Longitudinal analysis of three-dimensional facial shape data

Sarah Jane Elizabeth Barry

A Dissertation Submitted to the
University of Glasgow
for the degree of
Doctor of Philosophy

Department of Statistics

May 2008

Abstract

Shape data encompass all the information that is left to describe a shape following removal of location, rotation and scale effects. Much work has been done in the analysis of two-dimensional shapes depicted by anatomical landmarks placed at points of importance. Less has been carried out in the area of three-dimensional shapes, particularly in terms of growth or change over time.

This thesis considers the analysis of such longitudinal three-dimensional shape data. In doing so, two well established but normally unrelated areas of Statistics are brought together: those of longitudinal data analysis (specifically, linear mixed effects models) and shape analysis. A recently proposed method of analysing longitudinal high-dimensional data is presented in a novel application within the area of shape analysis, illustrated by a study comparing the facial shapes of cleft-lip and palate children with controls as they grow from three months to two years of age. Both anatomical landmarks and facial curves are considered.

Chapter 1 broadly introduces the areas of shape analysis, linear mixed effects models and dimension reduction. Standard methods for measuring shapes are introduced, along with the difficulties inherent in analysing the resulting data. A broad overview of the methods of aligning individual shapes to remove the unwanted effects of location, rotation and scale is given, along with related geometrical issues in terms of the high-dimensional space in which a set of shapes resides. A general introduction to linear mixed effects models compares and contrasts them with simple linear models, explaining the reasons behind using them and presenting the different specifications of the conditional and marginal models. The area of dimension reduction is touched upon, specifically introducing B-splines and principal components analysis, with reference to the analysis of curves consisting of many points at

small increments to one another. The data from the cleft-lip and palate study are introduced, along with a discussion of the primary interest of the analysis and the issue of missing data.

Chapter 2 presents the statistical definition of a shape and introduces the area of statistical shape analysis in detail, specifically presenting the technicalities of shape space and distances, and methods such as Procrustes alignment of a set of shapes to remove unwanted effects. The concept of tangent coordinates is introduced as a projection of shape data into a Euclidean space, to enable the use of multivariate methods, and an outline given of thin-plate splines and deformations for the analysis of surfaces. Recent literature in the area of shape analysis is presented.

Further recent literature addressing the modelling of growth in shapes is presented in Chapter 3, which goes on to discuss the use of linear mixed models on univariate and multivariate longitudinal data. The difficulties of applying mixed models to multivariate data are discussed and a recently proposed alternative method introduced, which involves fitting mixed models to the responses on pairs of outcomes rather than the full set. A description of the `R` function written as part of this thesis to fit such pairwise models follows, and this is applied to simulated triangles and quadrilaterals as an illustration.

The initial application of the pairwise method to the cleft-lip and palate landmark data is presented in Chapter 4. The landmarks are described and the models are fitted to the tangent coordinate responses with different covariance structures for the random effects. The problems that arise and the deficiencies of the fitted models are extensively discussed.

Chapter 5 goes on to address the issues raised in Chapter 4. A method of aligning the individual shapes based upon a subset of landmarks is suggested, along with a model that assumes independence of coordinates between dimensions but correlation within, and the benefits of these approaches compared. A simulation study is carried out to investigate the reasons behind and effects of random effects correlations that are estimated as being close to one, concluding that the problem lies in small variances that are poorly estimated, but that this is unlikely to be of severe detriment to the fixed effects estimates. A method of taking the principal components of the tangent coordinates is suggested, where the model responses are the principal components scores, and this proves to be the most appropriate way of applying the pairwise models in terms of model fit and computational efficiency.

In Chapter 6, recent literature on the topic of curve analysis is presented, along with the way the facial curves are measured and the need for dimension reduction. Two methods are presented to this end: B-splines and principal components analysis, with the former suffering similar problems to the landmark analyses in terms of poorly estimated random effects variances, and the latter proving more successful. The application of the pairwise models to the principal components scores of the tangent coordinates provides a detailed analysis of the cleft-lip and palate data.

Issues surrounding model comparison are addressed in Chapter 7, with several hypothesis tests presented and applied to simulated data. Drawbacks with some of the tests when applied to high dimensional or longitudinal data result in poor performance, but a method suggested by Faraway (1997) and a modification of the likelihood ratio test, both using bootstrapping, show similarly successful results. These are subsequently used to test for any differences in the time trends for the cleft and control groups post-surgery and find that there are significant differences.

Condensed forms of this thesis have been presented at invited seminars and international conferences, and may be found in published form in Barry & Bowman (2006), Barry & Bowman (2007) and Barry & Bowman (2008).

Acknowledgments

The last three years have been fun, interesting and frustrating in equal measure, and there are many people that have made the frustrating times bearable and the fun times even more enjoyable. I would like to thank all of them and I have no doubt that I have forgotten one or two along the way. To them I apologise.

Firstly, I thank the Engineering and Physical Sciences Research Council (EPSRC), without whose funding there would have been no Ph.D. My thanks also to those of the Department of Computing Science and the Dental School who collected and provided the data analysed in this thesis.

To my supervisor, Professor Adrian Bowman, I am extremely grateful for your guidance and support throughout these three years. Thank you for always having faith in me when I lost it in myself; for being calm, wise and reassuring, especially in the last fraught few months; and for never pushing, but always quietly advising and supporting.

To Mum and Dad, thank you for always being there and for your ever-welcoming retreat in the North (whether it be house, hut or palace); for dogs, ponies and peace, and for your constant faith and confidence in me. To Cutie, for being my sidekick and for café solo and churros; and to Grandma, for always looking out for me.

To my Edinburgh friends, particularly Gill and Sally, for always being on the other end of the phone when I needed you and for years of fun and silliness; and Peet, for keeping me in gin. To Caroline, for putting up with me for a lovely year and a half! To all my friends at the Western tennis club, for hitting hundreds of balls at me and thereby keeping me sane; especially Matt, Kate, Russ and the rest of the Tuesday night doubles crowd.

To my fellow Ph.D. students in the Department of Statistics, thank you for welcoming me when I arrived and for the fun times throughout, particularly the heady days of the Glasgow RSC and WASA! For advice sought and kindly given, many thanks to Mitchum Bock, Ian Dryden, Steffen Fieuws, Geert Molenberghs, Agostino Nobile, Simon Rogers, Geert Verbeke and Simon Wood; and for further valuable advice two anonymous reviewers and many of the delegates of the 2006 and 2007 International Workshops on Statistical Modelling. Thank you also to my friends at the workshops for making my trips extremely enjoyable and especially for accompanying me into the Atlantic Ocean!

Finally, to Simon, my immeasurable thanks for your patience and logic, for gorgonzola and wine and intrepid adventure, for listening and being so tolerant and for making me laugh when I needed it most.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Shape

Shapes are all around us and we, as humans, can easily describe different shapes in many ways. Perhaps not as obvious is how to quantitatively measure, describe or compare a set of different shapes. Landmarks, curves and, more recently, surfaces have all been used to describe shapes. Landmarks may be marked onto objects or images by humans or computers, with the points chosen usually being positions of particular anatomical or geometrical importance. The placed landmarks must have consistency across objects so that their positions may reasonably be compared and this is not always an easy task. Curves are generally constructed by placing "pseudo-landmarks" at small increments along a curve between anatomical landmarks and, again, it is important that these are consistent between objects. The landmarks and curves can be placed on or extracted from an image using surface meshes of the individual objects, which may allow greater consistency, and these surface meshes may be used themselves to provide a more detailed description of an image for analysis. This thesis deals with landmarks and curves only, although the methods could be extended to surfaces.

An important aspect of statistical shape analysis is the necessary alignment of the objects in terms of location, rotation and possibly scale, so that any differences left between objects are attributable entirely to their shapes. Various methods of alignment have been proposed and

these are described in detail in Dryden & Mardia (1998) and will be covered more briefly in Chapter 2. These methods are not foolproof, however. For example, the Bookstein method of alignment involves selection of two landmarks as the reference and adjusting the remaining coordinates such that they correspond to the two as they did originally. The alignment is not symmetric in the choice of reference landmarks and correlations may be induced into the remaining coordinates as a result of the alignment (Dryden & Mardia, 1998). Care must also be taken with all alignment processes when attributing any differences between objects to discrepancies between their shapes and not to aspects of the alignment.

Shape data are by their very nature high-dimensional and this adds further difficulties into the analysis. It is not always clear how to treat landmarks that are defined by coordinates in three different dimensions; for example, is there any reason to believe that a y-coordinate of one landmark is correlated with a z-coordinate of another? Another issue is how to treat the landmark set as a whole, as often we are concerned with how the entire shape of the object is changing or how it compares to others, rather than in the comparison of individual points. Sometimes interest lies in particular areas of the object as well as the entire shape and so some relevant breakdown is required. One useful way of both considering the entire shape and decomposing it into different regions is presented by Bock & Bowman (2006), who calculate asymmetry scores of faces in order to compare the facial shapes of cleft-lip and palate and control infants. The scores may be computed using the landmarks from the entire face but also for different regions (such as the upper face, nose rim, upper lip) to find the areas that are contributing most highly to scores indicating high asymmetry. This approach, however, necessarily loses some of the shape information by aggregating all of the individual landmark positions.

If we wish to analyse the shape of an object as a whole, we encounter geometrical issues. With a set of landmarks in three dimensions, the entire shape resides in a high-dimensional non-Euclidean shape space. The shape space for pre-shapes (shapes which have been translated and scaled but not rotated) is a high-dimensional sphere or a hypersphere, and the shape resides as a point on the surface of this hypersphere. The landmark positions themselves, therefore, may not be treated as ordinary multivariate data and must undergo some transformation prior to analysis. One approach is to use tangent coordinates, which in-

volves projection of the point representing the shape onto a plane at a tangent to the high-dimensional shape space and this renders the distances between objects Euclidean so that normal multivariate methods may be applied. The resulting tangent coordinates are approximations to their shape space equivalents, but provided the shapes are close together in space this method ensures that the approximation is a good one. With a set of similar shapes in high dimensions, such as faces, the assumption of closeness is generally reasonable.

1.2 Linear mixed effects models

Generally there are two different types of data that are measured over time: time series, which takes many measurements on a (usually) small set of individuals, and longitudinal data, which takes a small number of measurements on a (usually) large number of individuals. We wish to deal with the latter, in which there is a small number of time points at which the individuals give responses, but in this case there is a large amount of data measured at any given time point.

Longitudinal data can be analysed using various different methods, such as latent variable models, generalised estimating equations and multivariate analysis of variance, amongst others, but the approach we will take is to fit linear mixed effects (LME) models. For one outcome variable, these models are similar in essence to simple linear models in which $Y = X\beta + \epsilon$ decomposes the response variable, Y, down into a fitted mean, $\mu = X\beta$, and a random error term, ϵ, which is assumed distributed as $N(0, \sigma^2)$. The vector of parameters, β, is estimated by $\hat{\beta} = (X'X)^{-1}X'Y$, where X is the known design matrix ($1 \times p$ in the univariate case, where p is the number of parameters in β). The difference between simple linear models and LME models lies in the error term, where in the former this is defined as purely measurement error and in the latter it is decomposed into a random effect part as well as measurement error. This generally arises because there have been multiple measurements taken on an individual unit. In our case, these measurements are taken at different points in time (ignoring the multiple outcomes, for now) but there is often a hierarchy involved; for example, measurements being taken on children within schools which are within regions, or multiple siblings within the same family. The issue is that the individual measurements are

no longer independent, a crucial assumption in the fitting of a simple linear model, and the result is likely to be that, although the mean effect of a covariate may be estimated correctly, its variance will usually be underestimated. This is partly due to the correlation between measurements on an individual being assumed zero and the effective sample size being larger than it really is because of the multiple measurements over time.

The LME model therefore becomes $Y = X\beta + Zb + \epsilon$, where Z is a known matrix of coefficients arising from the covariates and b is a vector of random effects, which is assumed to be distributed as $N(0, D)$, where D is a covariance matrix. The random error term is distributed as in the simple linear model. The random effects vector may contain solely a random intercept, or may also include a random slope. In the context of measurements over time, this allows each individual to have their own deviation from an overall or group mean at baseline (random intercept) and their own deviation from the overall or group trend over time (random slope). Generally the individual deviations themselves are not of interest, but the variance of the random effects will give a measure of the level of heterogeneity amongst individuals and is used in the estimation of the fixed effects parameters, β.

The LME model is usually specified in terms of the responses conditional on the random effects, so that in general terms (for Y univariate or multivariate) $Y|b \sim N(X\beta + Zb, \Sigma)$, where the covariance matrix of the random error, Σ, may be equal to $\sigma^2 I$ for I the identity matrix. However inferences are made based upon the marginal model, $Y \sim N(X\beta, ZDZ' + \Sigma)$, unless a Bayesian framework is employed for the analysis. The fixed effects parameter vector is estimated by (restricted) maximum likelihood by conditioning on the set of variance parameters, denoted as α, in $V = ZDZ' + \Sigma$, and the estimate is given by

$$\hat{\beta}(\alpha) = (X'V^{-1}X)^{-1}X'V^{-1}Y \tag{1.1}$$

(Verbeke & Molenberghs, 2000). The conditional model implies the marginal model, but the models are not identical and the reverse is not necessarily true. The marginal model does not explicitly assume that the variation between individuals is due to the random effects. However, this is only a drawback to the analysis if the values of the estimated random effects are of particular interest (Verbeke & Molenberghs, 2000).

In the case of a multivariate response where each individual has multiple outcomes measured at each time point, each of the individuals may have a different random intercept (and trend) estimated for each outcome and correlation allowed between their different responses. Individuals are assumed to be independent. The general multivariate model is defined as $Y = XB + U$, where Y is an $(n \times k)$-matrix for n individuals and k outcomes, X is the usual $(n \times p)$ design matrix, B is an $(p \times k)$ matrix of parameters and U is a matrix of normally distributed error terms (Mardia et al., 1979). The multivariate regression model may be written as $Y^v = X^* B^v + U^v$, where the v superscript indicates that the k columns of the respective matrices have been stacked on top of one another into a vector (Mardia et al., 1979). In the LME model case with random intercepts for the different outcomes (but no random slope), the error term, U^v, is normally distributed with mean zero and a covariance matrix, V, which has blocks on the diagonal containing the random intercept variance for a particular outcome plus the random error variance on the diagonal and the random intercept variance elsewhere. On the off-diagonal V has blocks that contain the covariance between the two relevant outcomes. The size of the blocks depends on the number of measurements over time. The fixed effects parameter vector is estimated as in (1.1).

1.3 Dimension reduction

The high dimensionality of shape data means that some method of dimension reduction is often required before multivariate methods may be applied in analysis. Principal components analysis (PCA) is a widely used method which finds the linear combinations of the original variables that explain the largest amount of the variance in the data. The principal components (PCs) are orthogonal to one another, and thus the process transforms the axes on which the data are measured. PCA is often useful in the case of high-dimensional data which are also highly correlated, as it uses the correlations between outcomes to identify underlying lower dimensional features of the data which provide a suitable summary of the variables (Mardia et al., 1979). The first few PCs, which explain the greatest variance in the data, are often plotted against one another as an exploration of the main directions of variation and for discrimination between different groups of individuals.

PCA is widely used in shape analysis, particularly as an exploratory measure. Dryden & Mardia (1998) recommend finding the PCs of the covariance matrix of the tangent coordinates as a useful way of exploring the main aspects of the variation of a shape. As a dimension reduction technique it is also very useful, since the inherently high correlation within shapes results in a low number of necessary components explaining a large amount of the variation in the data.

