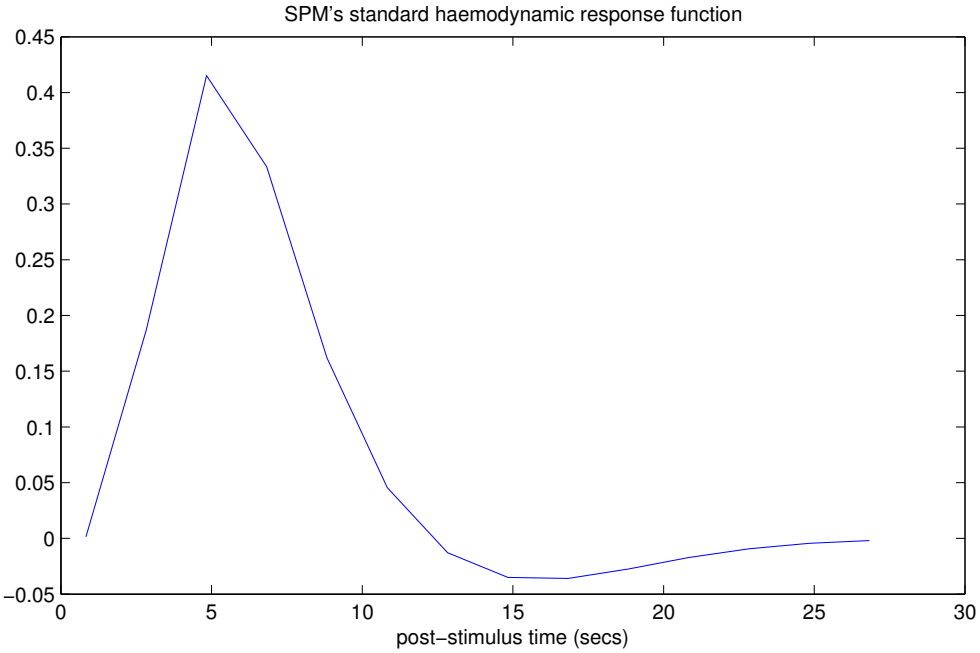

Figure 5.1: SPM2's standard haemodynamic response function evaluated at two second intervals.

makes a mistake in one trial, there is a systematic relationship to the behaviour in the next trial. fMRI data has been studied by Duann *et al.* (2002) using independent components analysis to "reveal dramatic and unforeseen haemodynamic response variations not apparent to researchers analysing their data with event-related response averaging and fixed haemodynamic response templates" in the visual cortex, across a number of subjects. A variation of the SPM2 model-driven approach has been suggested by Lu *et al.* (2005) but they still incorporate prior knowledge by only allowing the haemodynamic response to fluctuate in a neighbourhood of the original basis function. Despite these initial successes in this new field, a model-driven approach, free of *a priori* assumptions, has not been previously advocated.

In our study, gradient echo planar images (EPI) of a volunteer's motor cortex were taken at the Sir Peter Mansfield Magnetic Resonance Centre

(SPMMRC) on the Philips 7T scanner. A dynamic scan was taken at 2 second intervals for 5 minutes and each dynamic scan was composed of 12 slices, each 3mm thick with a 0.7mm slice gap, taken sequentially at 1/6 second intervals. Each slice is composed of 64 x 64 pixels of size 1mm x 1mm. A total of ten visual stimuli were presented at 28.25 second intervals, this cued the volunteer to press a button either once or five times as indicated. The different responses were requested alternately, starting with a single button press.

Our experiment paradigm, with single and multiple button presses, also raises interesting questions regarding the relationship between neuronal activity and haemodynamic response. It is well documented that the relationship is non-linear, but predicting the change in response to a longer period of activation, for example, is less well understood (Mechelli *et al*, 2001). SPM2 adjusts its basis function by convolving the period of neuronal activity with its standard haemodynamic response function, using the Volterra series (Friston *et al.*, 1998). Independently, Buxton *et al.* (1998) developed the balloon/Windkessel dynamical non-linear model of how the haemodynamic response is influenced by the underlying physical changes in blood vessel volume and deoxyhemoglobin content. The model suggests that increased blood flow inflates a venous "balloon", diluting and expelling deoxygenated blood causing an increase in the BOLD signal. As the flow decreases, the balloon deflates reducing the discharge and increasing the concentration of deoxygenated blood, causing the post-stimulus undershoot. Friston *et al.* (2000) showed that the Volterra and balloon/Windkessel models were consistent, in that the Volterra kernels which best represented those derived from empirical evidence, also had biologically plausible estimates for the balloon/Windkessel model parameters. In this study, we seek to develop a model that distinguishes between different responses, caused by different neuronal activity, without making assumptions regarding the underlying non-linear relationship.

5.2 Image preprocessing

Usually, fMRI images are subjected to three preprocessing steps, namely slice-timing, realignment and smoothing, before areas of activation can be examined. Slice-timing corrects for the time difference between each slice being recorded within each dynamic scan and transforms all the data within each scan to a single time point. More details are given below. Realignment corrects for the small movement in the volunteers position between each dynamic scan by translating and rotating each scan, in relation to the first, with a six-parameter rigid-body transformation. Finally, the images are smoothed to reduce noise and to allow for any remaining imperfections in anatomical alignment. The size of the spatial smoothing kernel should match the size of any potential activation we wish to detect. Typically, this is 1.5 times the voxel size, so our images were smoothed with an isotropic Gaussian kernel of 1.5mm full-width half-maximum (FWHM).

Slice-timing corrects for the difference in acquisition times of each slice within a dynamic scan. Modelling the time series from each slice as a linear combination of sinusoids of different phases and frequencies, the data is shifted forwards or backwards in time by effectively adding a constant to the phase of every frequency. Conventionally, each slice is shifted so that the time-series has the values that would have been obtained had the slice been acquired at the same moment as a reference slice. Typically, the reference slice is chosen to be the middle slice and all other slices in every scan are corrected to this reference slice.

A schematic diagram of slice-timing can be seen in Figure 5.2. In this simplified experiment paradigm, each dynamic scan is composed of three slices (acquired at 1 second intervals), with a 3 second inter-scan interval. The stimuli is presented every 5.5 seconds. Let slice 2 be the reference slice, then with ordinary slice timing, slice one is lagged back 1 second and slice three is brought forward 1 second. So the reference time points are at 1, 4, 7 and 10 seconds. However, the post-stimulus times are 1 and 4 seconds for

the first trial, but 1.5 and 4.5 seconds for the second trial. This negates the ability to examine single-trial variability.

For this reason, we implement slice-timing with a different time shift for each slice/trial combination so that all slices are corrected to the same post-stimulus time points. Returning to Figure 5.2, with slice/trial-timing the slices are lagged by 1, 0 and -1 seconds in the first trial and 0.5, -0.5 and -1.5 seconds in the second trial. This makes the reference time points 1, 4, 6.5 and 9.5 seconds, maintaining the post-stimulus times of 1 and 4 seconds in both trials.

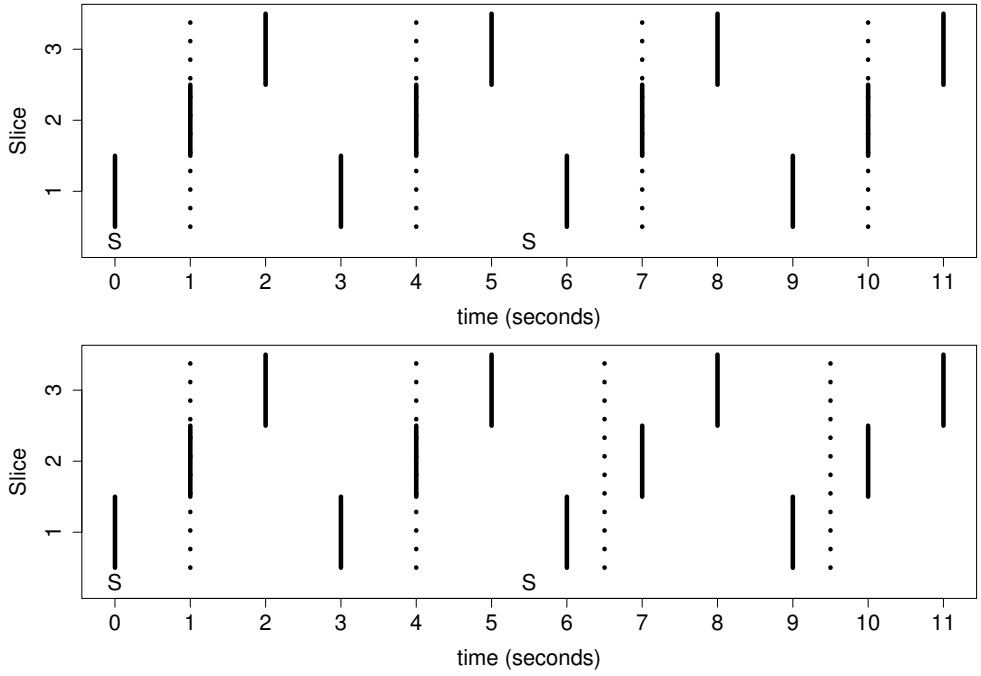

Figure 5.2: Schematic diagrams showing the difference between slice timing (top) and slice/trial timing (bottom) for a simplified paradigm with a dynamic scan of 3 slices and a 5.5 second interval between stimuli. Slice acquisition times are shown with a solid line, and the corrected timings with a dotted line. The stimuli were given at times marked "S".

Let T be the number of post-stimulus time points, and E be the number

of events or trials. If the number of scans, N, exceeds $T \times E$, then we discard those scans requiring a large time lag to bring them into alignment with a reference time point. The post-stimulus times chosen for our study were $\frac{5}{6}$, $2\frac{5}{6}$, $4\frac{5}{6}$, ..., $26\frac{5}{6}$ seconds, as these were the default SPM2 post-stimulus times during the first trial. Therefore, $T = 14$, and there were $E = 10$ events, so we assume $N = T \times E = 140$.

Let X^* be an $N \times q$ design matrix composed of covariate information, with the mean of each column being zero. Possible covariates might include the realignment parameters from preprocessing or physiological data such as pulse and respiratory rate as movement, blood flow and oxygen levels all affect the grey-level recorded in an MR image. In our study, we let $q = 6$ using the six realignment parameter vectors as covariates, as physiological data was unavailable.

Let $\widetilde{Y}_{ij}$ be the grey-level at the ith voxel at the jth time point, $j = 1, ..., N = T \times E$. The design matrix and data are both subjected to temporal filtering, a standard SPM2 preprocessing step, to remove low-frequency confounds. A matrix, K, is formed of the first k non-constant basis functions of a one-dimensional discrete cosine transform, of length N. Let $t = 2$ seconds be the inter-scan time interval and let $c = 128$ seconds be a cutoff time, then k is $2(N * t)/c$ rounded down to the nearest integer. The filtered design matrix and data are then $X = X^* - K(K^T X^*)$ and $Y_i^* = \widetilde{Y}_i - K(K^T \widetilde{Y}_i)$. We set the baseline signal for each voxel to be zero by subtracting the voxel's mean grey-level value, i.e. $Y_{ij} = Y_{ij}^* - \sum_{j=1}^{N} Y_{ij}^*/N$. Therefore, Y_{ij} is the haemodynamic response at the ith voxel that we wish to model. Voxels at locations exterior to the brain are automatically masked out by SPM2 and not included in the analysis.

5.3 The model

We fit a linear model at each voxel to the data and design matrix described in Section 5.2. The pre-processed data for two voxels are shown in Figure

5.3. Clearly, not all voxels within the brain are activated in response to the stimuli and we assume no prior information regarding which voxels are active. Therefore, we formulate a multivariate Gaussian model for each of the two classes of voxel (active and inactive). For active voxels we model the response Y_i with mean $\mu_1 = \beta_i\widetilde{\mu} + Xb_i$ and covariance matrix $\Sigma_1 = \Sigma_E \otimes \Sigma_T$, and for inactive voxels let mean $\mu_2 = Xb_i$ and covariance matrix $\Sigma_2 = \sigma^2 I_N$. Dividing the voxels into the two groups to maximise the likelihood of this model is equivalent to minimising $|\widehat{\Sigma}_1|^{V_1}|\widehat{\Sigma}_2|^{V_2}$, where Σ_i is the covariance matrix of the V_i voxels in group $i = 1, 2$. However, this would require calculating the maximum likelihood estimate (MLE) for 2^{V-1} combinations, where V is the total number of voxels, which is computationally impractical.

Instead, all the model parameters will be estimated from the data by maximising the likelihood using an Expectation-Maximisation (EM) algorithm. The vector, $\widetilde{\mu}$, represents a "standard" haemodynamic response to a stimuli. We expect the response at each stimuli to be similar, so we constrain $\widetilde{\mu} = 1_E \otimes \mu$, where μ is a vector of length T, to detect true responses rather than noise. The scale parameter, β_i, denotes the scale of the response at the ith voxel. To remove the arbitrary scaling of these parameters we constrain $\|\mu\| = 1$. Lastly, b_i is the parameter vector for the design matrix, and Σ_E and Σ_T denote variability between and within events, respectively, where the subscript indicates the size of the matrix.

The log likelihood of the Gaussian mixture model,

$$f(Y) = pf_1(Y) + (1-p)f_2(Y),$$

where $f_1 \sim N(\mu_1, \Sigma_1)$ and $f_2 \sim N(\mu_2, \Sigma_2)$ is,

$$l(p, \mu_1, \mu_2, \Sigma_1, \Sigma_2|Y) = \sum_{i=1}^{V} \log\left[pf_1(Y_i|\mu_1, \Sigma_1) + (1-p)f_2(Y_i|\mu_2, \Sigma_2)\right], \quad (5.1)$$

where V is the number of voxels.

As noted in Chapter 4, image preprocessing and modelling assumptions

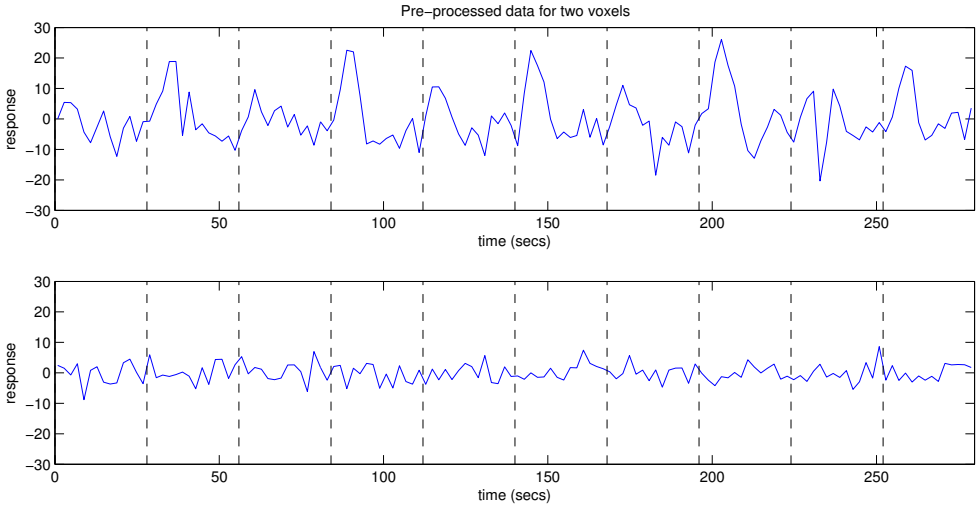

Figure 5.3: The pre-processed data from two voxels. One shows signs of activation (top), whilst the other is mainly noise (bottom). The stimuli was presented at the times marked with dotted lines.

influence the resulting parameter estimates. In this chapter, we base our assumptions entirely on the models used in SPM2, with the exception of modifications outlined above, for ease of comparison.

5.4 Parameter estimation using the EM algorithm

Maximising the likelihood in Equation (5.1) is non-trivial, although a solution exists if we constrain $\Sigma_1 = \Sigma_2$ (Day, 1969). Proceeding with an EM algorithm, we augment the data with latent variables, which are effectively missing data. Let Z_i be a Bernoulli random variable with $P(Z_i = 1) = p$. Let $Y_i \sim f_1$ if $Z_i = 1$ and $Y_i \sim f_2$ if $Z_i = 0$. The joint density of (Z_i, Y_i) is $f(z_i, Y_i) = [pf_1(Y_i)]^{z_i} [(1-p)f_2(Y_i)]^{1-z_i}$, and the log likelihood for the data

is,

$$l(p, \mu_1, \mu_2, \Sigma_1, \Sigma_2 | x, Y) = \sum_{i=1}^{V} z_i \log \left[p f_1(Y_i|\mu_1, \Sigma_1) \right] + (1 - z_i) \log \left[(1-p) f_2(Y_i|\mu_2, \Sigma_2) \right].$$

We obtain maximum likelihood estimates of the model parameters by taking expectations over the missing data and maximising the likelihood over the model parameters.

Estimating the model parameters is an iterative procedure with each iteration composed of expectation and maximisation steps. The expectation step estimates the values of the latent variables and then the conditional log-likelihood is maximised by differentiation. Let $\Theta' = (p, \mu_1, \mu_2, \Sigma_1, \Sigma_2)$ be the current parameter estimates and let Θ denote the parameter values at the next iteration then,

$$\begin{aligned} Q(\Theta|\Theta', Y) &= \mathbb{E}\left[l(\Theta|Z, Y) | Y, \Theta' \right], \\ &= \sum_{i=1}^{V} \mathbb{E}(Z_i|Y, \Theta') \log \left[\frac{p f_1(Y_i|\Theta)}{(1-p) f_2(Y_i|\Theta)} \right] + \log \left[(1-p) f_2(Y_i|\Theta) \right], \end{aligned}$$

where Q has its usual EM interpretation and,

$$\mathbb{E}(Z_i|Y, \Theta') = f(Z_i = 1|Y, \Theta') = \frac{p f_1(Y_i|\Theta)}{p f_1(Y_i|\Theta) + (1-p) f_2(Y_i|\Theta)} = \widetilde{p}_i.$$

Due to the small values of f_1 and f_2, it is computationally more feasible to calculate $\widetilde{p}_i = 1/(1 + e^c)$, where $c = \log(1-p) - \log(p) + \log(f_2) - \log(f_1)$. Recalling $\Sigma_1 = \Sigma_E \otimes \Sigma_T$ and $\Sigma_2 = \sigma^2 I_N$, we can write Q in terms of our model parameters,

$$Q(\Theta|\Theta', Y) = \sum_{i=1}^{V} \widetilde{p}_i \log \left[p f_1(Y_i|\mu_1, \Sigma_1) \right] + (1 - \widetilde{p}_i) \log \left[(1-p) f_2(Y_i|\mu_2, \Sigma_2) \right],$$

$$
\begin{aligned}
= \; & const + \sum_{i=1}^{V} \Big\{ \widetilde{p}_i \log(p) - \frac{\widetilde{p}_i E}{2} \log|\Sigma_T| - \frac{\widetilde{p}_i T}{2} \log|\Sigma_E| \\
& -\frac{\widetilde{p}_i}{2}(Y_i - \beta_i\widetilde{\mu} - Xb_i)^T\Sigma_1^{-1}(Y_i - \beta_i\widetilde{\mu} - Xb_i) \\
& +(1-\widetilde{p}_i)\log(1-p) - \frac{(1-\widetilde{p}_i)ET}{2}\log(\sigma^2) \\
& -\frac{(1-\widetilde{p}_i)}{2\sigma^2}(Y_i - Xb_i)^T(Y_i - Xb_i)\Big\}.
\end{aligned}
$$

We now estimate the model parameters in turn. For each, we maximise Q conditional on the current values of the other parameters by setting the derivative of Q with respect to the parameter equal to zero and solving for the parameter.

$$
\begin{aligned}
\frac{\partial Q}{\partial p} = \sum_{i=1}^{V}\left\{\frac{\widetilde{p}_i}{p} - \frac{(1-\widetilde{p}_i)}{(1-p)}\right\} &= 0, \\
\implies \qquad (1-p)\sum_{i=1}^{V}\widetilde{p}_i &= p\sum_{i=1}^{V}(1-\widetilde{p}_i), \\
\implies \qquad \widehat{p} &= \frac{1}{V}\sum_{i=1}^{V}\widetilde{p}_i. \qquad (5.2)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial Q}{\partial \beta_i} = -\widetilde{p}_i\left(\beta_i\widetilde{\mu}^T\Sigma_1^{-1}\widetilde{\mu} - \widetilde{\mu}^T\Sigma_1^{-1}(Y_i - Xb_i)\right) &= 0, \\
\implies \qquad \widetilde{p}_i\widetilde{\mu}^T\Sigma_1^{-1}\left(\beta_i\widetilde{\mu} - (Y_i - Xb_i)\right) &= 0.
\end{aligned}
$$

Therefore,

$$
\widehat{\beta}_i = (\widetilde{\mu}^T\widetilde{\mu})^{-1}\widetilde{\mu}^T(Y_i - Xb_i). \qquad (5.3)
$$

For the parameter vector,

$$
\frac{\partial Q}{\partial b_i} = \widetilde{p}_i X^T\Sigma_1^{-1}\left(Xb_i - (Y_i - \beta_i\widetilde{\mu})\right) + (1-\widetilde{p}_i)X^T\Sigma_2^{-1}(Xb_i - Y_i) = 0.
$$

Therefore,

$$\widehat{b}_i = (A^T A)^{-1} A^T \left(\widetilde{p}_i X^T \Sigma_1^{-1} (Y_i - \beta_i \widetilde{\mu}) + (1 - \widetilde{p}_i) X^T \Sigma_2^{-1} Y_i \right), \qquad (5.4)$$

where $A = \widetilde{p}_i X^T \Sigma_1^{-1} X + (1 - \widetilde{p}_i) X^T \Sigma_2^{-1} X$. Let Y_{ij} be the response from voxel i at event j, X_j the rows of X corresponding to event j, and let e_{jk} be the jkth entry of Σ_E^{-1} then differentiating Q with respect to μ gives,

$$\begin{aligned} \frac{\partial}{\partial \mu} \left\{ \sum_{i=1}^{V} \sum_{j=1}^{E} \sum_{k=1}^{E} \widetilde{p}_i e_{jk} (Y_{ij} - \beta_i \mu - X_j b_i) \Sigma_T^{-1} (Y_{ik} - \beta_i \mu - X_k b_i) \right\} &= 0, \\ \sum_{i=1}^{V} \sum_{j=1}^{E} \sum_{k=1}^{E} \widetilde{p}_i e_{jk} \Sigma_T^{-1} \left(\beta_i^2 \mu - \beta_i (Y_{ij} - X_j b_i) \right) &= 0. \end{aligned}$$

Therefore,

$$\widehat{\mu} = \frac{\left(\sum_i \sum_j \sum_k \widetilde{p}_i e_{jk} \beta_i (Y_{ij} - X_j b_i) \right)}{\left(\sum_i \widetilde{p}_i \beta_i^2 \right) \left(\sum_j \sum_k e_{jk} \right)}. \qquad (5.5)$$

The relative sizes of β_i and $\widetilde{\mu} = 1_E \otimes \mu$ are arbitrary so we can artificially rescale them at each iteration such that $\|\mu\| = 1$ without changing the value of the likelihood. Let R_i be a $T \times E$ matrix of residuals, where the jth column is $Y_{ij} - \beta_i \mu - X_j b_i$, and let $S_i = R_i \Sigma_E^{-1} R_i^T$. Then,

$$\frac{\partial Q}{\partial \Sigma_T^{-1}} = \frac{\partial}{\partial \Sigma_T^{-1}} \left\{ \sum_{i=1}^{V} \widetilde{p}_i E \log |\Sigma_T^{-1}| - \widetilde{p}_i \mathrm{tr} \left[\Sigma_T^{-1} S_i \right] \right\} = 0.$$

Given, $\partial \log |\Sigma_T^{-1}| / \partial \Sigma_T^{-1} = 2\Sigma_T - \mathrm{Diag}(\Sigma_T)$, and $\partial \mathrm{tr}(\Sigma_T^{-1} S) / \partial \Sigma_T^{-1} = 2S - \mathrm{Diag}(S)$, if Σ_T and S are symmetric, then,

$$\sum_{i=1}^{V} \left\{ \widetilde{p}_i E \left[2\Sigma_T - \mathrm{Diag}(\Sigma_T) \right] - \widetilde{p}_i \left[2S_i - \mathrm{Diag}(S_i) \right] \right\} = 0,$$

$$[2\Sigma_T - \mathrm{Diag}(\Sigma_T)] - \frac{1}{E\sum_{i=1}^{V}\widetilde{p}_i}\sum_{i=1}^{V}\widetilde{p}_i\,[2S_i - \mathrm{Diag}(S_i)] \;=\; 0.$$

Let $M = \Sigma_T - \frac{\sum_i \widetilde{p}_i S_i}{E\sum_i \widetilde{p}_i}$, then we require $2M - \mathrm{Diag}(M) = 0$. Hence $M = 0$ and,

$$\widehat{\Sigma}_T = \frac{1}{E\sum_{i=1}^{V}\widetilde{p}_i}\sum_{i=1}^{V}\widetilde{p}_i S_i. \tag{5.6}$$

By analogy, let $T_i = R_i^T\Sigma_T^{-1}R_i$ then,

$$\widehat{\Sigma}_E = \frac{1}{T\sum_{i=1}^{V}\widetilde{p}_i}\sum_{i=1}^{V}\widetilde{p}_i T_i. \tag{5.7}$$

Finally,

$$\frac{\partial Q}{\partial\sigma^2} = \sum_{i=1}^{V}\left\{\frac{-(1-\widetilde{p}_i)ET}{2\sigma^2} + \frac{(1-\widetilde{p}_i)}{2\sigma^4}(Y_i - Xb_i)^T(Y_i - Xb_i)\right\} \;=\; 0.$$

Hence,

$$\widehat{\sigma}^2 = \frac{1}{ET\sum_{i=1}^{V}(1-\widetilde{p}_i)}\sum_{i=1}^{V}(1-\widetilde{p}_i)(Y_i - Xb_i)^T(Y_i - Xb_i). \tag{5.8}$$

EM algorithms are sensitive to the starting estimates of the model parameters. Suitable starting estimates for this algorithm are obtained by initially estimating a reduced version of the model, $Y_i \sim N(\beta_i\widetilde{\mu}+Xb_i, \Sigma_E\otimes\Sigma_T)$, where all voxels are assumed to be active. The model parameters can be found with a simpler version of the algorithm presented above with no expectation step. For the reduced model, the starting mean response, μ, is taken as the standard response function from SPM2 and the covariance is initially assumed to be isotropic. Statistical tests, outlined below, of the resulting β_i parameters provide an initial classification of active and inactive

voxels in the EM algorithm. This classification is used to produce starting estimates of the proportion of active voxels, p, and the covariance matrices for the two groups, Σ_1 and Σ_2. The EM algorithm takes approximately 50 iterations to converge for the model and data used in this study. This only takes a few minutes to compute but it will require more iterations if the amount of data or covariates is increased.