Another method of dimension reduction in the case of curves involves the use of B-splines. This is a nonparametric method, which describes a curve by a linear combination of basis functions and spline coefficients. If $x(t)$ represents the curve parameterised along $t \in [0, 1]$, then $x(t) = \sum_{k=1}^{K} c_k \phi_k(t)$, where the $\phi_k(t)$ are the basis functions and the c_k the corresponding spline coefficients (Ramsay & Silverman, 1997). A simple example is a quadratic polynomial, which may be written as a linear combination of the basis functions $1, t, t^2$. The B-spline joins a sequence of piecewise polynomials together smoothly at a chosen number of knots, which are points at particular intervals along the curve. The curve may be made smoother or rougher by varying the number of knots, K, where if K equals the number of points on the curve then it will interpolate it exactly and as K decreases, the fitted curve will become smoother. This may give a very close representation of a curve, but it will generally require a higher number of knots than PCA will require components. B-splines have not, thus far, been widely applied in shape analysis, but they have potential in terms of substantially reducing the dimension of the data.

1.4 Cleft-lip and palate data

The data that are used in the application come from a study comparing the facial shapes of a group of 49 infants with unilateral cleft lip and palate with those of a group of 100 age-matched controls, who had no facial deformity. Images of the children in three dimensions were captured at age 3, 6, 12 and 24 months using a sophisticated stereophotogrammetry system, which takes photographs from two different angles and thereby builds up a three-dimensional image, much in the same way that two eyes work together in human sight (Ayoub et al., 2003). The image is transformed into a surface mesh and some examples of these

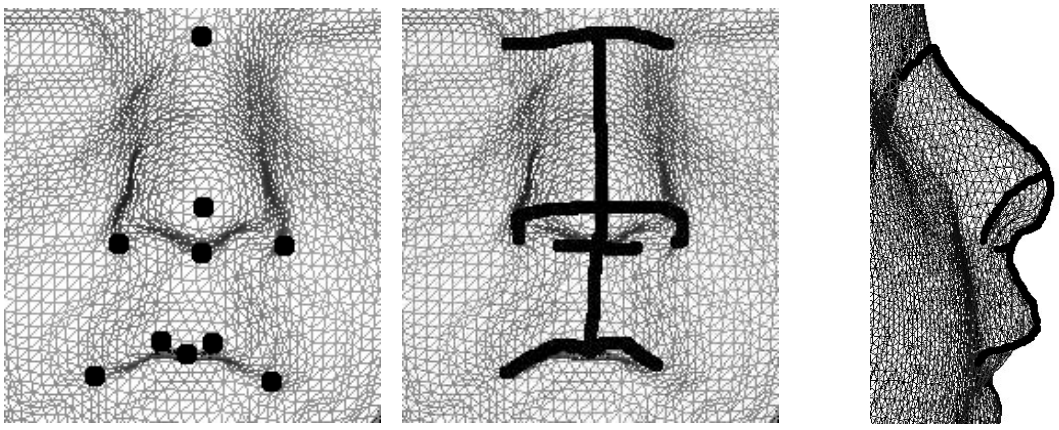

Figure 1.1: Low resolution surface mesh of an infant with landmarks (left) and curves in frontal (centre) and profile (right) views marked onto the mesh

meshes (with landmarks and curves marked) are displayed in Figure 1.1. The cleft group underwent surgical repair between the 3 and 6 month points. The full background of the study is given in Hood et al. (2004).

The data consist of both landmarks and curves marked onto surface meshes, as illustrated in Figure 1.1. Data are available on 34 landmarks on the face and 197 points constituting different curves. Important anatomical landmarks were manually marked onto the images of each of the individuals and the positions calculated of pseudo-landmarks at many small increments between various landmarks to produce the curves that cut the surfaces. It is clear that there will be substantial differences in terms of facial shape between the cleft children and the controls at three months of age since the cleft individuals have had no surgery at this point. However, the primary interest in this analysis is whether, following surgery, there are systematic differences in the facial shapes of the individuals within the cleft and control groups and whether these differences persist or reduce as the children grow. If there are persistent differences, then the particular areas of the face in which they lie and their extent will also be of interest.

1.5 Missing data

The cleft-lip and palate data has a reasonably high level of missingness, in large part due to the fact that the alignment process requires an entire set of data on an image at any particular time point (although it is not necessary that individuals have data at all time points). This is a particular problem for the curves, for which at 3, 6, 12 and 24 months there are respective response rates of 54%, 62%, 75% and 68% for the controls and 43%, 43%, 52% and 45% for the cleft group. For the landmarks the response rates are 82%, 92%, 91% and 90% for the control group and 60%, 58%, 63% and 67% for the cleft group at 3, 6, 12 and 24 months, respectively. This issue has not been explicitly dealt with here. To some extent, the LME model implicitly takes missing data into account because it uses information both from the group mean and from the individuals to estimate the time trends and other effects (Diggle et al., 2002). For example, if most of the dropouts at a later time point had high values of the response at earlier time points, then the estimated mean at the later point is likely to be higher than the empirical mean. This is unlikely to be enough adjustment if the probability of dropping out is related to the response, but there is no particular reason to believe that this is the case for the cleft study. It is possible that some of the more severely affected clefts may be more likely to drop out because of their medical problems, but otherwise it seems likely that there would be little relationship. In part the missing data is not due to dropout but to incomplete landmark or curve sets for an individual and there is certainly no reason to believe that this would be related to the response.

1.6 Overview of thesis

Chapter 2 will cover the basics of statistical shape analysis, including Procrustes' methods and based largely on Dryden & Mardia (1998), and will go on to present some recent developments in the field. Chapter 3 will review developments in the area of modelling and monitoring growth of shapes and will proceed by discussing mixed models for univariate and multivariate data, including problems with the implementation of these models for high-dimensional data. A pairwise methodology proposed by Fieuws & Verbeke (2006) will be

introduced as an alternative and an `R` program described for implementation of the method, while a small simulation study will illustrate the application to shape data. Chapter 4 involves the application of the pairwise approach to the cleft-lip and palate landmark data, with different models fitted and various problems demonstrated with the fitting of such models to shape data. Chapter 5 proposes and compares different ways of avoiding such problems and Chapter 6 uses the resulting conclusions to apply the pairwise approach to the cleft-lip and palate curve data, with a comparison of B-splines and principal components analysis for dimension reduction. Chapter 7 discusses some issues around model comparison for high-dimensional longitudinal data and illustrates various tests using simulations prior to application to the curve data. Chapter 8 gives a summary and suggests future work.

Chapter 2

Review of statistical shape analysis

2.1 Statistical Shape Analysis

This section will outline the foundations of statistical shape analysis, as a basis for the subsequent work in this thesis. Full details may be found in Dryden & Mardia (1998) and may be assumed to be the reference for all of the contents of the section.

2.1.1 Shape

The definition of shape is, as described by Dryden & Mardia (1998), "all the geometrical information that remains when location, scale and rotational effects are filtered out from an object." Shapes are generally defined by landmarks, points of correspondence on each object that match between and within populations. These are often anatomical landmarks with a biological meaning, such as the bridge of the nose or corner of the lips on a face, but they can also be mathematical landmarks (points with geometric meaning such as extreme curvature) or pseudo-landmarks which generally connect the other types of landmark.

A configuration is defined as a set of landmarks on a particular object and a configuration matrix X is of size $k \times m$ where k is the number of landmarks in the configuration and m the number of dimensions, usually two or three.

Since we generally wish to scale the object to remove any effect of size from the shape analysis, we require some measure of size. One commonly used measure is the *centroid size*,

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

where $\bar{X}_j = \frac{1}{k}\sum_{i=1}^{k} X_{ij}$, $C = I_k - \frac{1}{k}\mathbf{1}_k\mathbf{1}_k'$ and $||X|| = \sqrt{tr(X'X)}$.

Effectively therefore, $S(X)^2$ is the sum of squared distances from the mean point in each dimension. An alternative measure of size in two dimensions is *baseline size*:

$$D_{12}(X) = ||(X)_2 - (X)_1||,$$

the length between landmarks 1 and 2, where $(X)_i$ is the ith row of the configuration X.

2.1.2 Shape coordinate systems in two dimensions

Bookstein coordinates

In two dimensions different shape coordinate systems may be adopted to enable translation, rotation and scaling of the configuration and the consequent analysis of shape. One such system, *Bookstein coordinates*, for $k \geq 3$, fixes two of the landmarks (eg. at $(\frac{1}{2}, 0)$ and $(0, \frac{1}{2})$) and allows the other $k-2$ landmarks to describe the shape. If (x_j, y_j) in $m = 2$ dimensions, for $j = 1, \ldots, k$, describe the landmarks and the first two landmarks, (x_1, y_1) and (x_2, y_2), are fixed at, say, $(u_1, v_1) = (-\frac{1}{2}, 0)$ and $(u_2, v_2) = (\frac{1}{2}, 0)$ respectively, the remaining coordinates (u_j^B, v_j^B), $j = 3, \ldots, k$, are the Bookstein coordinates describing the shape, where

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

and $D_{12}^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2 > 0$.

This method of translating, rotating and scaling the configuration is often used in the first stages of an analysis but there is a lack of symmetry in choosing the baseline along with correlations induced into the coordinates, and points close to the origin tend to have lower variability than those further away. A mean shape of all the configurations may be calculated by taking the arithmetic mean of the Bookstein coordinates but this is not an appropriate method to use when interpreting shape variability because of the correlations induced into the shape variables.

Kendall coordinates

A further system, Kendall coordinates, transforms the configuration using a Helmert sub-matrix, H, where a Helmert matrix, H^F, is a $k \times k$ orthogonal matrix with its first row of elements equal to $1/\sqrt{k}$ and the remaining rows orthogonal to this. The Helmert sub-matrix drops the first row of H^F so that the transformed configuration HX does not depend on the original location, and therefore H is $(k-1) \times k$. The j^{th} row of H is

$$(\underbrace{h_j, \ldots, h_j}_{j}, -jh_j, \underbrace{0, \ldots, 0}_{k-j-1}),$$

where $h_j = -\{j(j+1)\}^{-1/2}$ and $j = 1, \ldots, k-1$.

The coordinate systems are similar and there are correspondences between them, but location is removed differently by each.

2.1.3 Procrustes methods

Procrustes methods allow for the removal of all location, rotation and scale effects for configurations of dimension higher than two and can be used to calculate a mean shape configuration.

Shape space and distances

Pre-shapes are configurations that have had the effects of location and scale removed, but not rotation. The pre-shape of a configuration X, under the Kendall coordinate system, is

$$Z = \frac{HX}{||HX||},$$

which is a $(k-1) \times m$ matrix. All of the possible pre-shapes containing k points in m dimensions, which have been translated and scaled, are contained in the *pre-shape space*, S_m^k, which is a hypersphere of unit radius in $(k-1)m$ real dimensions (since $||Z|| = 1$).

The *shape* of X is $[X] = \{Z\Gamma : \Gamma \in SO(m)\}$, where $SO(m)$ is the set of $m \times m$ special orthogonal matrices, Γ, such that $\Gamma'\Gamma = \Gamma\Gamma' = I_m$ and $|\Gamma| = 1$. The shape of X is invariant under location, rotation and isotropic scaling. The shape space, Σ_m^k, is of dimension $km - m - 1 - m(m-1)/2$.

The *full Procrustes distance* between two shape configurations, X_1 and X_2 is

$$d_F(X_1, X_2) = \inf_{\Gamma \in SO(m), \beta \in \mathbf{R}} ||Z_2 - \beta Z_1 \Gamma||,$$

where $Z_r = \frac{HX_r}{||HX_r||}$, $r = 1, 2$ and β is a scale parameter. The full Procrustes distance may be written as

$$d_F(X_1, X_2) = \left\{ 1 - \left(\sum_{i=1}^{m} \lambda_i \right)^2 \right\}^{1/2},$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{m-1} \geq |\lambda_m|$ are the square roots of the eigenvalues of $Z_1' Z_2 Z_2' Z_1$ and the smallest eigenvalue, λ_m, is the negative square root if and only if $\det(Z_1' Z_2) < 0$. The rotation matrix that minimises d_F is given by

$$\hat{\Gamma} = UV',$$

where $U, V \in SO(m)$ and $Z_2' Z_1 = V \Lambda U'$, $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_m)$.

The scale factor that minimises d_F is

$$\hat{\beta} = \sum_{i=1}^{m} \lambda_i,$$

where $0 \leq \hat{\beta} \leq 1$.

The *partial Procrustes distance* between the two configurations is

$$\begin{aligned} d_P(X_1, X_2) &= \inf_{\Gamma \in SO(m)} ||Z_2 - Z_1\Gamma|| \\ &= \sqrt{2}\left(1 - \sum_{i=1}^{m} \lambda_i\right)^{1/2}. \end{aligned}$$

The full and partial Procrustes distances differ in that the former is the Euclidean distance between pre-shapes that have been rotated and scaled to minimise the sum of squared distances between them, whereas for the latter the pre-shapes have only been rotated.

A further measure of distance, the *Procrustes distance*, is the closest great circle distance between Z_1 and Z_2 on the pre-shape sphere:

$$\begin{aligned} \rho(X_1, X_2) &= 2\arcsin(d_P(X_1, X_2)/2) \\ &= \arccos\left(\sum_{i=1}^{m} \lambda_i\right). \end{aligned}$$

The relationships between the Procrustes distance, ρ, and the full and partial Procrustes distances, d_F and d_P respectively, are displayed in the cross-section of the pre-shape sphere in Figure 2.1.

Full ordinary Procrustes analysis (OPA)

Full OPA involves least squares matching of two configurations, say X_1 and X_2, with k landmarks each in m dimensions, by minimising the distance between them following rotation, scaling and centring.

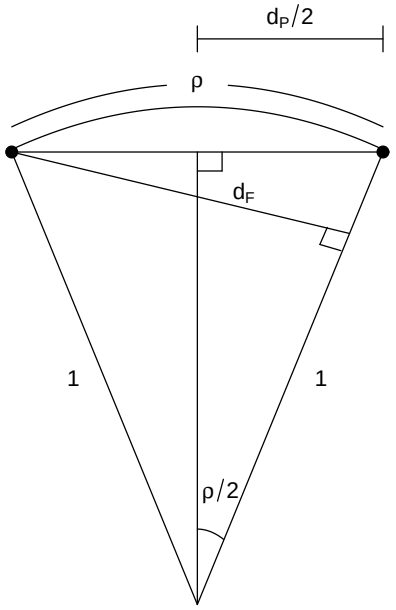

Figure 2.1: Cross-section of the pre-shape sphere and the relationship between the various Procrustes distances

Define

$$D^2_{OPA}(X_1, X_2) = ||X_2 - \beta X_1 \Gamma - \mathbf{1}_k \gamma'||^2,$$

where Γ is an $m \times m$ rotation matrix, $\beta > 0$ is a scale parameter and γ is an $m \times 1$ location vector. The configurations X_1 and X_2 are assumed to be centred, using $C = I_k - \frac{1}{k}\mathbf{1}_k\mathbf{1}'_k$.

The *ordinary (Procrustes) sum of squares*, $OSS(X_1, X_2)$, minimises D^2_{OPA} and gives:

$$\hat{\gamma} = 0;\ \hat{\Gamma} = UV',\ U, V \in SO(m);\ \hat{\beta} = \frac{tr(X'_2 X_1 \hat{\Gamma})}{tr(X'_1 X_1)}$$

and $OSS(X_1, X_2) = ||X_2||^2 sin^2 \rho(X_1, X_2)$, where $\rho(X_1, X_2)$ is the Procrustes distance defined above.

The *full Procrustes fit* of X_1 onto X_2 is

$$X_1^P = \hat{\beta} X_1 \hat{\Gamma} + \mathbf{1}_k \hat{\gamma}', \tag{2.1}$$

where X_1^P are defined as the Procrustes coordinates for configuration one and the parameter estimates result from the minimisation of the distance between the configurations,

$D^2_{OPA}(X_1, X_2)$.