5.5 Model evaluation

The model proposed above is obviously more complex than the simpler linear model currently implemented by SPM2. In this section we compare 5 models of ranging complexity using the deviance and Akaike information criterion (AIC), calculated using the log-likelihood, $\log L$, and the number of model parameters, p. The first two linear models we consider are,

$$\begin{aligned} &\text{Model 1:} \quad Y_i \sim N_{TE}(\beta_i \mu_s + X b_i, I_{TE}); \\ &\text{Model 2:} \quad Y_i \sim N_{TE}(\beta_i \widetilde{\mu} + X b_i, I_{TE}); \end{aligned}$$

where μ_s is the HRF used by SPM2 and $\widetilde{\mu}$ is an HRF estimated from the data. In both models the shape of the HRF is constant for each voxel and trial but the magnitude of the response varies at each voxel through the estimation of β_i, $i = 1, ..., V$. We compare these to Gaussian mixture models where active voxels are modelled by,

$$\begin{aligned} &\text{Model 3: Active voxels,} \quad Y_i \sim N_{TE}(\beta_i \widetilde{\mu} + X b_i, \Sigma_E \otimes I_T); \\ &\text{Model 4: Active voxels,} \quad Y_i \sim N_{TE}(\beta_i \widetilde{\mu} + X b_i, I_E \otimes \Sigma_T); \\ &\text{Model 5: Active voxels,} \quad Y_i \sim N_{TE}(\beta_i \widetilde{\mu} + X b_i, \Sigma_E \otimes \Sigma_T); \end{aligned}$$

and, in each case, inactive voxels are modelled by $Y_i \sim N_{TE}(X b_i, \sigma^2 I_{TE})$. In these models the expected response and covariance matrix structure is different at active and inactive voxels, and the models only differ in the

complexity of the covariance structure for active voxels.

The deviance, $Dev = -2 \log L$, and Akaike information criterion, $AIC = Dev + 2p$, for each of the five models are shown in Table 5.1. It can be seen that making the model more complex adds comparatively few parameters. Therefore, the more complex models are favoured by a low AIC. The reduction in AIC from model 1 to model 2 highlights the benefit of allowing the data to estimate the haemodynamic response. Similarly, the significant reduction in deviance from model 2 to models 3-5 shows the need for separating the active and inactive voxels with a mixture model. A mixture model allows the variability of active responses to be estimated using only voxels classed as "active". The reduction in AIC between model 5 and model 4 supports our hypothesis that sequential events cannot be treated as independent responses to a stimulus and gives evidence that responses are correlated.

Model	p	$\log L$	Dev	AIC
1	70434	-12481601	24963201	25104069
2	70447	-12348551	24697102	24837996
3	80566	-3898088	7796176	7957308
4	80616	-3894852	7789705	7950937
5	80671	-3890070	7780140	7941482

Table 5.1: The deviance and AIC for five models.

To evaluate whether the difference in log-likelihood between models 3-5 was due to the model parameterisation or an artifact of the EM algorithm's ability to find the global maximum from a given starting point, the EM algorithm was run a further 10 times for each model. On each occasion the starting point was adjusted by choosing 10% of the voxels considered at random and altering their initial active/inactive classification. Further, random noise was added to the initial estimate of μ by the addition of a $N_T(0, I)$ random variable. Although the algorithm generally took longer to converge from these random starting points, in each case the evaluated

log-likelihood was identical (to 6 significant figures) to the values quoted in Table 5.1. It, therefore, seems reasonable to assume that the differences in the likelihood between the models is due to the model parameterisation and not the algorithm. Based on Table 5.1, it seems valid to model the data using model 5.

5.6 Hypothesis testing

To examine single-trial variability, we first need to extract from the data the voxels activated by the stimuli. Figure 5.4 shows that voxels with large β_i nearly always have $\widetilde{p}_i = 1$, which seems to justify using the Gaussian mixture model. The model suggests that active voxels will have a large positive β_i parameter and this parameter can be tested with a student-t test on the null hypothesis $\beta_i = 0$ versus the alternative hypothesis that $\beta_i > 0$ for $i = 1, ..., V$. Under our model, $Y_i \sim N_{TE}(\widetilde{\mu}\beta_i + Xb_i, \Sigma_1)$. A transformation gives $Y_i^* \sim N_{TE}(\mu^*\beta_i + X^*b_i, I_{TE})$ where $Y_i^* = \Sigma_1^{-1/2}Y_i$, $\mu^* = \Sigma_1^{-1/2}\widetilde{\mu}$ and $X^* = \Sigma_1^{-1/2}X$. The test statistic under $TE-q-1$ degrees of freedom is,

$$T_i = \frac{\widehat{\beta}_i}{\sqrt{S_i^2(\mu^{*T}\mu^*)^{-1}}},$$

where $S_i^2 = \|Y_i^* - \mu^*\beta_i - X^*b_i\|^2/(TE - q - 1)$.

There has been much debate in the neuro-imaging community regarding how to set the statistical significance level in fMRI studies (Marchini and Ripley, 2000), due to problems in estimating the magnitude of the response and the correlation in spatially and temporally correlated data. SPM2 utilises set-level inference, using distributional approximations from the theory of Gaussian random fields (Friston *et al.*, 1996). The method assesses the probability of obtaining c or more clusters, containing v or more voxels. Our voxel-by-voxel approach is the simplest case of this, allowing clusters of just one voxel. The use of Gaussian random fields has potential

advantages over SPM2's previous methods (Friston *et al.*, 1994, Friston *et al.*, 1995d, Worsley *et al.*, 1995), which corrected the significance level for temporal correlations only. Our method of pre-whitening the data using the estimated covariance matrix is similar to that of Worsley *et al.* (2002), although they restrict the covariance matrix to be of auto-regressive form. We correct for multiple comparisons by adjusting the initial p-value threshold of 0.001 to control the false discovery rate (Benjamini and Hochberg, 1995). We use Benjamini and Hochberg's (2000) adaptive method, and defer analysis of the spatial correlations until later.

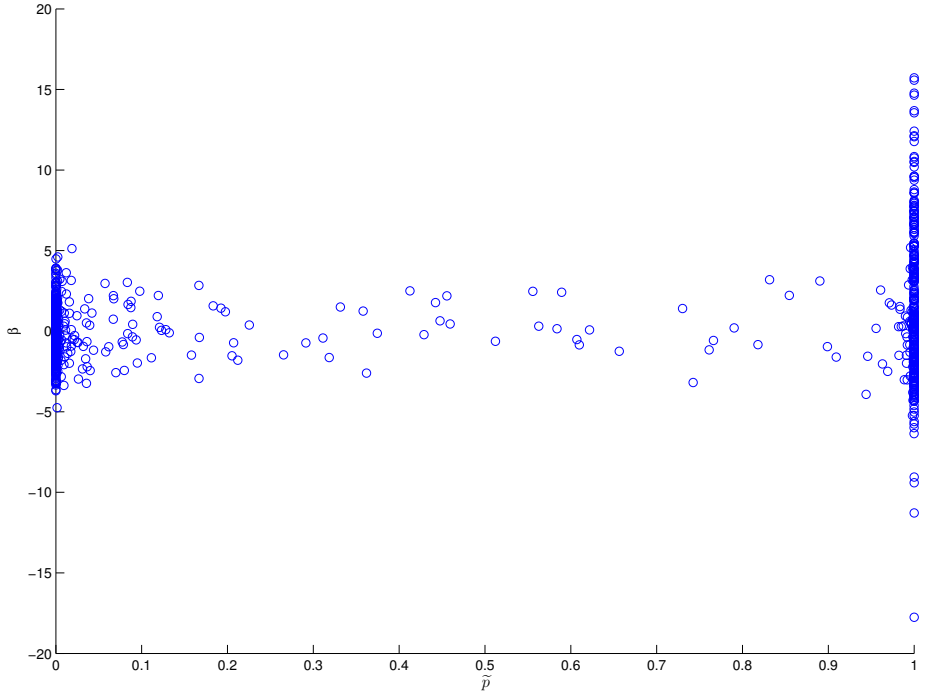

Figure 5.4: A scatter plot of $\widetilde{p}_i$ versus β_i.

The goal of single-trial variability analysis is to identify trends in the way active voxels respond in space and time to stimuli. Principal components (PC) analysis of Σ_T will highlight key differences in the response of voxels at

different events. The PC score at each voxel/event is, $s_{ijk} = \gamma_k^T(Y_{ij} - \beta_i\mu - X_j b_i)$, where γ_k is the kth PC loading and i denotes voxel and j denotes event. These are used as the response, S_k, $k = 1, ...T$, in the linear models,

$$S_k = Z\xi_k + U_k, \tag{5.9}$$

where ξ_k is a parameter vector, U_k is an error matrix, and Z is a design matrix with covariates,

$$\left[1,\ x,\ y,\ z,\ e,\ s,\ (x*y),\ (x*z),\ (y*z),\ (e*s),\ x^2,\ y^2,\ e^2\right],$$

where x, y and z are the voxel co-ordinates, e is the event number and s is a binary variable indicating whether the volunteer pressed the button once ($s = 0$) or five times ($s = 1$). Exploratory data analysis showed the assumption of isotropic variability in the model seemed false, so we transform the response and the covariates using the maximum likelihood estimate of Σ_E from the PC scores and weight the voxels by their values of $\widetilde{p}_i$. The model parameters were tested for significance using a two-sided student-t test at the 0.01 significance level. This was repeated for each PC.

5.7 Results

The EM algorithm described above was run until the change in parameter estimates between successive iterations fell below a specified tolerance of 0.001. Figure 5.5 shows the mean response, $\beta\mu$, and the effect of the first five principal components of Σ_T on the mean, $\beta\mu \pm \lambda_i^{1/2}\gamma_i$, where $\beta = 10$ and λ_i, γ_i are the ith eigenvalue and eigenvector of Σ_T. Likewise, Figure 5.6 shows the first five principal component loadings of Σ_E. The loadings do not display much structure, however maximum likelihood estimation of Σ_E shows a significant increase in the likelihood compared to $\Sigma_E = I_E$.

The voxels where β_i is significantly greater than zero are highlighted on

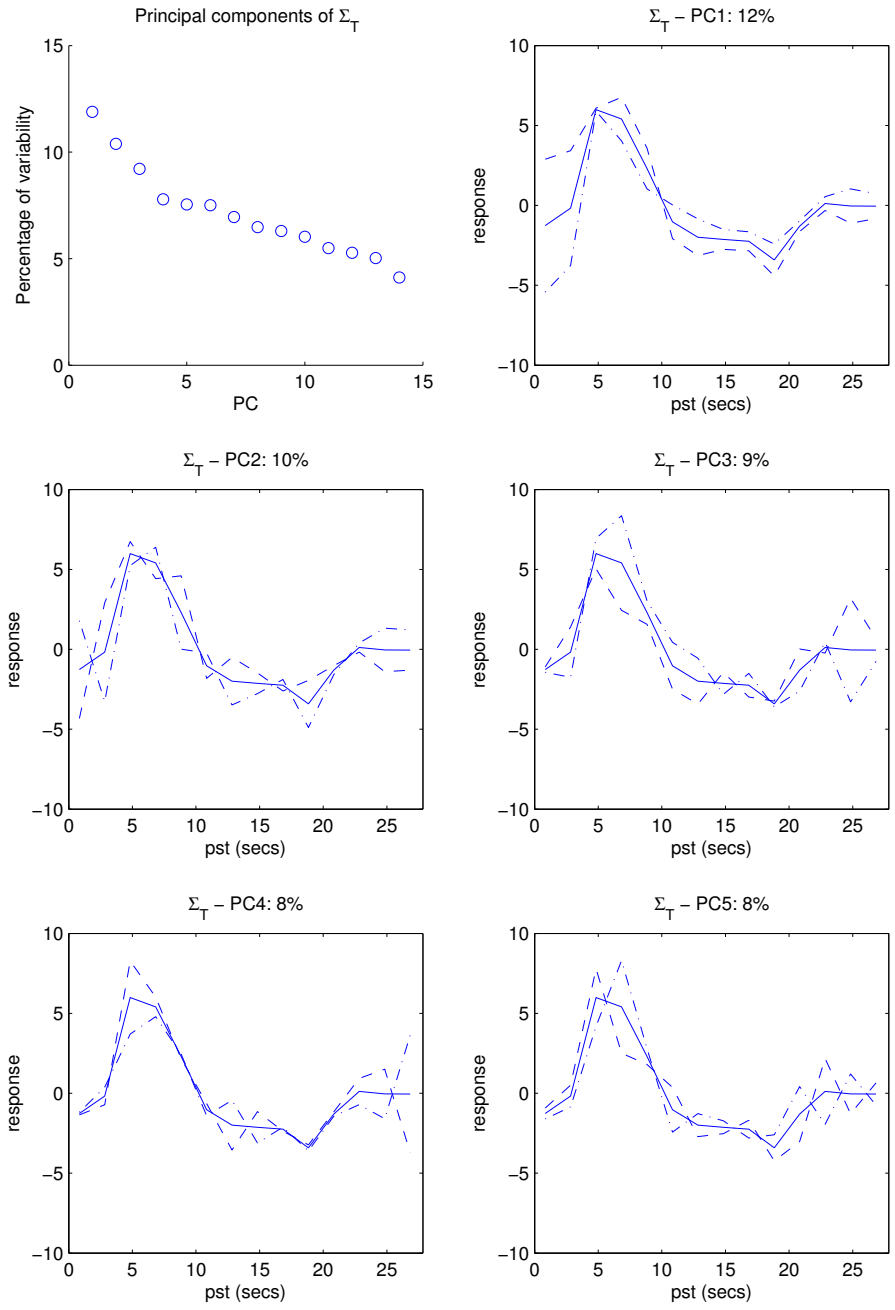

Figure 5.5: The first plot shows the percentage of variability in each principal component of Σ_T. The remainder show the mean response plus(dash/dot)/minus(dash) the first five principal components of Σ_T.

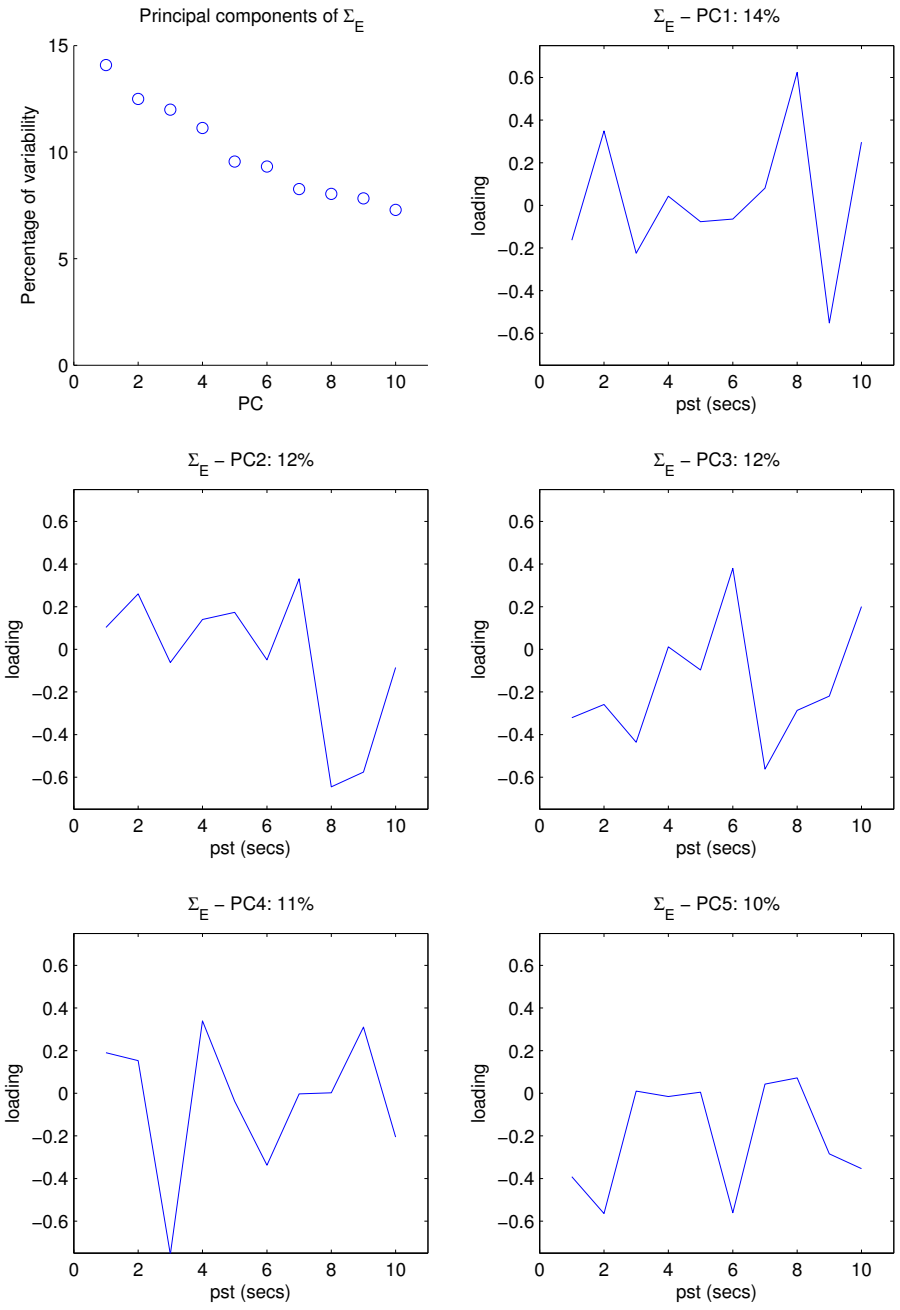

Figure 5.6: The first plot shows the percentage of variability in each principal component of Σ_E. The remainder show the first five principal component loadings of Σ_E.

the activation map in Figure 5.7. Cluster analysis of the voxel co-ordinates revealed two large group of voxels concentrated on slices eight to ten around $(x = 30, y = 20)$ and $(x = 60, y = 50)$. Interestingly, the former is similar to the area of activation detected using the existing methods of analysis presented in Section 5.8. Note that head movement during scanning has caused part of the image in slice 12 to not be consistently recorded throughout the experiment and consequently this portion of the image is masked during analysis.

Figure 5.5 shows that the first three principal components of Σ_T explain much more variability than the others. Consequently, linear models with spatial and temporal covariates, were fitted to each of the first three PC scores in turn, see Equation (5.9). Figure 5.8 shows the PC1 scores changing through time. It also shows the fitted function from the linear model, evaluated through time and space. Voxels with a low PC1 score show a much earlier rise in the response and a slightly later decay. PC2 accounts for the width of the response, with low scores corresponding to a longer period of activation. Voxels with a high PC3 score have a much later peak and a stronger response. Tables 5.2 to 5.4 show that these principal components change significantly with time and type of event, but less so spatially.

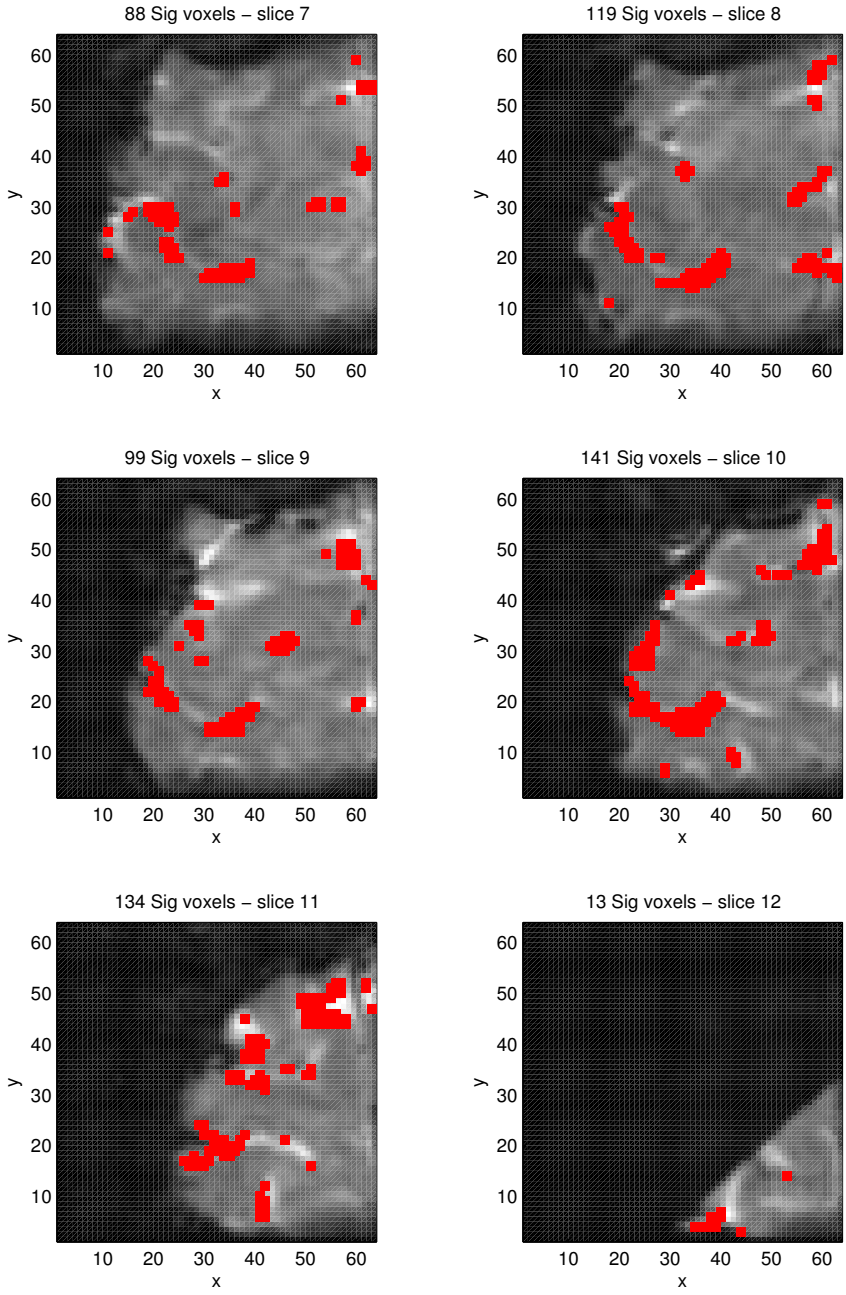

Figure 5.7: The locations of active voxels in the motor cortex.

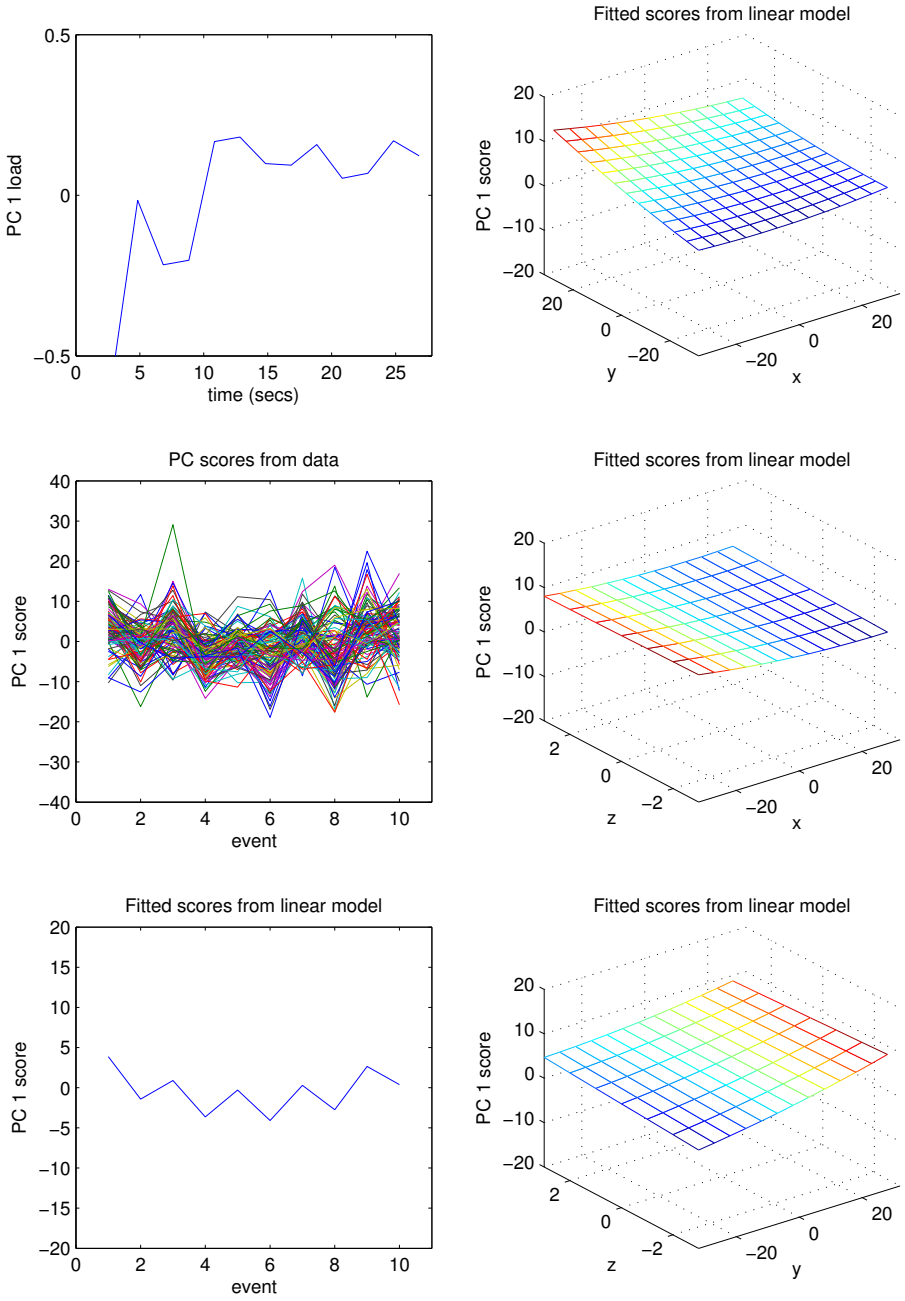

Figure 5.8: Raw and fitted PC 1 scores from the linear model in Equation (5.9). The response in odd numbered events was a single button press and even numbered events required multiple button presses.