The ordinary Procrustes fit is not reversible, i.e. $OSS(X_1, X_2) \neq OSS(X_2, X_1)$, unless the figures are the same size, so $\sqrt{OSS(X_1, X_2)}$ cannot be used as a distance measure. Instead the full Procrustes distance defined above is used, where

$$OSS\left(\frac{X_1}{||X_1||}, \frac{X_2}{||X_2||}\right) = d^2_F(X_1, X_2).$$

Full generalised Procrustes analysis (GPA)

Full GPA involves Procrustes matching of n configurations, $X_1, \ldots, X_n$, where $n \geq 2$. The following quantity, which is proportional to the sum of squared norms of pairwise differences, is minimised:

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=i+1}^{n} ||(\beta_i X_i \Gamma_i + \mathbf{1}_k \gamma_i') - (\beta_j X_j \Gamma_j + \mathbf{1}_k \gamma_j')||^2,$$

such that the centroid size $S(\bar{X}) = 1$, where the average configuration is $\bar{X} = \frac{1}{n}\sum_{i=1}^{n}(\beta_i X_i \Gamma_i + \mathbf{1}_k \gamma_i')$.

Unlike OPA, GPA is symmetric for two configurations, even if the objects are not the same size.

The *generalised (Procrustes) sum of squares* is defined by

$$\begin{aligned} G(X_1, \ldots, X_n) &= \inf_{\beta_i, \Gamma_i, \gamma_i} \frac{1}{n}\sum_{i=1}^{n}\sum_{j=i+1}^{n} ||(\beta_i X_i \Gamma_i + \mathbf{1}_k \gamma_i') - (\beta_j X_j \Gamma_j + \mathbf{1}_k \gamma_j')||^2 \\ &= \inf_{\beta_i, \Gamma_i, \gamma_i} \sum_{i=1}^{n} ||(\beta_i X_i \Gamma_i + \mathbf{1}_k \gamma_i') - \frac{1}{n}\sum_{j=i+1}^{n} (\beta_j X_j \Gamma_j + \mathbf{1}_k \gamma_j')||^2 \\ &= \inf_{\mu: S(\mu)=1} \sum_{i=1}^{n} OSS(X_i, \mu) = \inf_{\mu: S(\mu)=1} \sum_{i=1}^{n} \sin^2 \rho(X_i, \mu). \end{aligned}$$

Using these methods, we can calculate the *full Procrustes estimate of mean shape*:

$$\hat{\mu} = \arg \inf_{\mu: S(\mu)=1} \sum_{i=1}^{n} \sin^2 \rho(X_i, \mu) = \arg \inf_{\mu: S(\mu)=1} \sum_{i=1}^{n} d^2_F(X_i, \mu).$$

The full Procrustes mean may be estimated by calculating the arithmetic means of each co-ordinate, following full Procrustes matching, so $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i^P$, where X_i^P are the Procrustes coordinates for individual i.

Partial Procrustes analysis involves Procrustes matching of configurations that are not of unit size and is generally used for studying size and shape together, but this is not of particular interest here.

2.1.4 Tangent coordinates

The *tangent space* is a linearised space to the shape space, about the average shape. If, for simplicity, we think of the shape space as a sphere (as is the case for triangles in two dimensions), then the tangent space is the two-dimensional plane at a tangent to the point on the sphere that (usually) corresponds to the average shape. If points close together in the shape space are then projected onto the tangent space they can be assumed to approximate the actual points. The Euclidean distance in tangent space is therefore a good approximation to shape distances in shape space and general multivariate methods may be applied to the points in the tangent space. The tangent projection does not depend on the original rotation of the shape.

Define the vectorize operator, $\mathrm{vec}(X)$, such that

$$\mathrm{vec}(X) = (x_1', x_2', \ldots, x_m'),$$

where, if X is $l \times m$ with columns $x_1, x_2, \ldots, x_m$ then $\mathrm{vec}(X)$ is an lm-vector.

A pole, γ, $((k-1) \times m)$, which is often the mean shape, is chosen on the pre-shape sphere. The pre-shape, Z, is rotated to match γ as closely as possible by multiplying by $\hat{\Gamma}$, choosing $\hat{\Gamma}$ to minimise $||\gamma - Z\Gamma||^2$. The projection onto the tangent plane at γ gives the tangent coordinates:

$$v = (I_{km-m} - \mathrm{vec}(\gamma)\mathrm{vec}(\gamma)')\mathrm{vec}(Z\hat{\Gamma}).$$

A useful approximation to the tangent coordinates is to use the Procrustes residuals

$$r_i = w_i^P - \left(\frac{1}{n}\Sigma_{i=1}^n w_i^P\right), i = 1, \ldots, n, \tag{2.2}$$

where $w_1, \ldots, w_n$ are the original coordinate points and $w_1^P, \ldots, w_n^P$ the corresponding Procrustes adjusted points.

2.1.5 Thin-plate splines, deformations and warping

Although not explicitly relevant to this thesis, a brief summary of thin-plate splines, deformations and warping is given here as an introduction to the following section on recent literature. A thin-plate spline may be applied in multiple dimensions to interpolate data, much in the same way that a natural cubic spline is used in one dimension (Dryden & Mardia, 1998). If a deformation of the shape configuration $T = (t_1, \ldots, t_k)'$ into the configuration $Y = (y_1, \ldots, y_k)'$ is defined as

$$y = \Phi(t) = (\Phi_1(t), \ldots, \Phi_m(t))',$$

where $t_j, y_j \in \mathbf{R}^m$, then in two-dimensional space a pair of thin-plate splines may be written as

$$\Phi(t) = (\Phi_1(t), \Phi_2(t))' = c + At + W's(t),$$

for t a (2×1) vector, $s(t) = (\sigma(t - t_1), \ldots, \sigma(t - t_k))'$ is $(k \times 1)$ and

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

This can be written in matrix form, along with a further six constraints, as

$$\begin{bmatrix} S & 1_k & T \\ 1_k' & 0 & 0 \\ T' & 0 & 0 \end{bmatrix} \begin{bmatrix} W \\ c' \\ A' \end{bmatrix} = \begin{bmatrix} Y \\ 0 \\ 0 \end{bmatrix}, \tag{2.3}$$

for $(S)_{ij} = \sigma(t_i - t_j)$. If Γ is defined as the matrix on the furthest left in (2.3), then its inverse may be written as

$$\Gamma^{-1} = \begin{bmatrix} \Gamma^{11} & \Gamma^{12} \\ \Gamma^{21} & \Gamma^{22} \end{bmatrix}$$

and the *bending energy* matrix is defined as $B_e = \Gamma^{11}$. The partial warps are used to explain the components of a thin-plate spline and are defined as

$$R_j(t) = Y'\lambda_j\gamma_j\gamma_j's(t),$$

for $j = 1, \ldots, k-3$ and λ_j, γ_j, the jth eigenvalue and eigenvector, respectively, of B_e and $s(t)$ as defined above. Full details may be found in Dryden & Mardia (1998).

2.2 Recent Literature in Shape Analysis

In recent years there have been various developments in the field of shape analysis, for data in two and three dimensions. The following section presents some of the developments which are most relevant to this thesis. Work concerning the modelling of growth is introduced in the next chapter on longitudinal shape modelling. Principal components analysis (PCA) occurs frequently in the cited literature as it is widely used in shape analysis as an effective method of both reducing the dimensionality of a problem, and of analysing the main modes of variation in shape. The principal components may be interpreted as a measure of size and/or shape and change along each PC can be plotted to study growth or differences in shape (Dryden & Mardia, 1998).

2.2.1 Statistical shape analysis

Statistical shape analysis became a recognised branch of statistics in the late 1970s and early 1980s, following a period when it was considered as a special case of multivariate analysis (Bookstein, 1991). Seminal work by Bookstein (1978, 1984*a*,*b*), Kendall (1984) and Goodall (1983) brought together the fields of geometry, biology and statistics to provide a framework for analysing shapes, realising the initial aims of Thompson (1917). Dryden &

Mardia (1998) provide an authoritative synthesis of recent work and other useful references are Bookstein (1991) and Small (1996). Shape analysis is currently used for many different applications in fields such as computer science, medicine and forensic science, as well as statistics.

In recent work, Mardia et al. (2000) tested a hypothesis of bilateral symmetry of shapes, $H_0 : \mu = QA\mu$, where μ is a centred and scaled configuration with $k + l$ landmarks in m dimensions, Q is a permutation matrix that relabels the points in the configuration after reflection and A is some orthogonal matrix with determinant -1 that reflects the points. Therefore, under the null hypothesis, if μ is reflected and relabelled, the resulting transformed configuration is the same as its original.

Let X be a configuration and $Y = QAX$ its reflection. In the isotropic case, $\text{vec}(X_i^P) \sim N_{(k+l)m}(0, \sigma^2\Sigma)$, where X^P are the Procrustes tangent coordinates for configuration X and Σ is a multiple of the identity. The total sum of squared distances, SS_T, between the Procrustes coordinates of the configurations can be broken down into $SS_T = SS_B + SS_W$, representing respectively the SS between sides, or directional asymmetry, and the SS within cases from the group mean, or fluctuating asymmetry.

The test of bilateral asymmetry in this case is:

$$\frac{SS_B/d}{SS_W/(n-1)d} \sim F_{d,(n-1)d}$$

under H_0, where $d = mk/2 + l - m$.

In the non-isotropic case, $\text{vec}(X_i^P) \sim N_{(k+l)m}(0, \sigma^2\Sigma)$, where Σ is allowed to be any positive semi-definite matrix. A Hotelling's T^2-test of the difference between the mean configuration of $X_1, \ldots, X_n$ and the mean of $Y_1, \ldots, Y_n$ has a test statistic of $T^2 = n(\bar{x} - \bar{y})'P^-(\bar{x} - \bar{y})$, where $x_i = \text{vec}(X_i)$ and $y_i = \text{vec}(Y_i)$ and P^- is the Moore-Penrose inverse (Mardia et al., 1979) of P the singular estimator of $\text{cov}(x - y)$, which is of order $(mk - m) \times (mk - m)$. Under the null hypothesis of bilateral symmetry,

$$\frac{T^2}{n-1} \cdot \frac{n-d}{d} \sim F_{d,n-d}.$$

The authors also present a permutation version of this test by suggesting that if $n < d$, one can calculate SS_B and do a Monte-Carlo test by randomly permuting the assignment of X_i^P or Y_i^P to the reflected state. Finally a method is introduced to estimate the amount of variability of a reflection axis across shapes by considering the covariance structure of the relation between sides. The sum of cross-products without mean centring is $\sum(x-y)(x-y)'$, i.e. the matrix of cross-products about 0; and of the principal components of this matrix, the first represents the direction of greatest sum of squares of $x-y$ around 0, the symmetric form. A further paper by Kent and Mardia presents these methods in a more rigourous format (Kent & Mardia, 2001)

Larsen et al. (2002) considered alternatives to PCA for two- and three-dimensional tangent shape space coordinates using non-Euclidean metrics. Their paper compared ordinary PCA, maximum autocorrelation factor analysis (MAF) and minimum noise fraction transformations (MNF). Whilst PCA seeks linear combinations of variables that exhibit maximum variance, MAF analysis, where an order of observations is available (eg. pixels of images), seeks and applies linear combinations that maximise autocorrelation (ie. spatial correlation); and MNF is PCA in a metric space defined by a noise covariance matrix estimated from the data.

The first MAF is a linear combination of Z_k (where Z_k is a multivariate stochastic variable) that exhibits maximum correlation with the same linear combination of $Z_{k+\delta}$. Further MAFs are uncorrelated with the first. The first MNF is a linear combination of Z_k that exhibits maximum signal variance to noise variance ratio, where Z_k can be expressed as a sum of signal plus noise ($Z_k = S_k + N_k$). Further MNFs are also uncorrelated with the first.

The authors apply these techniques to a dataset of two-dimensional lung shapes and one of three-dimensional landmarks of surfaces of human mandibles over time. The main conclusion is that MAF provides better interpretability in two dimensions than PCA. In three dimensions the MNFs show different patterns of growth to the PCs and different relations between the different components and size and age, but there is no general conclusion about the relative advantages of the methods.

Theobald et al. (2004) used PCA to summarise variation in two-dimensional reversible images, such as longitudinally sliced halves of carrots and carcasses of live sheep. They were

particularly interested in the effect of including the mirror images of the originals in the analyses. In the carrot analysis this removed bias caused by non-random selection of initial halves and as a result made some of the PCs more interpretable. In the sheep analysis, inclusion of mirror images also simplifies the interpretation of the PC weights, but for larger sample sizes consistent departures from bilateral symmetry in animals may cause bias when assuming reversibility.

Dryden et al. (2008*a*) proposed a multi-dimensional scaling (MDS) approach as a method of calculating mean reflection shapes, as an alternative to the full or partial Procrustes mean. This approach provides a more easily computable mean shape and although the representation applies only to reflection shapes, the authors claim that this is equivalent to working in one half of the total shape space which is generally the case in most applications anyway. The MDS mean shape is shown, using a simulation study, to be generally similar to the full and partial Procrustes mean shapes and is a lot less computationally intensive. However MDS tangent projection is somewhat different to the Procrustes tangent projection, despite similar means, which results in different PC loadings.

The authors go on to present hypothesis tests based on MDS mean shapes and show that the results are very similar to equivalent tests on Procrustes mean shapes.

Bowman & Bock (2006) explored issues of graphical exploration of three-dimensional shape data. They used GPA to align and scale configurations $X_1, \ldots, X_n$, before using tangent space as an approximation to shape space and Procrustes residuals as an approximation to tangent coordinates. On data such as these, standard multivariate methods such as PCA may be applied. The PCA results can be transformed back into shape space to show the changes in shape along the principal components. The authors represent facial shape curves as dense sets of discrete points and plot curves that correspond to ± 3 standard deviations on PC scales to show the changes, rather than using arrows as is generally done in two dimensions.

Data from shapes are very high dimensional so the proportions of variation explained by individual components of variation are small. There are many ways in which faces may vary and it is difficult therefore to examine the full ranges of variation using individual

components. The authors suggest a tour approach to get around this problem, in which points on the shape hyperspace are randomly selected and elements of variation in the data are captured by moving smoothly from one to the next.

The methods presented in the paper are applied to a longitudinal study of facial development in young children. Measurements were taken on 100 controls and 23 and 26 unilateral cleft lip and cleft lip and palate cases, respectively, at 3, 6, 12 and 24 months of age. Interest lies in the differences between the cleft and control groups and this can be studied by summarising the pattern of all of the data simultaneously and then locating the clefts and controls on common scales, for example by plotting the PC scores. By working in the subspace defined by the first few PCs, the bulk of variation in the data can be captured. The authors compare individual cleft cases to the closest point in the normal range of controls, with this being defined as the range that includes all individuals with a Mahalanobis distance from the mean that is within the lower 95% of the χ^2_f distribution (for f the number of PCs). They can therefore ascertain the differences between cleft cases and the most extreme normal control in the same direction. The authors point out that some important clinical differences between the groups are only highlighted by the fourth or fifth PC, and so it is important to consider these higher components in the analysis.

The authors present a severity index to measure the extent of abnormality using the probability of lying beyond the Mahalanobis distance, $d(x) = (x - \bar{x}_c)'\hat{\Sigma}^{-1}(x - \bar{x}_c)$, between a cleft point, x, and the control mean, $\bar{x}_c$, where $\hat{\Sigma}$ is the estimated covariance matrix of the controls. The vector x is f-dimensional, where the first f PCs are being considered, and $d(x) \sim \chi^2_f$ approximately.

A permutation test is also constructed for differences in the PC scores between the two groups. This may be done for f PCs ($f \geq 1$) so it can detect differences between the groups even in high PCs, which is especially useful when individual PCs explain small proportions of variation.

Bock & Bowman (2006) presented a paper developing methods to measure and analyse asymmetry in faces. The interest lies in shape rather than size so all configurations, X, are scaled such that $||X|| = 1$. The reflected image of each configuration, X_R, is constructed in

an arbitrary plane and matched with X using ordinary Procrustes analysis. The columns of X and X_R are centred around 0 so both have the same centroid and the asymmetry score for an individual is defined as

$$A = \frac{||X - X_R\hat{\Gamma}||^2}{k},$$

for k landmarks. The score may be thought of as the average displacement between each configuration and its reflection, as a squared distance per landmark. The authors show that since the score is based upon a sum of squares, it can be decomposed into components based on different facial features. The score may also be decomposed into different sources of asymmetry: intrinsic asymmetry of a feature calculated by the asymmetry score in isolation from the object; positional asymmetry as in the centroid position difference between the configuration and its reflection; and orientational asymmetry, which allows a feature rotation matrix instead of a global rotation matrix.