Covariate	Parameter	S.E.	T-value	p-value
1	6.0160	0.5867	10.2543	0.0000
x	-0.0416	0.0290	-1.4317	0.1522
y	0.1065	0.0332	3.2035	0.0014
z	0.0420	0.0739	0.5676	0.5703
e	-2.3745	0.1070	-22.1840	0.0000
s	-3.4478	0.3587	-9.6131	0.0000
$x * y$	-0.0019	0.0007	-2.5603	0.0105
$x * z$	0.0047	0.0048	0.9776	0.3283
$y * z$	-0.0119	0.0046	-2.5987	0.0094
$e * s$	-0.0664	0.0662	-1.0030	0.3158
x^2	0.0010	0.0005	1.9615	0.0498
y^2	0.0008	0.0006	1.2421	0.2142
e^2	0.2222	0.0107	20.7177	0.0000

Table 5.2: t-test of parameters in the linear model with response PC score 1.

Covariate	Parameter	S.E.	T-value	p-value
1	-3.1716	0.6115	-5.1863	0.0000
x	-0.1083	0.0303	-3.5747	0.0004
y	0.0626	0.0347	1.8049	0.0711
z	0.1840	0.0772	2.3841	0.0171
e	0.9576	0.1165	8.2196	0.0000
s	-2.0552	0.2828	-7.2680	0.0000
$x * y$	-0.0036	0.0008	-4.5581	0.0000
$x * z$	0.0106	0.0050	2.1176	0.0342
$y * z$	0.0028	0.0048	0.5908	0.5547
$e * s$	0.5633	0.0480	11.7301	0.0000
x^2	0.0030	0.0005	5.6534	0.0000
y^2	0.0029	0.0007	4.3566	0.0000
e^2	-0.1282	0.0120	-10.6905	0.0000

Table 5.3: t-test of parameters in the linear model with response PC score 2.

Covariate	Parameter	S.E.	T-value	p-value
1	-2.7932	0.5537	-5.0449	0.0000
x	0.0456	0.0276	1.6500	0.0989
y	0.0203	0.0316	0.6425	0.5206
z	0.2644	0.0703	3.7583	0.0002
e	-0.3918	0.0965	-4.0591	0.0000
s	4.4343	0.2880	15.3948	0.0000
$x * y$	-0.0028	0.0007	-3.9251	0.0001
$x * z$	-0.0144	0.0046	-3.1557	0.0016
$y * z$	0.0040	0.0044	0.9125	0.3615
$e * s$	0.0550	0.0473	1.1618	0.2453
x^2	0.0007	0.0005	1.5143	0.1300
y^2	0.0026	0.0006	4.3140	0.0000
e^2	0.0206	0.0091	2.2558	0.0241

Table 5.4: t-test of parameters in the linear model with response PC score 3.

Changes in haemodynamic response for a particular cluster can also be seen through plotting the fitted values,

$$\widehat{Y}_{ij} = \bar{\beta}\mu + \sum_{k=1}^{T} \widehat{s}_{ijk}\gamma_k, \tag{5.10}$$

where $\widehat{s}_{ijk}$ are the fitted PC scores from the linear models in Equation (5.9), and $\bar{\beta}$ is the mean value of β_i for the voxels in the cluster. Figures 5.9 show that the response is much stronger and peaks later with a larger undershoot in trials where the volunteer presses the button multiple times. In the one-press trials the response is largest in the middle of the experiment. In both cases the strength of the response tails off towards the end of the experiment, and the undershoot is largest towards the middle of the experiment.

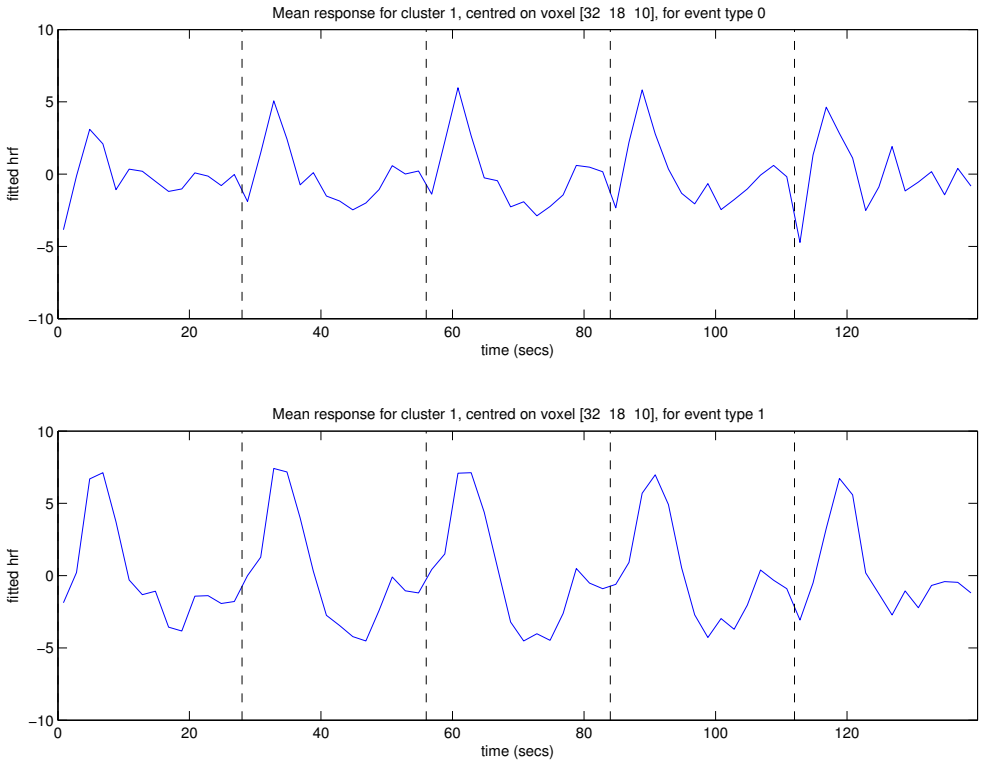

Figure 5.9: The fitted haemodynamic response for a particular cluster from Equation (5.10) for one-press trials (top) and five-press trials (bottom).

5.8 Current methods of analysis

We analyse the data using two current statistical software packages for comparison. The first is MELODIC, developed by Beckmann and Smith (2004). This uses probabilistic independent components analysis (ICA) to split the data into signal and noise. The raw data is first preprocessed by removing non-brain voxels, standardising the mean and variance of voxel-values at each location, and then projecting the data into a 36-dimensional subspace by probabilistic principal components analysis. ICA then decomposes the time-courses by optimising for non-Gaussian spatial source distributions (Hyvarinen *et al.*, 2001). We distinguish signal from noise by searching the 36 time-courses for those with a frequency near the reciprocal of the inter-stimuli interval, i.e. $1/28.25 = 0.0354$ Hz. The time-course, frequency distribution and spatial IC map for component 13 are shown in Figure 5.10. The highlighted regions are in the motor and somatosensory cortex, responsible for movement and sensing touch, respectively, where neurologists would expect activation.

The second software package, SPM2, fits a linear model at each voxel, with a design matrix composed of covariate information and a basis function of a typical haemodynamic response (Friston *et al.*, 1995e). The data and design matrix are preprocessed as described in Section 5.2 but with conventional slice-timing. A student-t test of the parameter corresponding to the haemodynamic basis function shows which voxels have significant non-zero activity. The active voxels are shown in yellow in Figure 5.11. Note the bottom-right image from the SPM2 output corresponds to slice 10 of the MELODIC output in Figure 5.10 but with the x-axis flipped in orientation. MELODIC's activation map is much smoother than SPM2's but both show similar regions of activation, particularly between where the cross-hairs are located and the surface of the brain.

Alternative approaches to analysing fMRI data have also been proposed. Friston *et al.* (1995b) suggests a multivariate approach, treating each scan

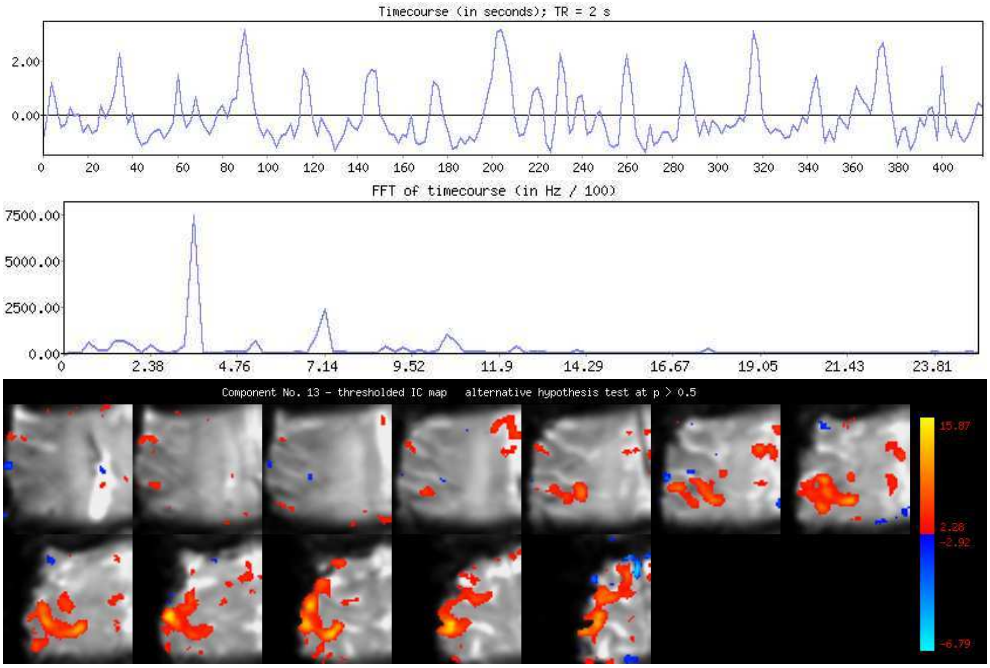

Figure 5.10: The time-course, with its frequency and spatial map, of component 13. Red areas indicate locations where this signal is present. The spatial map is plotted on slices 1-7 (top row) and slices 8-12 (bottom row).

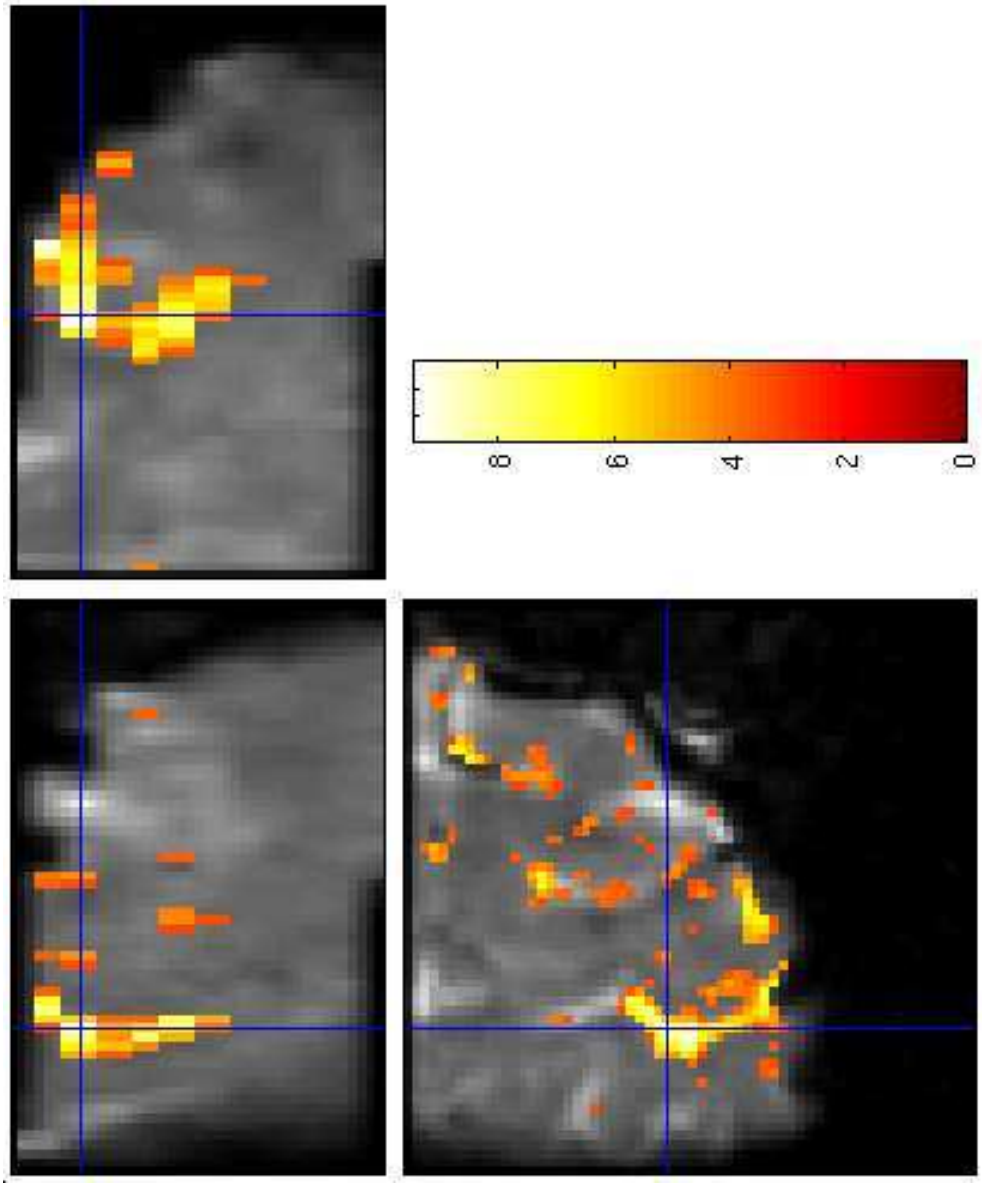

Figure 5.11: Highlighted voxels have significant non-zero activity under SPM2's model with a p-value below 0.001.

as a single observation, rather than the mass univariate approach of SPM2. While this method has advantages in modelling the temporal aspects of the haemodynamic response, its inability to make inferences about regional effects has meant it has not found favour in the neuro-imaging community. Friston *et al.* (1995c) uses two basis functions in the linear model to discriminate between early and late responses. This is an improvement over the usual SPM2 model when the form of the haemodynamic response is unknown. However, even more basis functions would be required to fully model the response, generating a larger number of possible contrasts and making physiological interpretation harder. Josephs *et al.* (1997) take this approach, employing multiple basis functions taken from terms of a Fourier series. Examination of the resulting F-statistic at each voxel tests whether the variance explained by the effects of interest is zero or not.

5.9 Discussion

In this chapter we have developed a statistical model for analysing single trial variability in fMRI data. The chosen model produced a significantly higher likelihood than other simpler models, see Table 5.1, including those implemented by software packages currently available. The model indicates areas of the brain active in performing the task of pressing a button, which concur with other methods of analysis, but the interpretation of results seems clearer here. In addition, the model gives us a maximum likelihood estimate of the variability in button presses within and between events. The three main sources of variation, displayed through principal components analysis, were found to be the timing of the initial rise in response, the length of the response and the strength of the peak. It remains unclear, however, if the variation is caused by changes in neuronal processes or by natural variation in the reaction time of the volunteer and the strength of the button press. The amount of noise in the data could be reduced with the inclusion of physiological data as covariates but this data was unavailable

for the experimental results presented here.

Importantly, through examining the PC scores, we are able to show how the response changes with the task and through time and demonstrate that these changes are statistically significant. The shape of the estimated mean response is similar to that of prior studies but displays a greater undershoot than previously reported, particularly when the peak is stronger and wider in the cases where the volunteer presses the button multiple times.

The methods and results given in this chapter obviously pertain to one experiment carried out with one volunteer. It would be interesting in further work to apply the model and methodology to other volunteers, or repeat it with the same volunteer, to examine if the sources of trial variability found in this study are common to all subjects or experiment repetitions. If this proves to be the case, it will greatly enhance the ability for neurologists to reach conclusions on voxel activation where a traditional repeated trial experiment paradigm is either impossible to conduct or impractical with the resources available.

Further work is also required to produce accessible software that will allow potential users to implement quickly and accurately the methodology presented in this chapter. Software packages such as SPM2 or MELODIC appeal to researchers because of their flexibility and robustness in working with different data sets, and their use has become widespread in the neuroimaging community. We wish to develop a similar software toolbox for use in single-trial variability studies that will deliver clear results to the user and facilitate easier interpretation.

Chapter 6

Conclusions and further work

6.1 Summary and Discussion

In this chapter we give the main findings and suggest possible extensions to the research presented in this thesis. Our objectives have been to reduce the error in shape registration and provide suitable models for the statistical shape analysis of symmetry and function in brain imaging data. This has required the generalisation and application of some existing statistical techniques, as well as suggesting novel methodology. In particular we have drawn heavily upon the concepts of shape analysis, Procrustes analysis and principal components analysis and carefully considered their application to high-dimensional data sets of the human brain, which possesses an intricate shape. The non-isotropic nature of the variability has frequently required the application of models with a large number of parameters, and techniques such as the EM algorithm and Markov chain Monte Carlo (MCMC) simulation have been used to estimate parameter values.

In Chapter 2 we discussed Procrustes shape registration using a known weighting matrix. Particular consideration was given to the special case where only a subset of landmarks were registered using isotropic Procrustes analysis. We derived expressions for the distribution of the Euclidean dis-

tance between the object and both a true template and an alternative template, which quantify theoretically the efficiency of the subset-matching estimator and examine its variance. In the more general weighted Procrustes case, we provide estimates of the scaling, rotation and translation that minimise the Mahalanobis distance between two or more configurations in the presence of both full and partial registrations. These estimators are shown to be identical to isotropic Procrustes estimators if the weighting matrix is proportional to the identity matrix. The variability in typical data sets is often far greater than any bias in the estimators.

In Chapter 3 we extended weighted Procrustes analysis to include estimation of an unknown shape covariance matrix which could be used to weight the registration. Both conditional maximum likelihood and MCMC algorithms were suggested for parameter estimation and a simulation study showed their ability to reduce the error in covariance matrix estimation compared to isotropic Procrustes. In the maximum likelihood case, consideration was given to the parameterisation of the covariance matrix and constraints were applied to avoid singularities by projecting the covariance matrix into the subspace orthogonal to the constraint vectors. In the MCMC case, the method was extended to cope with missing data through simulating the co-ordinates of missing landmarks as a step in the algorithm. Both the conditional maximum likelihood and MCMC algorithms are dependent on prior information but the MCMC method is more robust to weak or inaccurate prior information.

In Chapter 4 we applied statistical shape analysis to a data set of magnetic resonance imaging (MRI) brain scans consisting of controls and schizophrenia patients. In order to preserve shape information, maximum likelihood and Bayesian algorithms were developed to provide rigid-body registrations of each scan to a co-ordinate system based on Talairach space. The posterior distribution was found to be tightly concentrated around the maximum *a posteriori* estimate, so a simple grid search of parameter values to maximise the likelihood was chosen for speed of computation. Follow-

ing registration, a labelling was suggested that enabled the symmetry of the two groups to be investigated. Principal components analysis revealed more torque in the control group. A novel analysis of the inter-hemispherical join's location revealed the join to be further to the left in the patient group in the anterior region of the brain. There was some evidence that this provided an explanation for the observed asymmetry. The symmetry analysis was compared to voxel-based morphometry (VBM) analysis and the two methods were found to be complementary in locating large-scale and small-scale shape differences respectively.

In Chapter 5 we analysed brain activity using functional MRI scans. This required modelling voxel values with spatially and temporally dependent errors. Current model-driven methods using linear models with *a priori* response functions were generalised so that the response's form could be estimated. In addition, voxel values were modelled as a mixture of active and inactive components, thus covariance matrices could be estimated using active voxels alone. Modifications to image pre-processing and improved estimation of the error structure enabled single-trial variability to be analysed through principal component scores. The timing, length and strength of the response were shown to change through the sequence of consecutive trials, as well as being dependent on the task carried out in each trial. Reassuringly, the voxels that displayed significant activation were also shown to be active using two current software programs.

The research presented in this thesis has made significant progress in achieving our objectives. However, there remains plenty of potential for improving the analysis of shapes and brain imaging data, and we discuss some of these issues in Section 6.2.

6.2 Future work

The research presented in this thesis raises several questions which remain potential areas of future study. Firstly, the work relating to covariance-

weighted Procrustes analysis produces estimators of the mean shape, shape covariance and transformation parameters. Theoretical properties such as consistency will need to be investigated. This is a non-trivial problem because the estimates are obtained using a recursive algorithm.

Walker (2000) investigated the ability of morphometric methods to estimate covariance matrices with respect to factors such as the number of landmarks, the magnitude of variation, the proportion of landmarks with excessive variance compared to others and the magnitude of correlation. A similar analysis could be conducted using isotropic and covariance-weighted Procrustes analysis to gain a better understanding of the conditions that cause the latter to produce improved estimates of shape variability.

In Chapter 3 we extended the MCMC simulation to include the estimation of missing landmarks. The same objective could be achieved using maximum likelihood methods through an expectation-maximisation (EM) algorithm. The data could be augmented with latent data representing the location of the missing landmarks. It would then be possible to use the parameter estimates given in Chapter 3 to maximise the complete likelihood using the expected values of the landmark locations.

Further, covariance weighted Procrustes analysis could also be applied to data sets with unlabelled landmarks. Dryden *et al.* (2006) consider using isotropic Procrustes in conjunction with a match matrix that assigns labels to landmarks to register molecules. The match matrix is updated using an MCMC algorithm to determine possible labellings. Instead of using isotropic Procrustes, a constrained maximum likelihood estimate of the shape variability, as given in Chapter 3, could be used to weight the Procrustes registration for a particular labelling. Assuming the labelling is known, Figure 6.1 shows the isotropic Procrustes registration of several molecules, each consisting of 17 atoms arranged in four rings. A larger data set also includes molecules that have less than 17 atoms. Figure 6.2 shows the registration of these molecules using our hybrid MCMC algorithm, where some atom locations are simulated. The complete data set

consists of 31 molecules, each possessing up to 61 different atoms. It remains to be seen whether a good registration is possible with the complete data set of 61 atom locations and an unknown labelling.

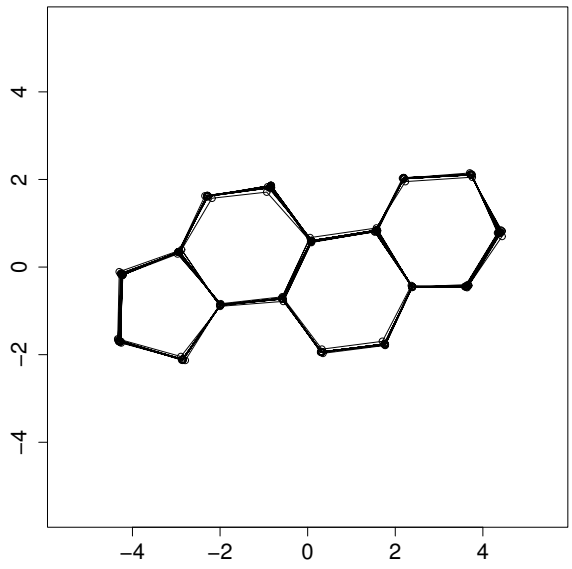

Figure 6.1: The location of 17 atoms registered using isotropic Procrustes.

The brain registration algorithm presented in Chapter 4 is reliant on the Brain Extraction Tool (BET) software correctly masking the non-brain voxels which is dependent on the choice of a tuning parameter. An alternative method is to use the Statistical Parametric Mapping 2 (SPM2) software to estimate a non-linear transformation of each image to a standard template, and then invert the transformation to map the mask corresponding to the template back to the image. Theoretically this possibility is potentially more robust and requires further investigation. Indeed, in further work we could also consider extending the Bayesian registration method to incorporate the estimation of many of the parameters currently used in image

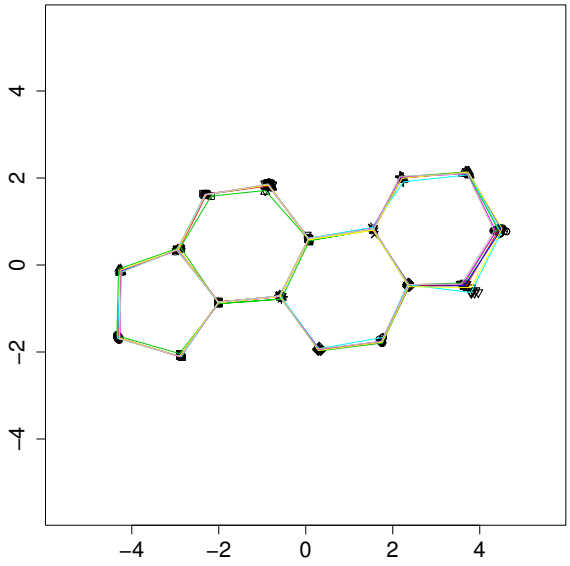

Figure 6.2: The location of 17 atoms registered using the hybrid MCMC algorithm with missing landmarks. The lines show the molecules with missing atoms.

preprocessing and modelling, such as the mask, weights and distances that define regions of interest. This would reduce the number of assumptions necessary, increasing the flexibility of the model and reduce the influence of parameters chosen *a priori*.