The authors present a cross-validation approach to decomposition, which matches an object to its reflection on landmarks from all the features except the one of interest. Describing shape by the use of curves is also considered by finely discretising the curves and using the same methods as for landmarks, or by analysing the curves themselves using functional representations of reflected and matched curves. To measure asymmetry in this latter case, the following score could be constructed:

$$\int_0^1 \left\{ \left[x(t) - x'(t)\right]^2 + \left[y(t) - y'(t)\right]^2 + \left[z(t) - z'(t)\right]^2 \right\} dt,$$

where $\{x'(t), y'(t), z'(t)\}$ is a functional representation of a reflected and matched curve and the dummy variable t runs in the opposite direction to the original (due to the reflection). However it is generally more convenient to finely discretise the curves and apply landmark methods. The data used to illustrate the methods in this paper are from the same facial development study as in the previously discussed paper (Bowman & Bock, 2006).

Kaliontzopolou et al. (2007) considered the size and shape of two species of lizard. Measurements were taken of distances between points on the bodies of 55 males and 48 females of one species and 53 males and 46 females of the other. Analysis of variance was used to

test for sexual differences in terms of various measured distances. The authors proceeded by comparing this method to the use of geometric morphometrics. Two-dimensional photographs were taken in two different views of the lizards' heads, with 30 landmarks marked on one view and 16 on the other. GPA was carried out on all of the images and MANOVA on the partial warp matrix. The authors calculated the relative warp scores for the males and females in their data and produced deformation grids by the regression of shape variables versus the scores. They used the Mahalanobis distance as a measure of the extent of sexual dimorphism. There was a discordance of results between the methods using distances between points on the lizards and those using the partial warps, but the authors believe that the two approaches are highlighting different effects of size and shape. Fink et al. (2005) used a similar partial warp approach and presented their results as deformation grids according to different second to fourth digit ratios. These authors also tested for a relationship between facial shape and the digit ratio by regressing the ratio on the Procrustes coordinates.

Hennessy et al. (2007) carried out a study of the relationship of schizophrenia to facial shape for both males and females. They used three-dimensional surface imaging on 37 male and 32 female sufferers of schizophrenia and 58 male and 34 female control subjects. Thin-plate splines were used to obtain a smooth interpolation of the whole surface of each face and to provide pseudo-landmarks. Twenty-four anatomical landmarks were also placed onto the images. The authors applied GPA to the images and performed PCA on the resulting approximate tangent coordinates to obtain the elements of highest variability in the sample and reduce the dimensionality. Goodall's F-test was applied to the landmark data and Hotelling's T^2 test to the PC scores, separately for males and females, to test for any difference between the schizophrenia and control groups in facial shape. The authors also used logistic regression with diagnosis as the dependent variable and the PCs as the independent variables, in order to find which PCs contributed to the likelihood of having schizophrenia. This was also performed separately for males and females. The results are illustrated for both males and females by producing mean schizophrenia and control facial surface images, thereby allowing observation of the main areas in which the groups differ. Goodall's F-test and Hotelling's T^2-test both showed significant differences between schizophrenics and controls amongst the females but not the males. Procrustes ANOVA showed a significant gender

by group interaction.

2.2.2 Shape analysis in other fields

There are a number of papers in the shape analysis area from the computing science perspective (Golland et al. (2001, 2005) and Srivastava et al. (2004) are examples), but these tend to be focused somewhat more on classifying and recognising shapes than modelling change.

Shape analysis has also been used for studies in medicine; some examples in which it has been applied to three-dimensional data in the study of cleft lip and palate infants and relationships between facial characteristics of non-cleft children with body measurements are Hood et al. (2003) and White et al. (2004). Other examples in medicine include Mutsvangwa & Douglas (2007), Styner et al. (2005), Free et al. (2001), Dean et al. (2000) and Hennessy et al. (2005). Examples in forensic science are Franklin et al. (2007*b*) and Kieser et al. (2007). Procrustes analysis along with principal components analysis have been applied in anatomy, in particular to the comparison in terms of cranial shape and shapes of other bones between hominids and different populations of modern humans, and in paleoanthropology, in the study of ancient relatives of the modern human, such as Neanderthals. Examples in anatomy include Lockwood et al. (2002), Harmon (2007) and Franklin et al. (2007*a*) and in paleoanthropology, Bacon (2000), Harvati (2003*a*,*b*) and Schaefer et al. (2004).

A particular example of the application of shape analysis using surfaces is Šefčáková et al. (2003), who considered sexual dimorphism in photographs of 27 human fossils ranging in age from 25 - 27000 BP and obtained from glass plate negatives. The authors standardised the shapes using the Bookstein coordinates method and the photographs were considered in various two-dimensional views: the frontal, lateral, occipital, basal and vertical. The authors' interest lay in the inter- and intra-sex differences. Glass plate negatives were obtained of the skulls of two males and three females and each skull was compared with every other in the five different views. The authors used thin-plate splines to detect affine (stretchable) and non-affine (deformational) changes and this enabled the comparison of two different objects. They were also able to consider all views together using deformation penalties resulting from bending matrices. The authors found that there were substantial differences

between the sexes, but that there were also differences between females while males were more homogeneous.

Chapter 3

Longitudinal shape modelling

3.1 Modelling growth in shapes

There have been various advances in the field of longitudinal shape modelling, largely related to modelling growth in shapes and most often in two dimensions. The work considered here is restricted to papers that have a statistical slant, but there has also been some considerable work done in computing science on geometric morphometrics and also facial recognition.

O'Higgins & Jones (1998) considered three-dimensional growth in mangabey skulls and, in particular, sexual dimorphism in terms of size-and-shape. They performed generalised Procrustes analysis (GPA) on the 31 facial landmarks of 49 individuals, and principal components analysis (PCA) on the tangent coordinates of the resulting translated and rotated shapes. A comparison was made between the results of carrying out PCA on the exact and approximate tangent coordinates, the latter being the Procrustes residuals. This comparison corresponds to a measure of the concentration of the data, since the more concentrated the data, the closer the Procrustes residuals will be to the exact tangent coordinates. The authors found little difference between the two sets of results but they did find evidence of sexual dimorphism in terms of size and shape changes due to increasing size just prior to maturity. They studied the main principal components (PCs) of the tangent coordinates, finding that size was the dominant factor (the first PC) and that there were greater differences between males and females in the second PC for older than younger specimens, indicating the divergence of

shape between the sexes shortly before reaching maturity. The authors presented graphical models showing the effects of the PCs on the skull shape deformation, using triangles to define the surfaces between marked landmarks.

Morris et al. (1999*b*) used the Iterative Closest Point (ICP) algorithm to study symmetry of faces as children grow, again in terms of size-and-shape. The ICP algorithm was used to find the best approximation to the medial plane of a face. A plane, P, is initially chosen by hand and the reflection in P denoted as R_P, while $X = \{x_i \in \mathbf{R}^3 : 1 \leq i \leq n\}$ is a large set of landmarks which describe the head. The closest point to x_i in $R_P(x_i)$ is denoted as y_i, where x_i is the i^{th} point in the configuration X. The distance between the two is defined as $d_i = ||x_i - y_i||$, and weights as $w_i = 1$ if $d_i < d$ or 0 otherwise, where d is chosen as twice the width of the nose. A plane, Q, is then found that minimises

$$\frac{1}{\Sigma w_i} \Sigma w_i |x_i - R_Q(y_i)|^2$$

over Q. Set $P = Q$ and iterate.

The data came from a total of 28 scans on five subjects between the ages of 5 and 16 years. On the faces 14 landmarks were marked in two dimensions, using bi-tangent lines to find points of extreme curvature. GPA was carried out on the landmarks, followed by PCA. The authors found that after PCA was performed on the tangent coordinates of the shapes the most important change as the subjects grew was in size. They speculate that their finding of a more sloping forehead shape with increased age was likely due to the method of choosing the landmarks, since as the nose grows the bi-tangent line joining it to the forehead will slope backwards more, and this emphasises the importance of landmark choice in studies such as this.

A further conference proceedings paper by Morris et al. (1999*a*) along similar lines plotted the second PC against the first and showed backtracking with age along the first because of misregistration of profiles resulting from different stances, once again highlighting the problems with gathering shape data from faces. The ICP algorithm was used in this case to register pairs of foreheads, since little change is expected in foreheads as children grow, and this method looked more stable in plots but did not give quantitatively better results,

in that the correlation between the first PC and age was no higher.

Morris et al. (2000) suggested a parallel curve model for change in 2-dimensional shape over time:

$$\gamma(t) + u_i,$$

where $u_i \in \mathbf{R}^{2K-4}$ for $i = 1, \ldots, n$ individuals with $k = 1, \ldots, K$ landmarks in two dimensions. Under this model $\gamma(t)$ is a linear combination of polynomials of degree $m = 1, \ldots, M$, which attempt to describe the shape changes over time. The model assumes that individuals follow parallel curves over time but they deviate from each other by some individual-specific constant.

If $y_{it} = (y_{it1}, \ldots, y_{it(2K-4)}) \in \mathbf{R}^{2K-4}$ are the Procrustes tangent coordinates describing the shape for individual i, then explicitly the model is

$$y_{it} = u_i + P_1(t)v_1 + P_2(t)v_2 + \ldots + P_M(t)v_M + e_{it},$$

where the basis $P_m(t) : m = 1, \ldots, M$ spans the function space $\mathcal{P}$, $v_m = (v_{1m}, v_{2m}, \ldots, v_{(2K-4)m})$ is a vector of coefficients and $e_{it} \in \mathbf{R}^{2K-4}$ is a vector of error terms.

Using MANOVA, the model was fitted to the rat data documented in Bookstein (1991), in which eight landmarks were measured in two dimensions on eighteen rat skulls at eight different ages. The authors found that the linear term in the model dominated with respect to describing the change in shape over time, but that the individual, quadratic and cubic terms were also statistically significant. A test of the hypothesis of parallelism showed significant evidence of individual linear time effects, but these were very small compared to the group effect so may not be scientifically relevant. The authors also applied the methods to Bookstein coordinates instead of Procrustes and obtained similar results.

Le & Kume (2000) took a different approach by avoiding the use of tangent coordinates and carrying out the analysis of growth directly in shape space before reinterpreting the results in a Euclidean context. The main idea is that each shape is represented by a point in shape space and that as the object changes it traces out a path in that space. The *least energy principle*, which implies that the movement of a particle from one point to another will take

the shortest route that external constraining forces will allow, is used to assume that the change follows preferred paths in the shape space such as geodesics. Hence in the short term the change in shape should follow a geodesic with random fluctuations about the path which correspond to noise, and if this is the case, we should be able to predict future change along that path.

The methods are applied to the same two-dimensional rat data as in the previously discussed paper (Morris et al., 2000), with the aim of modelling the shape change over time using principal coordinate analysis, which models data as the distances between points, rather than the points themselves. The authors found that the first principal coordinate explained about 92% of the variation in shape, which implies that the change in shape does take place approximately along a geodesic of shapes, with only small amounts of deviation. The differences in the first principal coordinates of the mean shapes at times $t = 1, \ldots, 8$ approximate the magnitude of the corresponding changes in shape.

Kent et al. (2000) suggested a further model for growth, where if $s \in \mathbf{R}^m$ is a point in an object at time zero, the position of s at time t is modelled by

$$\phi(s,t) = s + \sum_{l=1}^{r} f_l(s) g_l(t).$$

The functions $f \in \mathcal{F}$ and $g \in \mathcal{G}$ specify possible directions and rates of growth, respectively, for some rank r which represents the complexity of the model. If $\mathbf{x}_{nh}$ is a $k \times m$ matrix, for $n = 1, \ldots, N$ individuals and times $t_1, \ldots, t_H$ then the model takes the form

$$\mathbf{v}_{nh} = \mathbf{v} + \sum_{l=1}^{r} f_l^*(\mu) g_l(t_h) + \epsilon_{nh},$$

where $f^*(s)$ is the adjustment of $f(s)$ produced by adding a linear function of s.

Kent et al. (2001) went on to present functional models of growth for landmark data, by defining a roughness penalty on functions in space and time, specifying smoothness of directions and rates of growth, respectively. An $H \times H$ "bending energy" matrix, B, has eigenvectors γ_h that combine into an $H \times (H-1)$ matrix G. A further $K \times K$ bending energy matrix has eigenvectors that combine into an $K \times (2K-4)$ matrix F.

Then if W is the $KM \times H$ matrix of Procrustes coordinates for $h = 1, \ldots, H$ different times we fit the model:

$$W = \nu \mathbf{1}'_H + FAG' \Rightarrow F'WG = A,$$

where A is a $(2K-4) \times (H-1)$ matrix of coefficients and ν is an intercept of no interest.

When fitted to the rat data the best fitting model has linear growth in time in a general spatial direction and quadratic growth in time that is restricted to a single direction in space.

An alternative method of monitoring growth, when longitudinal data are unavailable, was employed by Hutton et al. (2003). The authors had three-dimensional facial shape data on individuals at different ages and they wished to compute average growth trajectories, despite having one image only from each participant in the study. They considered size-and-shape and used thin-plate splines to interpolate the surfaces between the measured landmarks and kernel smoothing to compute an average face for a given age while smoothing out individual variation; and Procrustes registration was carried out to remove the effects of location and rotation. Clear differences were displayed between males and females and size was dominant, particularly for males.

Kume (2007) extended the ideas of their earlier paper Le & Kume (2000) and presented a method of fitting smooth curves to a series of two-dimensional shapes measured over time, where the shape changes are large. They used Riemannian manifolds to produce shape space smoothing splines which fitted exactly to the data points if the smoothing parameter was zero, and corresponded to the geodesic curve if the parameter was infinitely large. The authors applied their method to various examples of three-dimensional human movement data (lower back, shoulder, wrist and index finger) described by four landmarks moving in time. A test was suggested of the null hypothesis that the mean path across time of the shape was a geodesic and there was strong evidence to reject this in favour of the smoothing spline. The authors suggest that if the shape changes are small, the method gives similar results to methods using Procrustes tangent projection.

Velemínská et al. (2006) studied change in facial shape of cleft-lip and palate children as compared to controls, from 10 to 15 years of age. Landmarks were obtained from two-dimensional X-ray films taken at each of the two time points and the coordinates were matched using

the Bookstein transformation. PCA was carried out on the resulting Bookstein coordinates and a paired Hotelling's T^2-test used to test for differences over time, while a two-sample T^2-test was used to test for differences between the groups.

3.2 Mixed models for multivariate longitudinal data

As in much of the work outlined above, we intend to employ Procrustes methods to longitudinal facial shape data in order to remove the effects of location, rotation and scale (interest lies here only in shape and not in size) and to obtain approximate tangent coordinates which allow multivariate analyses to be carried out. The change in facial shape over time will be modelled using linear mixed effects models. In the situation where we wish to model measurements $y_{ir}(t)$, from $i = 1, \ldots, n$ individuals, at time t for $r = 1, \ldots, m$ outcomes, we have the options of fitting multiple univariate mixed models for each of the outcomes, or of fitting a joint model to all m outcomes at one time. There are several benefits in joint modelling of multivariate outcomes: the association structure across outcomes may be of importance; we may be interested in the difference in evolution of the individual outcomes; or interest may lie in the joint testing of a treatment effect on a set of outcomes (Fieuws & Verbeke, 2006). This latter point is of particular importance in shape analysis when interest often lies in the treatment effect on the shape as a whole, rather than on individual landmarks. We proceed by describing the framework for univariate and joint mixed models, along with their disadvantages in the shape analysis setting, and continue by introducing an alternative pairwise modelling approach.