With respect to brain image registration we considered a single data set with all images obtained from one scanner. This is conventional practice to negate any confounding factors caused by the scanner. However, the ability to register scans of other types or modalities through the identification of the midplane and the anterior and posterior commissures is currently untested. Ashburner and Friston (1997) approach this problem through defining a template in each modality and estimating the transformation that maps each template to Talairach space. Our method provides a rigid-body registration directly to Talairach space without the need for a template, but training data would be required for each modality.

In Chapter 4 we considered a data set consisting of controls and patients with schizophrenia. Obviously, the registration and symmetry analysis could be applied to patients of other brain disorders, such as Alzheimer's disease, semantic dementia or lesions, which potentially influence the shape of the cortical surface. Gitelman *et al.* (2001) applies voxel based morphometry (VBM) to individuals who have recovered from herpes simplex encephalitis (HSE) and consequently have severely distorted brain shape. VBM encountered problems with its normalisation step due to the abnormalities and was forced to use the skull as well as the brain to produce a sensible registration. Our method is heavily dependent on the symmetry of the midplane and it is conceivable that it would encounter similar problems if a disease caused an asymmetric distortion to the shape of the ventricles. Any future application might require a masking of an affected brain region during registration.

The registration also included the identification of the anterior and posterior commissure using training data. The same method could be applied to other brain landmarks, to produce three dimensional co-ordinates of brain

features. A more conventional shape analysis could then be applied, possibly using covariance-weighted Procrustes, to discriminate between two or more patient groups. Unfortunately, the brain does not contain many landmarks that are clearly defined.

The study of single-trial variability is still in its infancy due to the limited availability of new technology required to improve the signal-to-noise ratio in functional MRI data. Our study has been handicapped by the limited data and further work is required to evaluate the model using more subjects to see if similar results are obtained. The errors could also be reduced if physiological data, such as pulse and respiration, could be included as covariates as these influence the natural oxygen levels in the blood.

A future experiment might focus on the somatosensory cortex that responds when a probe applies pressure to an area of skin, such as the tip of the index finger. This paradigm would help to remove variation caused by the volunteer's reaction time and the force used to press the button by minimising subject variability. The lag between the stimulus and the response would be more consistent and the pressure of the probe could be a carefully controlled variable. It would also be easier to compare responses between subjects.

Although the model proposed in Chapter 5 allows for variability between trials and can display spatially and temporally correlated trends in the principal components, we currently only display a single activation map. A more useful tool might show significant PC scores, providing a spatial map of voxels at each trial with, for example, an early peak in the response. A late response might be indicative of a draining vein rather than true neuronal activity. Further, while the mixture model produces separate estimates of variability for active and inactive voxels, the resulting statistical tests could be strengthened if the estimated variability is adjusted for each voxel, perhaps via an additional scalar parameter. Future research in this area could consider the fitting of a more complicated model.

The areas of activation identified with our model are slightly larger than

those shown by SPM2. This is probably due to SPM2 using Gaussian random fields to account for spatial correlation in the data. Incorporating random fields into the analysis would base the tests on the probability of obtaining c or more clusters with v or more voxels above a threshold. It could be used to reduce the probability of deeming voxels active in isolation and make our analysis more consistent with SPM2.

Functional MRI analysis is also limited by the poor temporal resolution of, typically, 2 seconds between scans. Electroencephalogram (EEG) data has much better temporal resolution, but poorer spatial resolution. It also does not require activation to be measured via the blood oxygen level which reacts much slower than the electrical signals. The concept of combining EEG and functional MRI scanning presents new challenges to both technology and statistical analysis but might provide the key to a better understanding of single-trial variability, given that neuronal firing occurs on a very small temporal and spatial scale.

The experiment paradigm considered in Chapter 5 consisted of a very simple task measured in the motor cortex. Individual brain regions, such as the motor cortex, are fairly well understood in isolation. The more complicated and realistic challenge is to understand the connectivity between brain regions in order to gain insight into the brain's decision-making processes. For example, the current experiment paradigm probably involves the visual cortex receiving the stimulus and a thought process in the frontal lobe before the motor cortex gives the response. We might expect the contribution of the frontal lobe to reduce during the experiment as the brain "remembers" the previous response. However, testing this hypothesis will require a more complicated model beyond the scope of the current study.

All of the research presented in this thesis has been conducted using functions written for the R (R Development Core Team, 2005) and Matlab (MathWorks, Natick, MA, USA) software programs for personal use. It would be a useful development for future studies to make the code available to applied researchers through a library or toolbox of functions, and this

will be one of the first tasks to undertake in further work. The benefit of producing free and intuitive software can be seen by the wealth of references in the literature to SPM2 and papers written by its authors. The distribution of our source code, combined with a graphical user interface, would aid the research of others and make it easier for researchers to adopt the techniques proposed in this thesis.

Bibliography

Arnold, S.F. (1981). *The Theory of Linear Models and Multivariate Analysis.* Wiley, New York

Ashburner, J., Andersson, J.L.R. and Friston, K.J. (1999). High-dimensional image registration using symmetric priors. *NeuroImage,* **9,** 618-628.

Ashburner, J., Andersson, J.L.R. and Friston, K.J. (2000). Image registration using a symmetric prior - in three dimensions. *Human Brain Mapping,* **9,** 212-225.

Ashburner, J. and Friston, K. (1997). Multimodal image coregistration and partitioning - a unified framework *NeuroImage,* **6,** 209-217.

Ashburner, J. and Friston, K. (1999). Nonlinear spatial normalization using basis functions. *Human Brain Mapping,* **7,** 254-266.

Ashburner, J. and Friston, K. (2000). Voxel-based morphometry - the methods. *NeuroImage,* **11,** 805-821.

Ashburner, J. and Friston, K. (2001). Why voxel-based morphometry should be used. *NeuroImage,* **14,** 1238-1243.

Ashburner, J., Hutton, C., Frackowiak, R., Johnsrude, I., Price, C. and Friston, K. (1998). Identifying global anatomical differences:

Deformation-based morphometry. *Human Brain Mapping,* **6,** 348-357.

Ashburner, J., Neelin, P., Collins, D.K., Evans, A. and Friston, K. (1997). Incorporating prior knowledge into image registration. *NeuroImage,* **6,** 344-352.

Barrick, T.R., Mackay, C.E., Prima, S., Maes, F., Vandermeulen, D., Crow, T.J. and Roberts, N. (2005). Automatic analysis of cerebral asymmetry: An exploratory study of the relationship between brain torque and planum temporale asymmetry. *NeuroImage,* **24,** 678-691.

Beckmann, C.F. and Smith, S.M. (2004). Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Transactions on Medical Imaging,* **23,** 137-152.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B,* **57,** 289-300.

Benjamini, Y. and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics,* **25,** 60-83.

Bilder, R.M., Wu, H., Bogerts, B., Degreef, G., Ashtari, M., Alvir, J.M.J., Snyder, P.J. and Lieberman, J.A. (1994). Absence of regional hemispheric volume asymmetries in first-episode schizophrenia. *American Journal of Psychiatry,* **151,** 1437-1447.

Bookstein, F.L. (1986). Size and shape spaces for landmark data in two dimensions. *Statistical Science,* **1,** 181-242.

Bookstein, F.L. (2001). "Voxel based morphometry" should not be used with imperfectly registered images. *NeuroImage,* **14,** 1454-1462.

Bookstein, F., Schafer, K., Prossinger, H., Seidler, H., Fieder, M., Stringer, C., Weber, G.W., Arsuaga, J., Slice, D.E., Rohlf, F.J., Recheis, W., Mariam, A.J. and Marcus, L.F. (1999). Comparing frontal cranial profiles in archaic and modern homo by morphometric analysis. *The Anatomical Record, The New Anatomist,* **257,** 217-224.

Brett, M., Leff, A.P., Rorden, C. and Ashburner, J. (2001). Spatial normalization of brain images with focal lesions using cost function masking. *NeuroImage,* **14,** 486-500.

Brignell, C.J., Browne, W.J. and Dryden, I.L. (2005). Covariance weighted Procrustes analysis. *In: Barber, S., Baxter, P.D., Mardia, K.V. and Walls, R.E. (eds.) Quantitative biology, shape analysis, and wavelets,* 107-110.

Brignell, C.J., Dryden, I.L., Gattone, S.A., Park, S.B.G., Leask, S.J., Browne, W.J. and Flynn, S. (2006). Surface shape analysis, with an application to brain cortical surface analysis in schizophrenia, to be submitted for publication.

Browne, W.J. and Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics,* **15,** 391-420.

Buxton, R.B., Wong, E.C. and Frank, L.R. (1998). Dynamics of blood flow and oxygenation changes during brain activation: The balloon model. *Magnetic Resonance in Medicine,* **39,** 855-864.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician,* **49,** 327-335.

Collinson, S.L., Mackay, C.E., James, A.C., Quested, D.J., Phillips, T., Roberts, R. and Crow, T.J. (2003). Brain volume, asymmetry and

intellectual impairment in relation to sex in early-onset schizophrenia. *British Journal of Psychiatry,* **183,** 114-120.

Crow, T.J. (1997). Schizophrenia as failure of hemispheric dominance for language. *Trends in Neurosciences,* **20,** 339-343.

Csernansky, J.G., Joshi, S., Wang, L., Haller, J.W., Gado, M., Miller, J.P., Grenander, U. and Miller, M.I. (1998). Hippocampal morphometry in schizophrenia by high dimensional brain mapping. *The Proceedings of the National Academy of Sciences,* **95,** 11406-11411.

Day, N.E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika,* **56,** 463-474.

Debener, S., Ullsperger, M., Siegel, M., Fiehler, K., Yves von Cramon, D. and Engel, A.K. (2005). Trial-by-trial coupling of concurrent electroencephalogram and functional magnetic resonance imaging identifies the dynamics of performance monitoring. *The Journal of Neuroscience,* **25,** 11730-11737.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B,* **39,** 1-38.

Dey, D.K. and Srinivasan, C. (1985). Estimation of a covariance matrix under Stein's loss. *The Annals of Statistics,* **13,** 1581-1591.

Dryden, I.L., Hirst, J.D. and Melville, J.L. (2006). Statistical analysis of unlabelled point sets: Comparing molecules in cheminformatics. *Biometrics,* to appear.

Dryden, I.L. and Mardia, K.V. (1991). General shape distributions in a plane. *Advances in Applied Probability,* **23,** 259-276.

Dryden, I.L. and Mardia, K.V. (1998). *Statistical Shape Analysis.* Wiley, Chichester.

Duann, J-R., Jung, T-P., Kuo, W-J., Yeh, T-C., Makeig, S., Hsieh, J-C. and Sejnowski, T.J. (2002). Single-trial variability in event-related BOLD signals. *NeuroImage,* **15,** 823-835.

Dutilleul, P. (1999). The MLE algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation,* **64,** 105-123.

Eysenck, H.J. and Frith, C.D. (1977). *Reminiscence, motivation and personality.* Plenum, New York.

Fischl, B., Liu, A. and Dale, A.M. (2001). Automated manifold surgery: Constructing geometrically accurate and topologically correct models of the human cerebral cortex. *IEEE Transactions on Medical Imaging,* **20,** 70-80.

Friston, K.J., Ashburner, J., Frith, C.D., Poline, J-B., Heather, J.D. and Frackowiak, R.S.J. (1995a). Spatial registration and normalization of images. *Human Brain Mapping,* **2,** 165-189.

Friston, K.J., Frith, C.D., Frackowiak, R.S.J. and Turner, R. (1995b). Characterizing dynamic brain responses with fMRI: A multivariate approach. *NeuroImage,* **2,** 166-172.

Friston, K.J., Frith, C.D., Turner, R. and Frackowiak, R.S.J. (1995c). Characterizing evoked hemodynamics with fMRI. *NeuroImage,* **2,** 157-165.

Friston, K.J., Holmes, A.P., Poline, J-B., Grasby, P.J., Williams, S.C.R., Frackowiak, R.S.J. and Turner, R. (1995d). Analysis of fMRI time-series revisited. *NeuroImage,* **2,** 45-53.

Friston, K.J., Holmes, A., Poline, J-B., Price, C.J. and Frith, C.D. (1996). Detecting activations in PET and fMRI: Levels of inference and power. *NeuroImage,* **4,** 223-235.

Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.-P., Frith, C.D. and Frackowiak, R.S.J. (1995e). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping,* **2,** 189-210.

Friston, K.J., Jezzard, P. and Turner, R. (1994). Analysis of functional MRI time series. *Human Brain Mapping,* **1,** 153-171.

Friston, K.J., Josephs, O., Ross, G. and Turner, R. (1998). Nonlinear event-related responses in fMRI. *Magnetic Resonance in Medicine,* **39,** 41-52.

Friston, K.J., Mechelli, A., Turner, R. and Price, C.J. (2000). Nonlinear responses in fMRI: The balloon model, Volterra kernels, and other hemodynamics. *NeuroImage,* **12,** 466-477.