3.2.1 Univariate mixed models

If y_i is a vector of responses from a single outcome across time for individual i, the model is assumed to be $y_i|b_i \sim F_i(\psi, b_i)$ for some pre-specified distribution F_i, a vector of unknown parameters ψ, and a vector of random effects, b_i, specific to each individual, where $b_i \sim N_q(0, D)$ for a $(q \times q)$ covariance matrix D (for a random intercept and slope, for example, D would be (2×2)). Conditional independence is often assumed in that the components

of y_i are independent, conditional on the random effects (Fieuws & Verbeke, 2006). The conditional linear mixed model assumes

$$y_i|b_i \sim N(X_i\beta + Z_i b_i, \Sigma_i), \quad i = 1, \ldots, N,$$

where, for n_i the number of observations from individual i,

- X_i is an $(n_i \times k)$ matrix of known covariates for the fixed effects;
- Z_i is an $(n_i \times q)$ matrix of known covariates for the random effects;
- Σ_i is an $(n_i \times n_i)$ covariance matrix. Under conditional independence $\Sigma_i = \sigma^2 I_{n_i}$ and so the parameters in Σ_i do not themselves depend on i.

In general, inference is based on the marginal model

$$f_i(y_i|\psi) = \int f_i(y_i|b_i, \psi) g(b_i)\, db_i,$$

where $f_i(y_i|b_i, \psi)$ is the conditional density of y_i given b_i and a vector of fixed effects, ψ, and $g(b_i)$ is the density of the random effects (Fieuws & Verbeke, 2006). The marginal distribution of the responses is

$$y_i \sim N(X_i\beta, Z_i D Z_i' + \Sigma_i).$$

3.2.2 Joint mixed models

If the m outcomes are to be modelled jointly, we assume that Y_{ir}, the n_{ir}-vector on subject i for outcome r, is distributed as

$$Y_{ir}|b_{ir} \sim f_{ir}(Y_{ir}|b_{ir}, \psi_r), \quad i = 1, \ldots, N, r = 1, \ldots, m,$$

for b_{ir} a q-dimensional vector of random effects and ψ_r a vector of fixed effects for outcome r (Molenberghs & Verbeke, 2005). If θ^* is the vector of all the unknown parameters (regression coefficients and random effects' covariances), then the log likelihood for individual

i is (Fieuws & Verbeke, 2006)

$$l_i(Y_{i1}, \ldots, Y_{im}|\theta^*). \tag{3.1}$$

Assume that, conditional on b_{ir}, $Y_{ir}, \ldots, Y_{im}$ are independent and that

$$b_i = \begin{pmatrix} b_{i1} \\ \vdots \\ b_{im} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} D_{11} & \ldots & D_{1m} \\ \vdots & \ldots & \vdots \\ D_{m1} & \ldots & D_{mm} \end{pmatrix}\right], \tag{3.2}$$

where b_{ir} is a vector of random effects for individual i and outcome r and the covariance matrix in (3.2) can be denoted as D (Molenberghs & Verbeke, 2005). Therefore the component D_{rs} is a matrix of covariances between the vectors b_{ir} and b_{is} if there is more than one random effect per individual and outcome (e.g. a random intercept and slope); otherwise D_{rs} is a scalar covariance between the scalars b_{ir} and b_{is} (e.g. a random intercept only) . It is clear that as the number of outcomes, m, increases, the dimension of the covariance matrix, D, in (3.2) becomes very large and thus difficult or impossible to estimate (Molenberghs & Verbeke, 2005).

One potential solution to this problem is to fit a shared-parameter model, which is a special case of the joint mixed model in which one set of random effects is assumed for all outcomes. This reduces the dimension of the random effects and prevents the dimension changing with the introduction of additional outcomes, but makes strong assumptions about the relationships between outcomes (Molenberghs & Verbeke, 2005). The following section presents a more flexible approach to dealing with the problem.

3.3 Pairwise modelling approach

The pairwise modelling approach, as suggested by Fieuws & Verbeke (2006) and documented in Molenberghs & Verbeke (2005), is aimed at preserving the advantages of the joint modelling approach but avoiding the computational difficulties inherent in estimating the parameters of D when the number of outcomes is large. It involves fitting $m(m-1)/2$ bivariate

models for all possible pairs

$$(Y_1, Y_2), (Y_1, Y_3), \ldots, (Y_1, Y_m), (Y_2, Y_3), \ldots, (Y_2, Y_m), \ldots, (Y_{m-1}, Y_m),$$

where Y_r is a vector of measurements of the r^{th} outcome across individuals and times. For Y_{ir} the subvector consisting of the vector of measurements on outcome r for individual i, the sum of log-likelihoods

$$\sum_{i=1}^{N} l_{irs}(Y_{ir}, Y_{is}|\theta_{r,s}) \tag{3.3}$$

is maximised for each $r = 1, \ldots, m-1$, and $s = r+1, \ldots, m$, where $\theta_{r,s}$ is the vector of all of the parameters in the bivariate model corresponding to the pair of outcomes (r, s).

The sum of log-likelihoods may be rewritten as $\sum_{i=1}^{N} l_{ip}(\theta_p)$ for $p = 1, \ldots, P$, where $P = m(m-1)/2$ represents the total number of possible pairs. Define θ as the stacked vector of all θ_p. Therefore $\hat{\theta}$ is the vector of all parameters estimated from the P bivariate models and $\hat{\theta}^*$, as defined in (3.1), is the vector of all parameters estimated from the joint model. It is important to note that $\theta \neq \theta^*$, an inequality that arises from some elements in θ^* having multiple counterparts in θ, for example the variance of a random effect for a particular outcome r (because it appears in the models containing both pairs (r, s) and (r, s').) It is possible to find a single estimate for each parameter by averaging over all the pair-specific estimates in θ and the inference for these estimates is presented in the next section.

3.3.1 Making inferences about θ

Although the estimates in θ^* may be found by averaging, the standard errors of the averaged estimates may not and, in any case, the pair-specific estimates (for example from pairs (r, s) and (r, s')) are correlated, a factor which must be taken account of. Fitting all of the pairwise models is equivalent to maximising a pseudo-likelihood (Besag, 1975) of the form

$$\begin{aligned} pl(\theta) &= l(Y_1, Y_2|\theta_{1,2}) + l(Y_1, Y_3|\theta_{1,3}) + \ldots + l(Y_{m-1}, Y_m|\theta_{m-1,m}) \\ &= \sum_{p=1}^{P} l_p(\theta_p), \end{aligned}$$

where a p represents a pair of outcomes (r, s) and $p = 1, \ldots, P$. Under the pseudo-likelihood framework, asymptotically,

$$\sqrt{N}(\hat{\theta} - \theta) \sim MVN(0, J^{-1}KJ^{-1}),$$

where J is block-diagonal with diagonal blocks J_{pp} of second derivatives of the log-likelihood for pair p, with respect to the vector of parameters pertaining to that pair:

$$J_{pp} = \frac{1}{N}\sum_{i=1}^{N} E\left[\frac{\partial^2 l_{ip}}{\partial\theta_p \partial\theta_p'}\right]$$

and K is symmetric with blocks K_{pq} of the products of first derivatives of the log-likelihoods for pair p and pair q, with respect to the vectors of parameters pertaining to each pair:

$$K_{pq} = \frac{1}{N}\sum_{i=1}^{N} E\left[\frac{\partial l_{ip}}{\partial\theta_p}\frac{\partial l_{iq}}{\partial\theta_q'}\right],$$

where $p, q = 1, \ldots, P$.

If A is a matrix of appropriate coefficients to calculate the average estimates, then $\hat{\theta}^* = A\hat{\theta}$ and is distributed approximately as

$$\hat{\theta}^* \sim MVN(A\theta, A\Sigma(\hat{\theta})A'), \tag{3.4}$$

for $\Sigma(\theta) = J^{-1}KJ^{-1}$.

3.3.2 Comparison with the fully joint approach

Fieuws & Verbeke (2006) undertook a simulation study using three outcomes to investigate the efficiency of the pairwise approach relative to fitting a fully joint model. They found high intra-class correlations between the estimates from the full and pairwise approaches and similar levels of bias in both. They also found that there was no efficiency loss by the pairwise approach in the scenario when all of the fixed and random effects were outcome-specific, but that there was some small efficiency loss when some effects were common to all 3 outcomes. There was no indication that the efficiency loss increased with the number of outcomes

sharing a parameter. We will corroborate this in the shape setting, in a simulated example in Section 3.5.3. The development of this approach in the context of longitudinal shape data is a novel yet natural one, given its high dimensionality and the benefits of treating the face as a whole, rather than breaking it down into constituent parts or aggregating all of the data into a single score.

3.4 The `lmepair` function

Fieuws & Verbeke (2006) implemented the pairwise method for their hearing data in SAS, but thus far we are unaware of any `R` code that has been written to fit the models. We therefore constructed a function, named `lmepair`, in order to do this. The function is available at `http://biostatistics.oxfordjournals.org/` with the published article containing some of the work of this thesis (Barry & Bowman, 2008).

The `lmepair` function is fairly general and can take any mean function for the fixed effects, specified in the "fixed" argument. This should be entered as a formula in the same way as for `lm` or `lme`, e.g. $y = group + time + group : time$. A dataframe must then be named, that contains the variables in the "fixed" argument and also the information on which response value corresponds to which outcome, using dummy variables for each outcome. The name of the subject identification variable (e.g. "ID") must be included, along with the number of outcomes. Either the maximum likelihood (ML) or the restricted maximum likelihood (REML) methods may be chosen to fit the linear mixed model. Various random effects structures may be selected, including independent or correlated random effects, different random effects variances for two groups (e.g. treatment groups or genders) and a setup where the covariance of the random effects is estimated if the empirical correlation between outcomes is greater than a certain cut-off point, and is assumed zero otherwise. Separate random error variances may also be estimated for different groups (e.g. treatment groups or coordinates). Finally, there is an option for fitting a reduced model, with the particular effect to omit specified.

The function then proceeds by selecting the data for one of the pairwise combinations of outcomes, then fitting the linear mixed effects model using either the `lme` function in the

`nlme` package, or the `lmer2/lmer` functions in the `lme4` package, as specified by the user. Currently, there does not appear to be a facility within `lmer` to fit different random error variances for two groups, so when this is required `lme` is used. Throughout the period of study for this Ph.D., the `lme4` package has been under development. Since it is more flexible with simpler syntax, and considerably faster computationally, it is the preferred package for fitting these models in `R`. However, there are still some facilities that are available using `lme` that cannot yet be programmed into `lmer` or `lmer2`. The `lmer2` function is the preferred function of the three, but it is the newest and will eventually replace `lmer` and be renamed as such. Where the `lmer` function is mentioned here, it currently refers to `lmer2`, with the intention of this being renamed as `lmer` in future. Now that the `lme` function has been updated to `lmer`, the linear mixed model computation can handle around 10 outcomes (or possibly more) with random effects variances for each.

Once the model has been fitted, the estimates of the fixed effects and random effects variances are extracted. The function finds the analytical first and second derivatives of the relevant log-likelihood for that particular pair by passing an expression for the log-likelihood to the function `deriv`, along with the mean function and the names of the variance parameters. Any mean function may be passed for the fixed effects. The derivatives are then evaluated using the parameter estimates obtained from the model and the individual-level responses and covariates.

There is a slight discrepancy between the method of obtaining the parameter estimates and their standard errors. The lme function takes the conditional likelihood and integrates out the random effects. Using decompositions to evaluate the likelihood and thereby obtain expressions for the estimates of the fixed effects and random error variance parameters (which are dependent on the random effects variance parameters), a profiled likelihood is ascertained for the vector of random effects variance parameters. This profiled likelihood is maximised with respect to the vector of random effects variance parameters, and the maximum likelihood estimator is plugged into the expressions for the fixed effects and random error variance parameters. The actual estimates are found by iteration; an initial value is chosen for the vector of random effects variance parameters, before a moderate number of EM iterations (Dempster et al., 1977) are performed to refine the starting estimates. Further iterations

with the Newton-Raphson algorithm (Laird & Ware, 1982) are then carried out to converge to the optimum. These optimisation methods estimate the fixed effects and random error variance parameters using the initial estimate for the random effects variance parameters, then use the new estimates for the fixed effects and random error variance parameters to find a new estimate for the random effects variance parameters, and so on until convergence is reached. Full details of this process are given in Pinheiro & Bates (2000).

The discrepancy thus arises from the way the likelihood is expressed. The `lme` function starts with the conditional likelihood and integrates out the random effects, before re-expressing the likelihood using decompositions; whereas the `lmepair` function expresses the pairwise likelihood for the pseudo-likelihood approach as the marginal likelihood outright. The latter results in the need to invert the variance matrix for the errors under the marginal likelihood, which include the random effects and random error variances. To allow inversion of this matrix, the correlations between time points were assumed to be zero for the purposes of calculating the standard errors of the parameter estimates. The parameter estimates and their standard errors are therefore calculated under slightly different assumptions, and although the sandwich estimator in the expression for the standard errors should correct for misspecification of the variance structure, some discrepancies may remain.

The second derivatives of each pairwise likelihood are summed over the time points for each individual, and subsequently averaged across subjects. The blocks, J_{pp}, of the J matrix are constructed from the averaged second derivatives and this is carried out for each pairwise combination of outcomes. The K matrix is constructed by summing the first derivatives over the time points for each individual, then multiplying the blocks of first derivatives for two pairs p and q together, and averaging across individuals to form the block K_{pq}. This is carried out for all the combinations of pairs and these are then stacked to form the full matrix.

The long vector of parameter estimates containing repetitions, $\hat{\theta}$, is constructed directly from the random effects variance and fixed effects estimates in the linear mixed effects model. The matrix of coefficients, A, corresponding to the repeated estimates of each parameter, is constructed and multiplied by $\hat{\theta}$ to obtain the required vector of parameter estimates, $\hat{\theta}^*$. Finally, the variance matrix, $A\Sigma(\hat{\theta})A'$, of $\hat{\theta}^*$ in (3.4) is evaluated, and the parameter and

variance estimates ordered such that they are appropriately displayed. The function returns $\hat{\theta}^*$, its covariance matrix and a vector of estimated standard errors for each parameter, the long vector of repeated parameter estimates, $\hat{\theta}$, the log-likelihood for each pairwise model and (if required) the J and $\Sigma(\theta)$ matrices for use in the pseudo-likelihood ratio test.

3.5 Application to simulated triangle and quadrilateral data

3.5.1 Introduction

Longitudinal shape data were simulated so that methods could be applied to these simple data before using them on more complex data. Initially coordinates were simulated for two-dimensional triangles over time in order to resemble the Bookstein method of fixing two points (Dryden & Mardia, 1998) and a linear mixed effects model was fitted to the two coordinates of the remaining point.

Coordinates were then simulated for quadrilaterals over time, again using the Bookstein method. This left four coordinates; two for each of the remaining points. A linear mixed effects model was fitted to the set of four outcomes for each subject at each time point and this was compared to the pairwise modelling approach which was used to analyse each bivariate combination of outcomes before combining the results. Different random effects variance structures were considered on both sets of data.

3.5.2 Triangles

Sets of coordinates for $n = 200$ triangles were simulated at each of three time points, a sample of which are illustrated in Figure 3.1. The coordinates of points 1 and 2 were fixed at (-0.5,0) and (0.5,0) respectively, according to the Bookstein coordinate method. The coordinates of point 3 were allowed to vary and were simulated as random Normal variables with mean 1 and standard deviation 0.3. Subject-specific variation was added in the form of a random intercept, which allowed each of the two coordinates at each time point to shift by a certain small amount for each subject (standard deviation 0.6 for the x-coordinate and

0.25 for the y-coordinate). The random intercepts were allowed to be different for the $x-$ and $y-$coordinates but had a correlation of 0.7 within a subject and were constant over time.

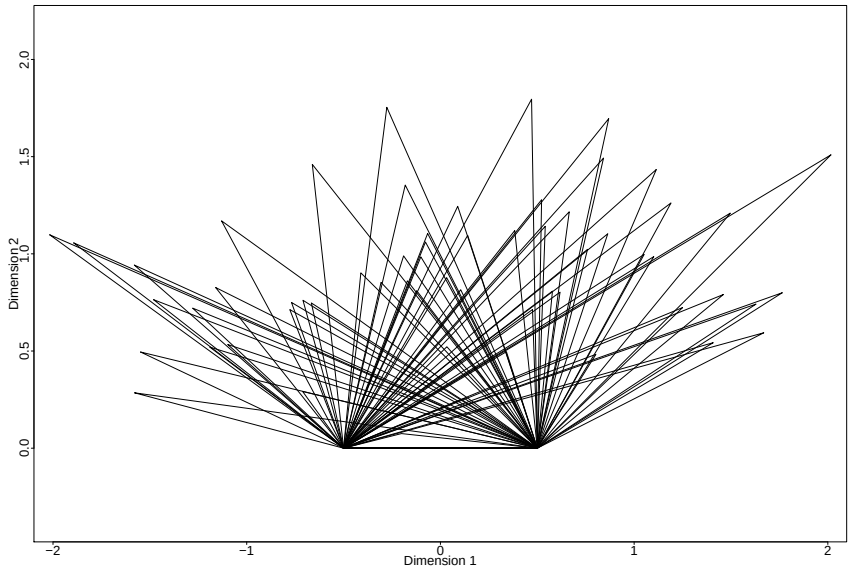

Figure 3.1: Sample of 20 individuals' simulated triangle data

Each subject had a 6-vector of responses, Y_i, consisting of two coordinates at each of three time points. The model for the value of a coordinate r measured on individual i was as follows:

$$y_{ir}(t) = \beta_{0r} + b_{ir} + \beta_{1r}t + \epsilon_{ir}(t) \tag{3.5}$$

for $i = 1, \ldots, n$ subjects, continuous time t, $r = 1, 2$ different coordinates ($r = 1$ corresponds to x and $r = 2$ to y) and fixed slopes β_{11} and β_{12}. The random error, $\epsilon_{ir}(t)$, is distributed as $N(0, \sigma^2)$. Each individual is assumed to have their own random intercept for each coordinate, b_{i1} and b_{i2} respectively, where

$$b_i = \begin{pmatrix} b_{i1} \\ b_{i2} \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.360 & 0.105 \\ 0.105 & 0.0625 \end{pmatrix}\right) \tag{3.6}$$

Data were simulated for 200 individuals from the conditional model $y_{ir}|b_{ir} \sim N(X_{ir}\beta + b_{ir}, \sigma^2)$ for $\beta = (\beta_{01}, \beta_{02}, \beta_{11}, \beta_{12})$, following simulation of the random effects from the distribution in (3.6). The linear mixed effects model in (3.5) was fitted to the data using the `lmer` command

in the `lme4` package (Bates, 2007) of the `R` programming language (R Development Core Team, 2007). The results of fitting the model are displayed in Table 3.1.

Parameter	Bias	Estimate	Std. error
β_{01}	-0.043	-1.043	0.047
β_{02}	0.018	1.018	0.025
β_{11}	0.001	1.001	0.015
β_{12}	-0.025	-0.025	0.015
σ_{1b}	0.003	0.603	-
σ_{2b}	-0.017	0.233	-
ρ_{12b}	-0.054	0.646	-
σ	-0.004	0.296	-

Table 3.1: Biases, with estimates and their standard errors for the fixed effects parameters in the random effects model, along with the estimate and bias for the standard error of the random effects for coordinates 1 (σ_{1b}) and 2 (σ_{2b}), the correlation between them (ρ_{12b}) and the random error SE

The biases are all reasonably small, with the exception perhaps of the coordinate 1 intercept. The correlation between the random effects is also reasonably well estimated. When 500 datasets (of 200 individuals) were simulated from the conditional model described in (3.5), the range of estimated correlations was from 0.57 to 0.81, with a mean of 0.70. This implies that on average the correlation between the random effects is very well estimated, and as the sample size in each dataset increases, the range becomes narrower. For example, for a smaller sample size of 30, the range of correlations was from 0.35 to 0.89. Histograms for the parameter distribution are displayed for the correlation along with the other random effects variance and fixed effects parameters in Figure 3.2. All show reasonably normal distributions centred approximately on their true means.

3.5.3 Quadrilaterals

Data

Sets of coordinates for $n = 200$ quadrilaterals were simulated at each of three time points. Each of the 200 individuals were allocated to one of two groups and these groups had different intercepts and mean trends. The simulated data for a sample of ten individuals at each of the three time points is displayed in Figure 3.3 as illustration.

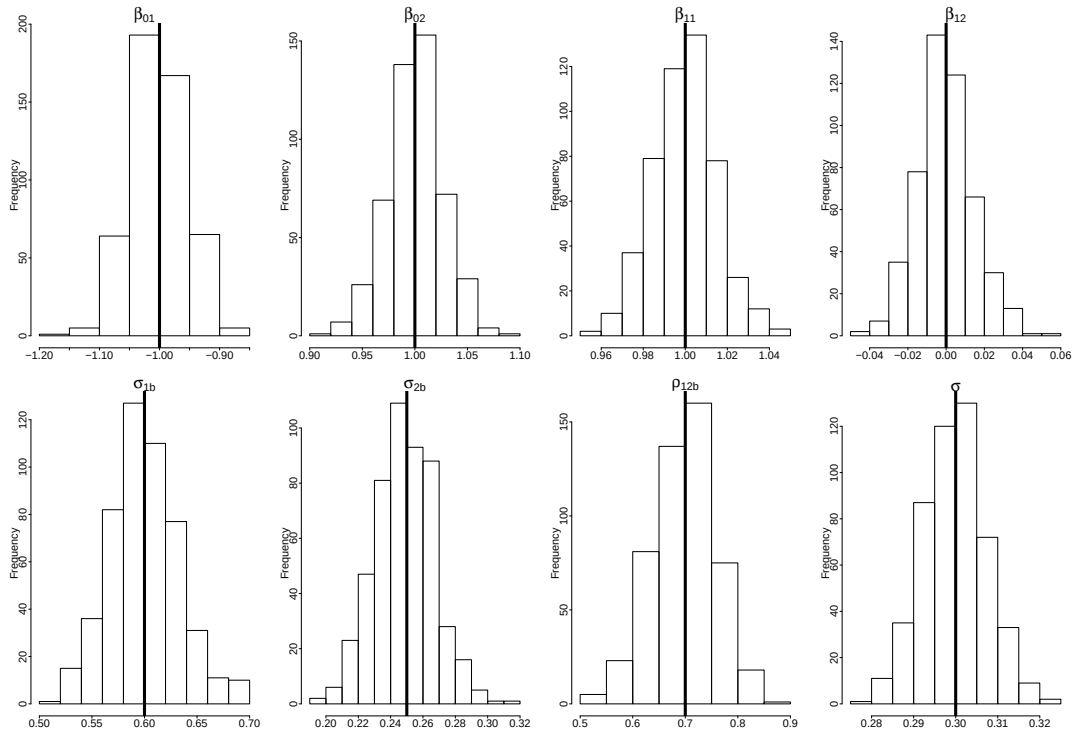

Figure 3.2: Distributions for the estimated parameters obtained from the linear mixed model fits to 500 simulated datasets of 200 individuals each. The true values for each of the parameters are marked by thick black lines.

The x- and y-coordinates of points 1 and 2 were again fixed at (-0.5,0) and (0.5,0) respectively, according to the Bookstein coordinate method. The two-dimensional coordinates of the two remaining points were allowed to vary and were simulated from the following model:

$$y_{ir}(t) = \beta_{0r} + b_{ir} + \beta_{1r}t + \beta_{2r}g_i + \beta_{3r}g_i \cdot t + \epsilon_{ir}(t), \tag{3.7}$$

for $y_{ir}(t)$ the value of coordinate r for individual i at time t, as in the triangles example. The coordinates were ordered such that 1 and 2 correspond to the x and y coordinates of point 1 (upper left point of the quadrilateral) and 3 and 4 correspond to the x and y coordinates of point 2 (upper right point). The random error was simulated as $\epsilon_{ir}(t) \sim N(0, 0.04)$. The true values of the fixed effects parameters were, for $\beta_p = (\beta_{p1}, \ldots, \beta_{p4})$, $(p = 0, \ldots, 3)$, the vector containing the parameters corresponding to the relevant fixed effect for each of the

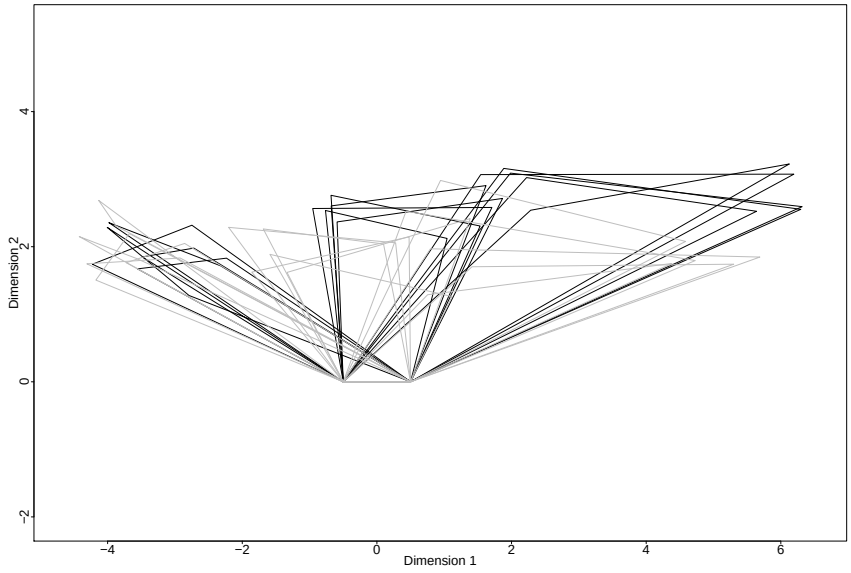

Figure 3.3: Sample of ten individuals' simulated quadrilateral data, five from group 0 (grey) and five from group 1 (black)

coordinates $1, \ldots, 4$:

$$\beta_0 = (-4, 2, -3, 2) \quad \beta_1 = (3, 0, 3.5, 0) \quad \beta_{2r} = \beta_2 = 0.05 \quad \beta_{3r} = \beta_3 = 0.575$$

Subject-specific variation was added in the form of a random intercept, the variance of which was coordinate specific and distributed, for individual i, as

$$b_i = \begin{pmatrix} b_{i1} \\ b_{i2} \\ b_{i3} \\ b_{i4} \end{pmatrix} \sim MVN \left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.040 & 0.070 & 0.025 & 0.006 \\ 0.070 & 0.123 & 0.012 & 0.063 \\ 0.025 & 0.012 & 0.123 & 0.011 \\ 0.006 & 0.063 & 0.011 & 0.090 \end{pmatrix} \right),$$

where $\text{var}(b_{ir}) = \sigma^2_{rb}$ and $\text{cov}(b_{ir}, b_{is}) = \sigma_{rs}$ for coordinates r and s. The covariance matrix

above is equivalent to the correlation matrix

$$\begin{array}{c} \\ x_1 \\ y_1 \\ x_2 \\ y_2 \end{array} \begin{array}{c} \begin{array}{cccc} x_1 & y_1 & x_2 & y_2 \end{array} \\ \begin{pmatrix} 1 & 0.1 & 0.7 & 0.1 \\ 0.1 & 1 & 0.1 & 0.6 \\ 0.7 & 0.1 & 1 & 0.3 \\ 0.1 & 0.6 & 0.3 & 1 \end{pmatrix} \end{array} \tag{3.8}$$

Comparison of fully joint and pairwise models

The model in (3.7) is assumed both in the fitting of the fully joint mixed effects model and the pairwise approach. The difference is that under the former, the index $r = 1, \ldots, 4$ represents all four coordinates, whereas under the pairwise approach, within each bivariate model r takes only two of the four values. It is important that the model is parameterised such that the mean value for each coordinate is calculated (subject to various covariate effects) rather than calculating differences between coordinates, because of the averaging of repeated estimates. Therefore there must be individual coordinate intercepts, onto which covariate effects may be added.

The parameter estimates along with their biases from the true values, and their standard errors, are displayed in Table 3.2. This shows that the fixed effects estimates are the same under the fully joint model and the pairwise approach. Since the models are parameterised such that the mean position for each coordinate is estimated (for various combinations of covariates), as opposed to differences between coordinates, one would expect the fixed effects parameter estimates to be the same across models. Therefore, under the pairwise approach, all repeated parameter estimates are equal and their average is identical to the equivalent estimate from the fully joint model. The standard errors of the fixed effects are not identical but are very similar for all parameter estimates and are not consistently larger for either model.

The biases in the fixed effects estimates are very small and are positive and negative in roughly equal proportions. Likewise the biases for the variance parameters under both models are also extremely small; however, these appear more likely to be positive than

	Fully joint model		Pairwise approach	
	Bias	Estimate (SE)	Bias	Estimate (SE)
x_1 - β_0	−0.027	-4.027 (0.027)	−0.027	-4.027 (0.024)
β_1	0.009	2.509 (0.014)	0.009	2.509 (0.014)
β_2	0.045	0.095 (0.038)	0.045	0.095 (0.038)
β_3	−0.030	0.420 (0.020)	−0.030	0.420 (0.020)
y_1 - β_0	0.010	2.010 (0.039)	0.010	2.010 (0.040)
β_1	−0.013	-0.013 (0.014)	−0.013	-0.013 (0.015)
β_2	0.039	0.089 (0.057)	0.039	0.089 (0.055)
β_3	0.037	0.487 (0.020)	0.037	0.487 (0.022)
x_2 - β_0	−0.028	-3.028 (0.039)	−0.028	-3.028 (0.039)
β_1	0.022	4.022 (0.014)	0.022	4.022 (0.015)
β_2	0.069	0.119 (0.056)	0.069	0.119 (0.056)
β_3	−0.046	0.404 (0.020)	−0.046	0.404 (0.021)
y_2 - β_0	0.017	2.017 (0.035)	0.017	2.017 (0.034)
β_1	−0.010	-0.010 (0.014)	−0.010	-0.010 (0.015)
β_2	−0.016	0.034 (0.050)	−0.016	0.034 (0.050)
β_3	0.044	0.494 (0.020)	0.044	0.494 (0.020)
σ_{1b}	−0.002	0.198 (−)	−0.003	0.197 (0.124)
σ_{2b}	0.003	0.353 (−)	0.003	0.353 (0.049)
σ_{3b}	0.001	0.351 (−)	0.001	0.351 (0.019)
σ_{4b}	0.003	0.303 (−)	0.004	0.304 (0.018)
σ_{12b}	0.002	0.009 (−)	0.002	0.009 (0.006)
σ_{13b}	−0.001	0.048 (−)	−0.001	0.048 (0.008)
σ_{14b}	0.003	0.009 (−)	0.003	0.009 (0.005)
σ_{23b}	0.004	0.016 (−)	0.004	0.016 (0.010)
σ_{24b}	0.003	0.066 (−)	0.003	0.066 (0.010)
σ_{34b}	0.001	0.012 (−)	0.001	0.012 (0.009)
σ	0.004	0.204 (−)	0.004	0.204 (0.056)

Table 3.2: Biases along with parameter estimates and standard errors for the fully joint and pairwise models fitted to the quadrilateral data

negative under both models. There is little difference in the amount of bias in the estimates provided by either model, for the random or the fixed effects parameters.