Gelfand, A.E., Smith, A.F.M. and Lee, T.-M. (1992) Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association,* **87,** 523-532.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **6,** 721-741.

Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1996). *Markov Chain Monte Carlo in Practice.* Chapman and Hall, London.

Gitelman, D.R., Ashburner, J., Friston, K.J., Tyler, L.K. and Price, C.J. (2001). Voxel-based morphometry of herpes simplex encephalitis. *NeuroImage,* **13,** 623-631.

Glasbey, C.A., Horgan, G.W., Gibson, G.J. and Hitchcock, D. (1995). Fish shape analysis using landmarks. *Biometrical Journal,* **37,** 481-495.

Good, C.D., Johnsrude, I.S., Ashburner, J., Henson, R.N.A., Friston, K.J and Frackowiak, R.S.J. (2001a). A voxel-based morphometric study of ageing in 465 normal adult human brains. *NeuroImage,* **14,** 21-36.

Good, C.D., Johnsrude, I.S., Ashburner, J., Henson, R.N.A., Friston, K.J and Frackowiak, R.S.J. (2001b). Cerebral asymmetry and the effects of sex and handedness on brain structure: A voxel-based morphometric analysis of 465 normal adult human brains. *NeuroImage,* **14,** 685-700.

Good, C.D., Scahill, R.I., Fox, N.C., Ashburner, J., Friston, K.J., Chan, D., Crum, W.R., Rossor, M.N. and Frackowiak, R.S.J. (2002). Automatic differentiation of anatomical patterns in the human brain: Validation with studies of degenerative dementias. *NeuroImage,* **17,** 29-46.

Goodall, C. (1991). Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society, Series B,* **53,** 285-339.

Goodall, C. (1995). Procrustes methods in the statistical analysis of shape revisited. *In: Mardia K.V. and Gill C.A. (eds.) Current issues in statistical shape analysis,* 18-33.

Gower, J.C. (1975). Generalized Procrustes analysis. *Psychometrika,* **40,** 33-51.

Gower, J.C. and Dijksterhuis, G.B. (2004). *Procrustes Problems.* Oxford University Press, New York.

Green, P.J. and Mardia, K. (2006). Bayesian alignment using hierarchical models, with applications in protein bioinformatics. *Biometrika,* **93,** 235-254.

Haff, L.R. (1980). Empirical Bayes estimation of the multivariate normal covariance matrix. *The Annals of Statistics,* **8,** 586-597.

Han, Y. and Park, H. (2004). Automatic registration of brain magnetic resonance images based on Talairach reference system *Journal of magnetic resonance imaging,* **20,** 572-580.

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika,* **57,** 97-109.

Highley, J.R., McDonald, B., Walker, M.A., Esiri, M.M. and Crow, T.J. (1999). Schizophrenia and temporal lobe asymmetry. A postmortem stereological study of tissue volume. *British Journal of Psychiatry,* **175,** 127-134.

Highley, J.R., Walker, M.A., Esiri, M.M., McDonald, B., Harrison, P.J., and Crow, T.J. (2001). Schizophrenia and the frontal lobes. *British Journal of Psychiatry,* **178,** 337-343.

Hobolth, A., Kent, J.T. and Dryden, I.L. (2002). On the relation between edge and vertex modelling in shape analysis. *Scandinavian Journal of Statistics,* **29,** 355-374.

Hurn, M., Husby, O. and Rue, H. (2003) *Advances in Bayesian image analysis. In: Green, P.J., Hjort, N.L. and Richardson, S. (eds.) Highly structured stochastic systems.* Oxford University Press, Oxford.

Hurn, M., Steinsland, I. and Rue, H. (2001). Parameter estimation for a deformable template model. *Statistics and Computing,* **11,** 337-346.

Hyvarinen, A., Karhunen, J. and Oja, E. (2001). *Independent Components Analysis.* Wiley, New York.

Izard, C., Jedynak, B.M. and Stark, C.E.L. (2005). Automatic landmarking of magnetic resonance brain images. *In: Fitzpatrick, J.M.*

and Reinhardt, J.M. (eds.) Medical Imaging 2005: Image Processing, 1329-1340.

Johnson, N.L. and Kotz, S. (1972). *Distributions in Statistics: Continuous Multivariate Distributions.* Wiley, New York.

Johnson, N.L., Kotz, S. and Balakrishnan, N. (1994). *Continuous Univariate Distributions Volume I.* Wiley, New York.

Johnson, N.L., Kotz, S. and Balakrishnan, N. (1995). *Continuous Univariate Distributions Volume II.* Wiley, New York.

Josephs, O., Turner, R. and Friston, K. (1997). Event-related fMRI. *Human Brain Mapping,* **5,** 243-248.

Kendall, D.G. (1977). The diffusion of shape. *Advances in Applied Probability,* **9,** 428-430.

Kendall, D.G. (1984). Shape manifolds, Procrustean metrics, and complex projective spaces. *The Bulletin of the London Mathematical Society,* **16,** 81-121.

Kendall, D.G., Barden, D., Carne, T.K. and Le, H. (1999). *Shape and Shape Theory.* Wiley, Chichester.

Kent, J.T. (1994). The complex Bingham distribution and shape analysis. *Journal of the Royal Statistical Society, Series B,* **56,** 285-299.

Kent, J.T., Dryden, I.L. and Anderson, C.R. (2000). Using circulant symmetry to model featureless objects. *Biometrika,* **87,** 527-544.

Kent, J.T. and Mardia, K.V. (1997). Consistency of Procrustes estimators. *Journal of the Royal Statistical Society, Series B,* **59,** 281-290.

Kent, J.T. and Mardia, K.V. (2001). Shape, Procrustes tangent projections and bilateral symmetry. *Biometrika,* **88,** 469-485.

Kertesz, A., Pold, M., Black, S.E. and Howell, J. (1990). Sex, handedness, and the morphometry of cerebral asymmetries on magnetic resonance imaging. *Brain Research,* **530,** 40-48.

Klassen, E., Srivastava, A., Mio, W. and Joshi, S.H. (2004). Analysis of planar shapes using geodesic paths on shape spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **26,** 372-383.

Klingenberg, C.P., Barluenga, M. and Meyer, A. (2002). Shape analysis of symmetric structures: Quantifying variation among individuals and asymmetry. *Evolution,* **56,** 1909-1920.

Koschat, M.A. and Swayne, D.F. (1991). A weighted Procrustes criterion. *Psychometrika,* **56,** 229-239.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis,* **88,** 365-411.

Lele, S. (1993). Euclidean distance matrix analysis (EDMA): Estimation of mean form and mean form difference. *Mathematical Geology,* **25,** 573-602.

Lele, S. and McCulloch, C.E. (2002). Invariance, identifiability, and morphometrics. *Journal of the American Statistical Association,* **97,** 796-806.

Lele, S.R. and Richtsmeier J.T. (2001). *An Invariant Approach to Statistical Analysis of Shapes.* Chapman and Hall, Boca Raton.

Lu, Y., Jiang, T. and Zang, Y. (2005). Single-trial variable model for event-related data analysis. *IEEE Transactions on Medical Imaging,* **24,** 236-245.

Mackay, C.E., Barrick, T.R., Roberts, N., DeLisi, L.E., Maes, F., Vandermeulen, D. and Crow, T.J. (2003). Application of a new image analysis technique to study brain asymmetry in schizophrenia. *Neuroimaging,* **124,** 25-35.

Marchini, J.L. and Ripley, B.D. (2000). A new statistical approach to detecting significant activation in function MRI. *NeuroImage,* **12,** 366-380.

Mardia, K.V., Bookstein, F.L. and Moreton, I.J. (2000). Statistical assessment of bilateral symmetry of shapes. *Biometrika,* **87,** 285-300.

Mardia, K.V. and Dryden, I.L. (1989a). Shape distributions for landmark data. *Advances in Applied Probability,* **21,** 742-755.

Mardia, K.V. and Dryden, I.L. (1989b). The statistical analysis of shape data. *Biometrika,* **76,** 271-281.

Mardia, K.V. and Dryden, I.L. (1994). Shape averages and their bias. *Advances in applied probability,* **26,** 334-340.

Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979). *Multivariate Analysis.* Academic Press, London.

Mathai, A.M. and Provost, S.B. (1992). *Quadratic Forms in Random Variables.* Marcel Dekker, New York.

McCarley, R.W., Wible, C.G., Frumin, M., Hirayasu, Y., Levitt, J.J., Fischer, I.A. and Shenton, M.E. (1999). MRI anatomy of schizophrenia. *Biological Psychiatry,* **45,** 1099-1119.

McDonald, B., Highley, J.R., Walker, M.A., Herron, B.M., Cooper, S.J., Esiri, M.M. and Crow, T.J. (2000). Anomalous asymmetry of fusiform and parahippocampal gyrus grey matter in schizophrenia: A postmortem study. *American Journal of Psychiatry,* **157,** 40:47.

Mechelli, A., Price, C.J. and Friston, K.J. (2001). Nonlinear coupling between evoked rCBF and BOLD signals: A simulation study of hemodynamic responses. *NeuroImage,* **14,** 862-872.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N, Teller, A.H. and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics,* **21,** 1087-1092.

Mitchell, M.W., Genton, M.G. and Gumpertz, M.L. (2003). A likelihood ratio test for separability of covariances. *Journal of Multivariate Analysis,* **97,** 1025-1043.

Rajarethinam, R., Sahni, S., Rosenberg, D.R. and Keshavan, M.S. (2004). Reduced superior temporal gyrus volume in young offspring of patients with schizophrenia. *American Journal of Psychiatry,* **161,** 1121-1124.

Ramsay, J.O. and Silverman, B.W. (2004). *Functional Data Analysis.* Springer, New York.

Rao, C.R. and Suryawanshi, S. (1996). Statistical analysis of shape of objects based on landmark data. *The Proceedings of the National Academy of Sciences,* **93,** 12132-12136.

R Development Core Team (2005). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna.

Richmond, N.J., Willett, P. and Clark, R.D. (2004). Alignment of three-dimensional molecules using an image recognition algorithm. *Journal of Molecular Graphics and Modelling,* **23,** 199-209.

Sibson, R. (1978). Studies in the robustness of multidimensional scaling: Procrustes statistics. *Journal of the Royal Statistical Society, Series B,* **40,** 234-238.

Sibson, R. (1979). Studies in the robustness of multidimensional scaling: Perturbational analysis of classical scaling. *Journal of the Royal Statistical Society, Series B,* **41,** 217-229.

Small, C.G. (1996). *The Statistical Theory of Shape.* Springer, New York.

Smith, S.M. (2002). Fast robust automated brain extraction. *Human Brain Mapping,* **17,** 143-155.

Smith, S.M., Jenkinson, M., Woolrich, M.W., Beckmann, C.F., Behrens, T.E.J., Johansen-Berg, H., Bannister, P.R., De Luca, M., Drobnjak, I., Flitney, D.E., Niazy, R.K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J.M. and Matthews, P.M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage,* **23,** S208-S219.

Stein, C. (1997). Lectures on the theory of estimation of many parameters. *In: Ibragimov, I.A. and Nikulin, M.S. (eds.) Studies in the Statistical Theory of Estimation, Part I, Proceedings of Scientific Seminars of the Steklov Institute, Leningrad Division,* **74,** 4-65.

Talairach, P. and Tournoux, J. (1988). *A stereotactic coplanar atlas of the human brain.* Thieme, New York.

Tanner, M.A. (1996). *Tools for Statistical Inference.* Springer, New York.

Ten Berge, J.M.F. (1977). Orthogonal Procrustes rotation for two or more matrices. *Psychometrika,* **42,** 267-276.

Theobald, C.M., Glasbey, C.A., Horgan, G.W. and Robinson, C.D. (2004). Principal component analysis of landmarks from reversible images. *Journal of the Royal Statistical Society, Series C,* **53,** 163-175.

Van Essen, D.C., Drury, H.A., Joshi, S. and Miller, M.I. (1998). Functional and structural mapping of human cerebral cortex: Solutions are in the

surfaces. *The Proceedings of the National Academy of Sciences,* **95,** 788-795.

Walker, J.A. (2000). Ability of geometric morphometric methods to estimate a known covariance matrix. *Systematic Biology,* **49,** 686-696.

Ward, G., Roberts, M.J. and Phillips, L.H. (2001). Task-switching costs, Stroop-costs, and executive control: A correlational study. *The Quarterly Journal of Experimental Psychology,* **54,** 491511.

Wible, C.G., Shenton, M.E., Hokama, H., Kikinis, R., Jolesz, F.A., Metcal, D. and McCarley, R.W. (1995). Prefrontal cortex and schizophrenia: A quantitative magnetic resonance imaging study. *Archives of General Psychiatry,* **52,** 279-288.

Worsley, K.J. and Friston, K.J. (1995). Analysis of fMRI time-series revisited - again. *NeuroImage,* **2,** 173-181.

Worsley, K.J., Liao, C.H., Aston, J., Petre, V., Duncan, G.H., Morales, F. and Evans, A.C. (2002). A general statistical analysis for fMRI data. *NeuroImage,* **15,** 1-15.