Figure 3.4 gives histograms for the distributions of parameter estimates obtained from simulating 500 sets of data on 200 individuals each, from the same model as discussed above, and fitting both the fully joint and pairwise models to each. The histograms of parameter estimates from the fully joint model are filled in with diagonal lines while those from the pairwise approach are shaded, so one can identify points at which they are different. All of the histograms show extremely similar results from both models and are centred roughly on

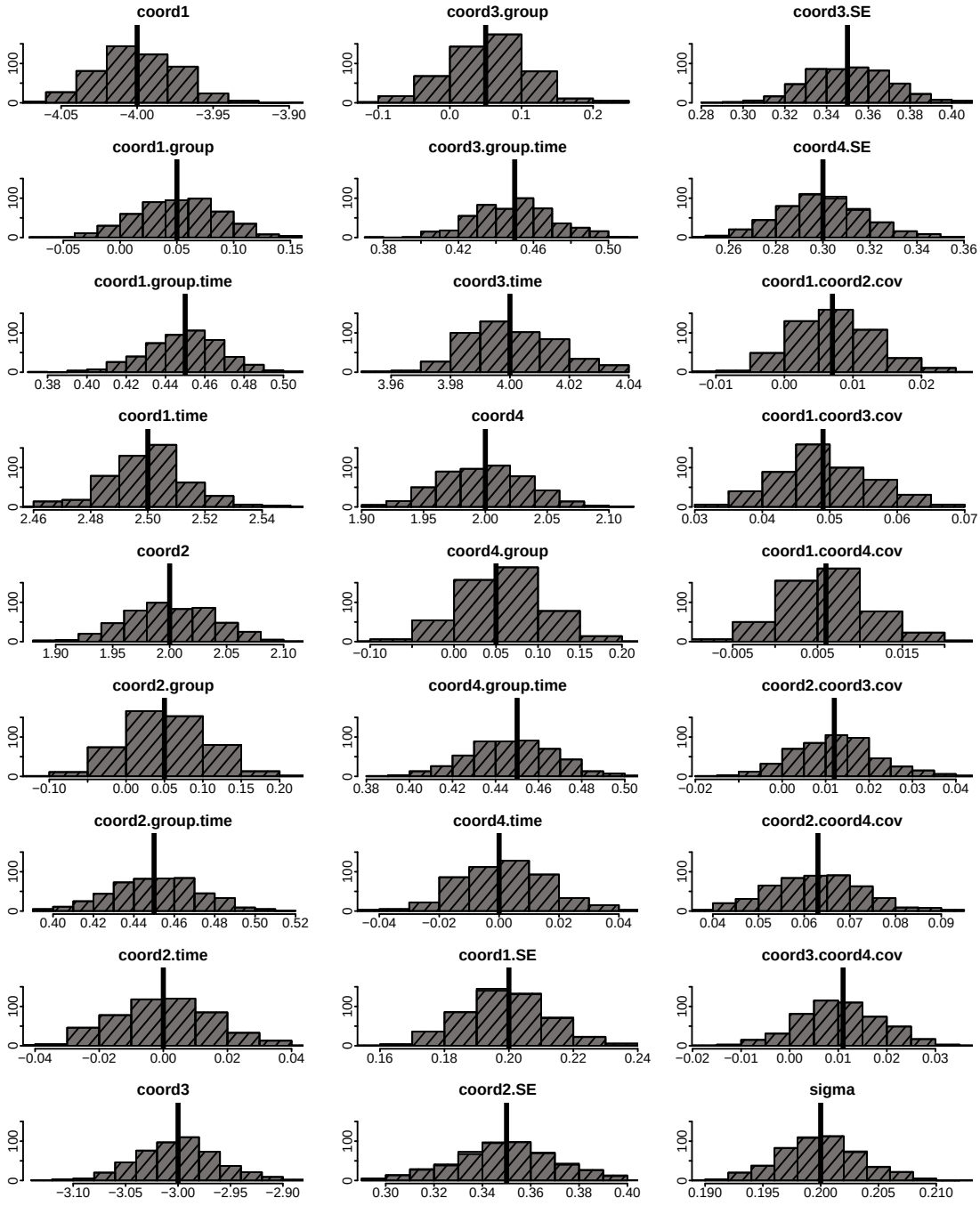

Figure 3.4: Histograms of the parameter estimates obtained by fitting the fully joint (diagonal lines) and pairwise models (dark grey shading) to 500 simulated datasets. The true values for each of the parameters are marked by thick black lines.

the true value for each parameter.

A misspecified covariance structure

If we simulate data from exactly the same model as above, but fit both the fully joint and the pairwise models assuming independent random effects (i.e. misspecifying the covariance structure), the approaches are still similarly robust, as illustrated in Figure 3.5. Both models, on average, correctly estimate the fixed effects and random effects standard errors, with little difference in variance to the previous model, where the covariance structure is correct. The random error variance is also estimated well under both approaches. It is reassuring that the pairwise approach appears to be reasonably robust to model misspecifications and that, despite the correlations being ignored, the remaining parameters are well estimated, including the random effects standard errors. This is likely to be assisted by the sandwich-type estimator of the covariance matrix for the fixed and random effects parameters, which should adjust for a misspecified covariance structure.

3.6 Discussion

Based on the simulations in this chapter, it appears that the pairwise approach gives results which mimic very closely those obtained by fitting a fully joint model. The fixed effects under the pairwise approach are estimated exactly as if the fully joint model was fitted, and their standard errors are very similar. The estimated SEs of the random effects are also close under the fully joint and pairwise approaches, as are the covariances between random effects. Both models show little bias in their parameter estimates compared to the true values and neither approach appears to give consistently higher bias. One advantage that the pairwise approach has over the fully joint model is that approximate standard errors are available for the random effects parameters without any extra effort. Both approaches show robustness when the covariance structure is misspecified, estimating the random effects variances correctly.

One potential issue that has not been addressed here is whether the pairwise approach performs as well as the fully joint model when there are larger numbers of outcomes. This,

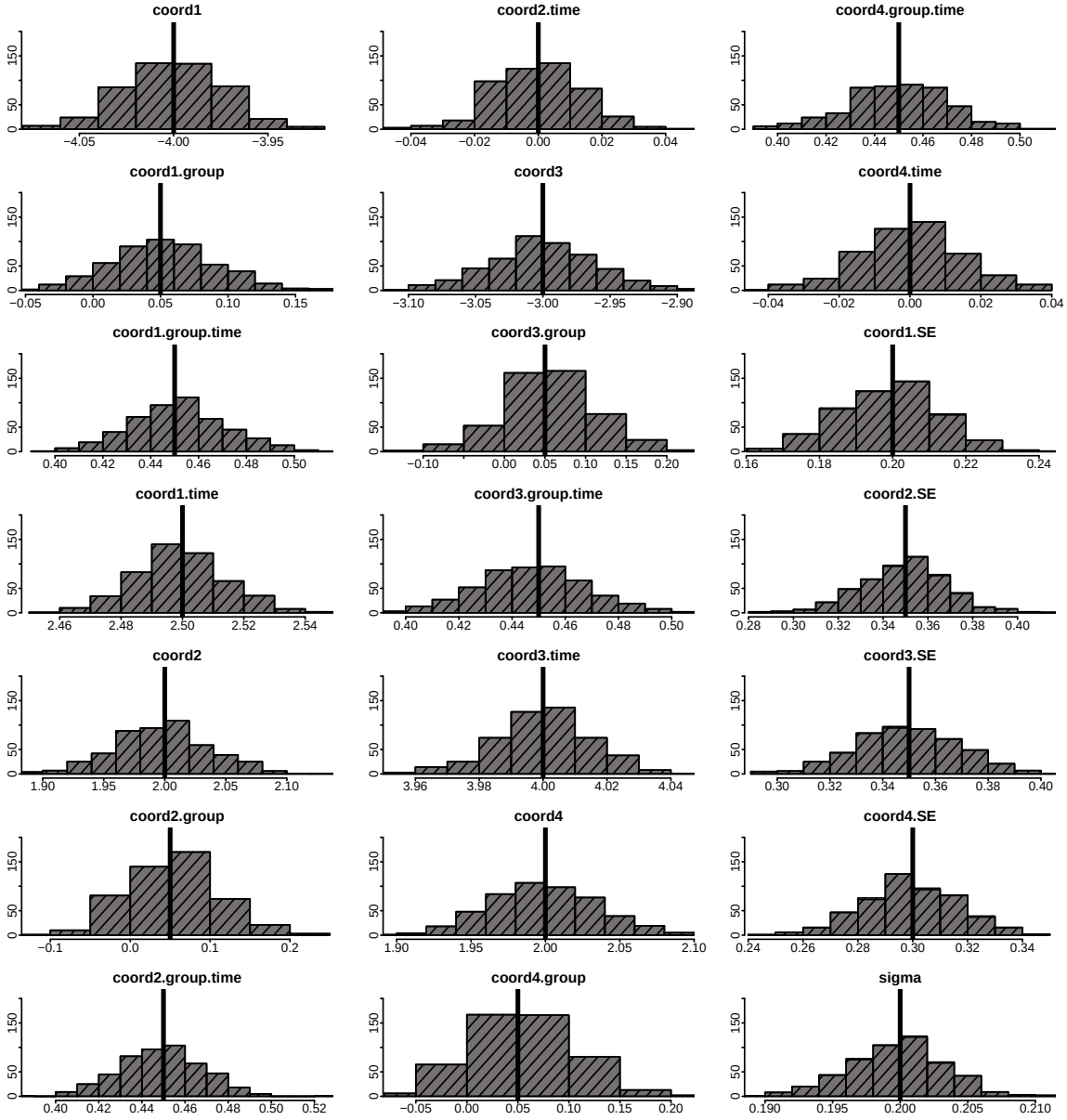

Figure 3.5: Histograms of the parameter estimates obtained by fitting the fully joint (diagonal lines) and pairwise models (dark grey shading) with misspecified variance structures to 500 simulated datasets. The true values for each of the parameters are marked by thick black lines.

however, is not an easily answered question, since the fully joint model cannot be fitted to large numbers of outcomes. There seems no reason why the pairwise approach would not continue to perform well for more outcomes, since it carries out the same calculations, just more of them, and the individual models do not become any more complex.

Chapter 4

Initial application of the pairwise approach to shape data

4.1 Landmarks

The cleft-lip and palate study was introduced in Chapter 1 and in this chapter the landmark data obtained from the study will be analysed using the pairwise approach of Fieuws & Verbeke (2006). As previously mentioned, the main aims of the analysis are to highlight areas of the face in which there are systematic differences that persist as the cleft children grow following their surgery.

Figure 4.1 displays the full set of the landmarks that are marked on each image. The uppermost five points mark out the corners of the eyes and the midpoint between them, whilst the middle collection of landmarks identify the rim and base of the nose and the nostrils. The points E, e, G, g and 2 represent the corners and "Cupid's bow" of the upper lip, and the four lowest midline points are placed on the midline of the lower lip and the chin.

4.2 Reflecting the clefts

The subjects in the cleft group all suffered from unilateral clefts (with or without cleft palate), but the cleft could be on either side in any particular patient. This means that if we are considering asymmetry in the frontal view of the face, on average the left-side clefts are likely to cancel out the right sides. Therefore we reflected images for those with clefts on the right side, resulting in a dataset with images of children with only left-side clefts.

This involves relabelling all the paired landmarks, so in Figure 4.1 those marked with capital letters (say, "E", "M", "H") would be relabelled with their equivalent lower case letters (e.g. "e", "m", "h") and vice versa. The x-coordinates of each paired landmark are then multiplied by -1 and the newly reflected landmarks are used for the rest of the analyses.

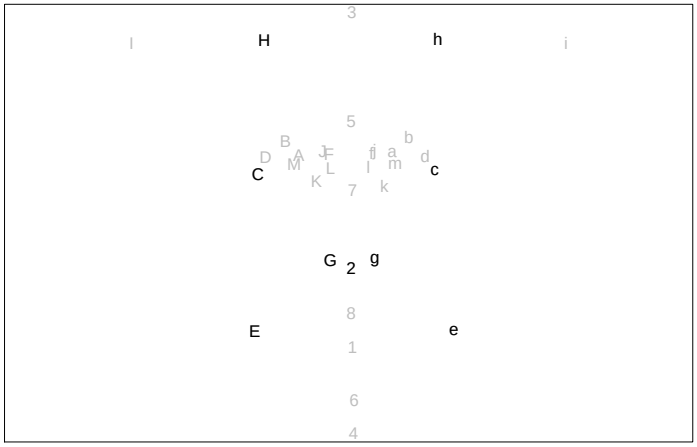

Figure 4.1: Landmarks used to describe facial shape, with lower case letters marking left-side landmarks and upper case marking the equivalent landmarks on the right side of the face. The numbered landmarks are placed along the midline of the face and the black points are those included in the following analysis, whereas the grey points are used only for the GPA.

4.3 Generalised Procrustes analysis and tangent coordinates

Full generalised Procrustes analysis (GPA), as outlined in Chapter 2, was carried out on the images using `procGPA` in `R`. Three-dimensional configurations of landmarks describe the individuals' faces at each of the four time points. The landmarks are labelled consistently

from one image to another and the GPA matches up corresponding landmarks across individuals, minimising the sum of squared differences between all of the images. This results in all effects of rotation, location and scale being removed and any differences remaining between the individuals should be entirely due to discrepancies in shape.

For the initial analyses included in this chapter, we use the selection of landmarks marked in black in Figure 4.1. However in carrying out the GPA, we require as much information as possible on each image in order to remove the unwanted effects of rotation, location and scale, but to retain all of the information about shape. Therefore we used all of the landmarks in Figure 4.1 in the GPA, and extracted the information on the subset of landmarks for inclusion in the examples below. This was especially relevant when only considering midline landmarks (numbered in Figure 4.1), since if the GPA was performed only on the unpaired landmarks, any effect of asymmetry was lost as the process minimised the differences between all images. Each set of cleft landmarks would be rotated so that, in the x-direction, it was in line with the control data. Using the other landmarks effectively anchored the midline of any image according to the relative position of the other landmarks. The GPA was carried out on the shapes from all time points simultaneously.

The landmark data are broken down into individual coordinates, and each coordinate treated as a separate outcome in the analyses. Since we wish to carry out multivariate analyses on these data, the most straightforward approach is to use approximate tangent coordinates instead of the original points, which reside in a high-dimensional shape space, as outlined in Chapter 2. We approximate the tangent coordinates by Procrustes residuals, defined in (2.2).

Occasionally problems arose with the convergence of the `lme` or `lmer` functions within the `lmepair` function in `R`. These problems may be less likely to occur when the magnitude of the outcomes is ranged around unity. The tangent coordinates resulting from the current dataset, in absolute value ranged from about 4.8×10^{-8} to $1.1\times10-2$. We therefore multiplied all tangent coordinates by 10000 in an attempt to avoid convergence problems, transforming back at the end to present the results.

4.4 Facial landmarks

The landmarks selected for the analyses describe the area of the face around the mouth and nose up to the corners of the eyes. Nine points were chosen, being on Figure 4.1 the paired points marked in black as "E", "G", "C", "H" (in lower and upper case) and the midline point "2". Three individuals had data removed (one individual at two time points and two individuals at one time point) due to paired landmarks mistakenly having the same sign for the x-coordinates (i.e. supposedly being on the same side of the face), and to one misplaced z-coordinate. The remaining individuals' Procrustes aligned landmarks are displayed in Figure 4.2 in the frontal view, with the cleft cases in the upper row and the controls in the lower. It is clear that there is more variation between individuals in the cleft group than amongst the controls, especially at three months, and particularly in light of the fact that the control group is approximately double the size. There are particular differences between the groups in the landmarks around the upper lip, where there is some overlap between the points in the cleft group but more distinct separation amongst the controls.

Following Procrustes alignment of the landmarks, the tangent coordinates were approximated by Procrustes residuals, and the trends over time of these are plotted for each coordinate in Figure 4.3. The plots are arranged such that the first row shows the x-coordinates, the second the y and the third the z-coordinates, and each column represents the three coordinates for a particular landmark, with the identifiers above the columns. Within dimension they were ordered as going (loosely) from lowest to highest in the y-direction, and then from left to right in the x-direction from a frontal view.

The plots in Figure 4.3 show the individual trends over time of each of the coordinates, whilst Figure 4.4 displays the mean trends with 95% confidence intervals for each of the two groups. The plots show that, for the majority of the coordinates, there is a jump in the mean cleft group trend between 3 and 6 months, but there is no such effect in the control group. This jump corresponds to the effect of surgery in the cleft group, and therefore must be accounted for in the model. Beyond the 6 month time point, however, both groups appear to follow a reasonably linear trend up to 24 months for most coordinates. This implies that a linear effect of time should suffice from 6 to 24 months for the cleft group and across the

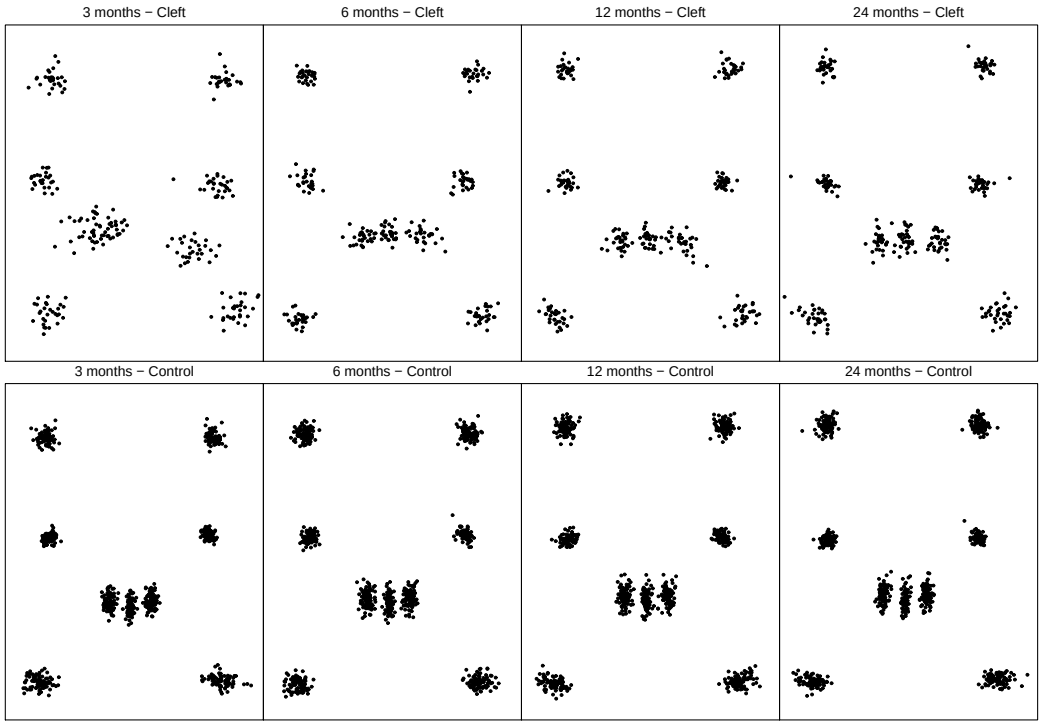

Figure 4.2: Procrustes aligned individual landmarks in the frontal view at 3, 6, 12 and 24 months, with cleft subjects included on the top row and controls on the bottom

entire time period for the control group.

An issue that arises from the plots in Figure 4.4 is that there appears to be differing extents of variation between individuals in each of the two groups for some coordinates and this is corroborated by Figure 4.5. In general there is larger variation amongst subjects in the cleft group, which would be expected due to the varying severity of cleft-lip and palate, although this is not the case for all coordinates and there are larger differences for some than others. It may be necessary to account for this effect in the model, and this will be addressed when comparing different model covariance structures. There are also some slight differences in the extent of variation over time (for the cleft group, at least); but since these are slight and restricted to a small selection of coordinates, no initial allowances will be made in this respect, although it will be discussed at the end of the chapter.

One aspect of interest in fitting the pairwise models is whether different covariance structures for the random effects will actually make much difference to the final results. If not, then in

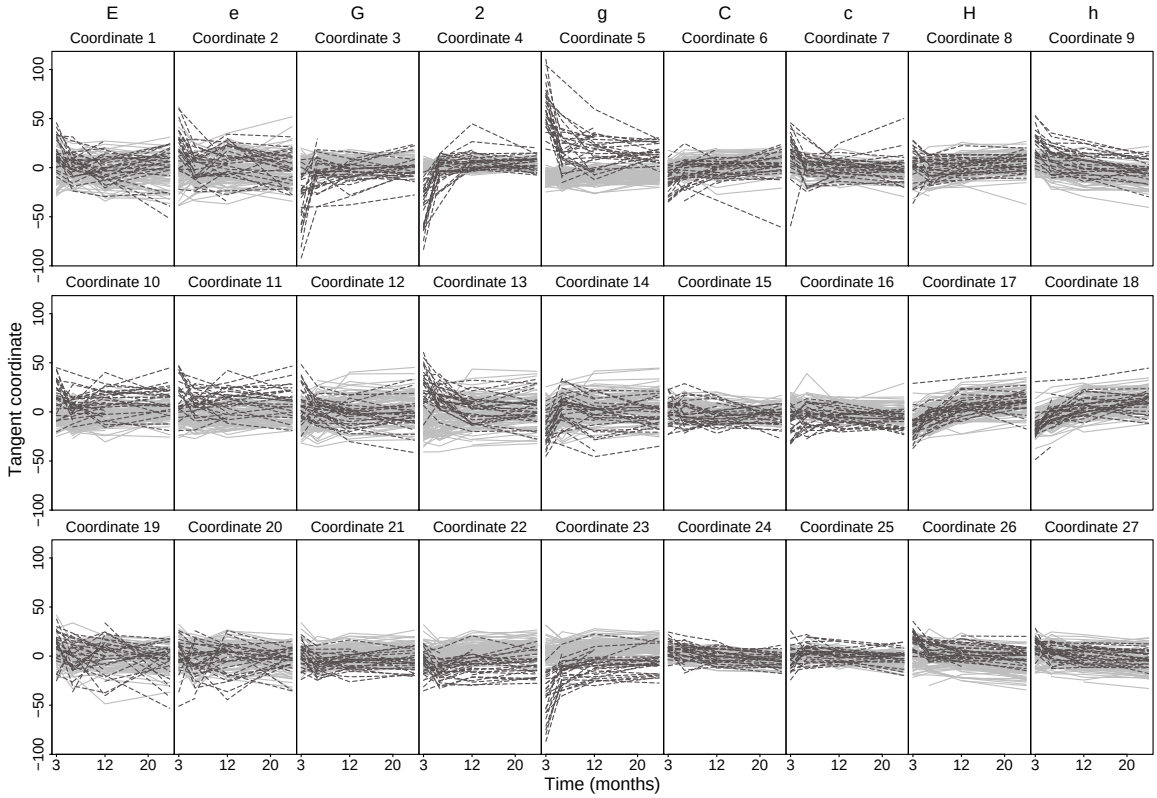

Figure 4.3: Traces over time of the tangent coordinates for each individual, stratified into clefts (dashed black lines) and controls (solid grey lines). The rows represent the x, y and z dimensions and the columns represent the landmarks (with the symbols from Figure 4.1 for identification)

the interests of parsimony, it may be most efficient to fit models with simple structures. This chapter proceeds by considering various model covariance structures and their outcomes.

4.4.1 Independent random effects

We define $y_{ir}(t)$ as the value of a single tangent coordinate r for individual i at time t months of age, where $r = 1, \ldots, km$ and $i = 1, \ldots, n$, where k and m are the numbers of landmarks and dimensions, respectively. Time is treated as continuous. The model for a single tangent coordinate is defined as follows:

$$y_{ir}(t) = \beta_{0r} + b_{ir} + \beta_{1r} g_i + \beta_{2r} t + \beta_{3r} g_i \cdot t + \beta_{4r} g_i \cdot S(t) + \epsilon_{ir}(t), \tag{4.1}$$

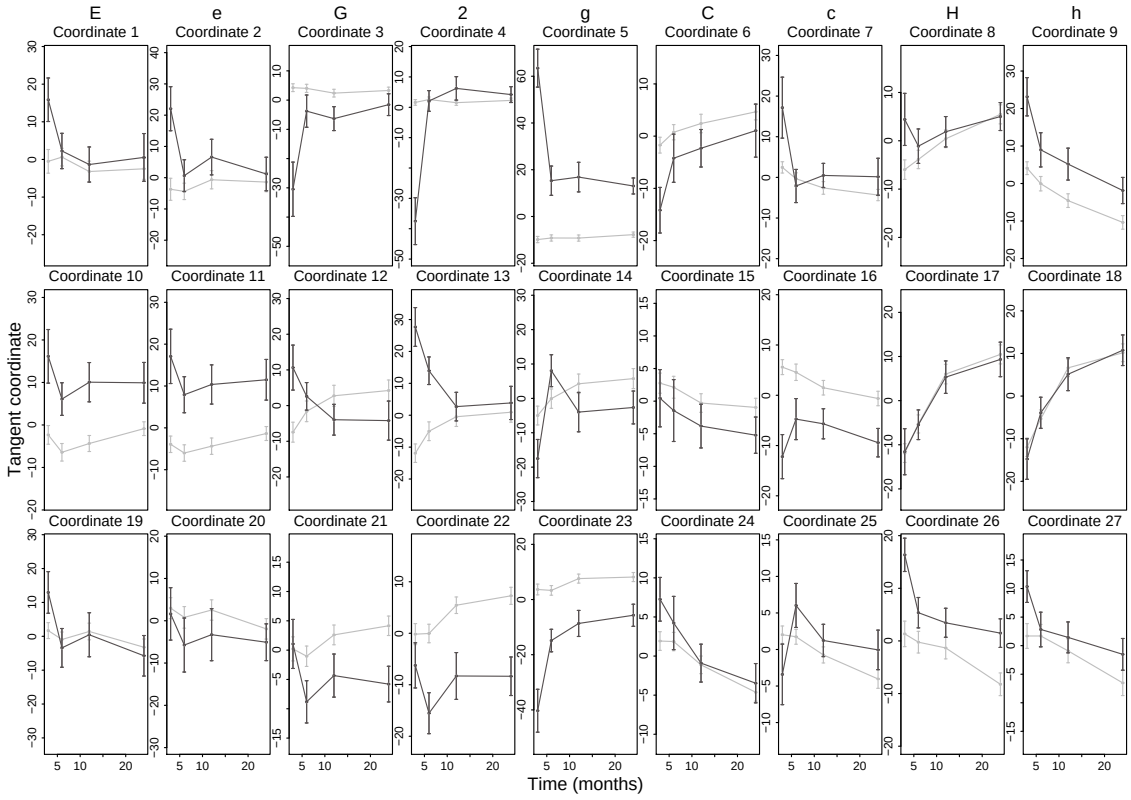

Figure 4.4: Mean trends over time of the tangent coordinates for the clefts (black line) and controls (grey line), with 95% confidence intervals.

where $S(t)$ is an indicator variable taking value 0 for $t < 6$ months and 1 otherwise to account for the effects of surgery. The effect of time is assumed to be linear for the controls and from 6 months onwards for the clefts and Figure 4.3 suggests that this is appropriate, as previously discussed. The covariate g_i is an indicator taking value 1 if individual i is in the cleft group and 0 otherwise and the time by group interaction allows a different slope between 6 and 24 months for the two groups. The random error term, $\epsilon_{ir}(t)$, is distributed as $N(0, \sigma^2)$.

The random intercepts, b_{ir}, follow a $N(0, \tau_r^2)$ distribution. Therefore each individual has a random intercept for each coordinate, but this is constant across time. Under the independent random effects model, the random intercepts are uncorrelated, which is the simplest covariance structure for the random effects whilst still allowing them to be coordinate-specific. This is also a rather unrealistic covariance structure, but it may be enough just to account

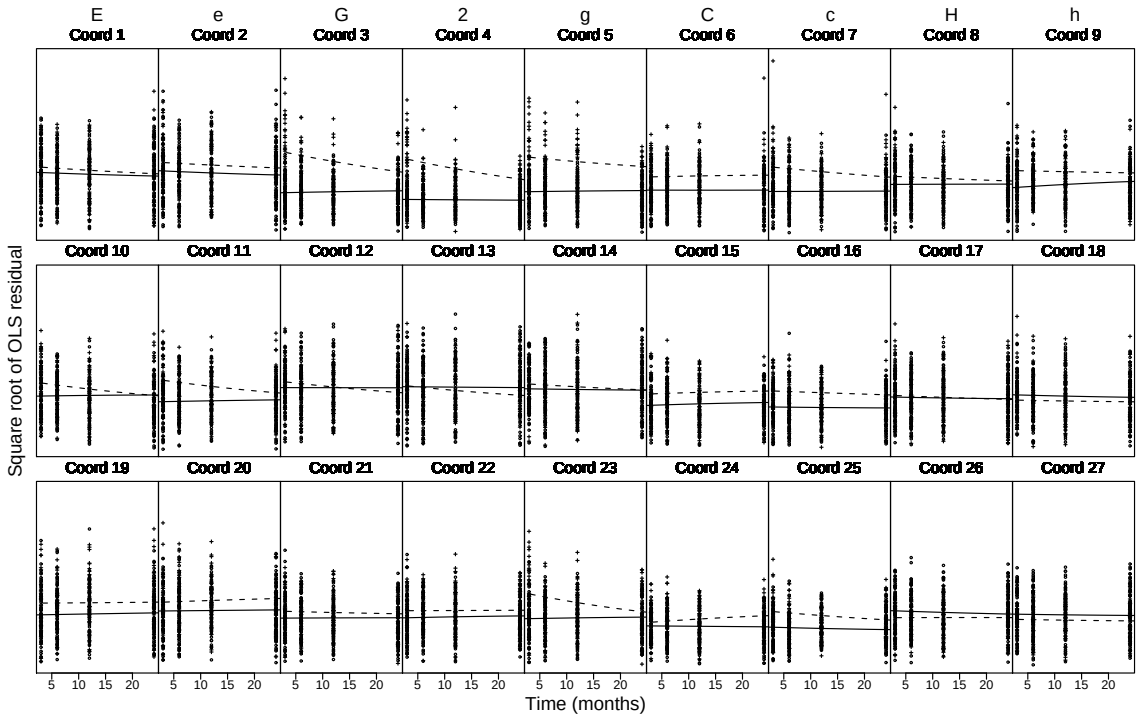

Figure 4.5: Square root of the OLS residuals for each coordinate over time with the means for the clefts (dashed line) and controls (solid line), after fitting a linear model to the data and ignoring random effects

for the random effects, even if their covariance structure is over-simplified.

For each pairwise combination of tangent coordinates $p = (r, s)$ then, the bivariate model is specified as

$$\mathbf{y}_{ip}(t) = \boldsymbol{\beta}_{0p} + \mathbf{b}_{ip} + \boldsymbol{\beta}_{1p} g_i + \boldsymbol{\beta}_{2p} t + \boldsymbol{\beta}_{3p} g_i \cdot t + \boldsymbol{\beta}_{4p} g_i \cdot S(t) + \boldsymbol{\epsilon}_{ip}(t), \tag{4.2}$$

where all bold vectors $\boldsymbol{\beta}_{jp} = (\beta_{jr}, \beta_{js})$, (where $j = 1, \ldots, 4$, representing the four fixed effects in this model), are of length two, the first entry corresponding to tangent coordinate r and the second to coordinate s. The two entries, $\epsilon_{ir}(t)$ and $\epsilon_{is}(t)$, of the random error vector, $\boldsymbol{\epsilon}_{ip}(t)$ are both distributed as $N(\mathbf{0}, \sigma_p^2)$. The vector of random intercepts, $\mathbf{b}_{ip} = (b_{ir}, b_{is})$ is distributed as $N(\mathbf{0}, D)$, where D is the 2×2 diagonal matrix

$$\begin{pmatrix} \tau_r^2 & 0 \\ 0 & \tau_s^2 \end{pmatrix}. \tag{4.3}$$

These bivariate models are fitted for all pairwise combinations of tangent coordinates $p = 1, \ldots, P$. The results are aggregated using the `lmepair` function and the parameter estimates, θ^*, and corresponding standard errors calculated. These are still on the tangent coordinate scale so they must be transformed back into shape space for the results to be interpretable. Since the tangent coordinates were initially approximated by the Procrustes residuals, it is necessary only to add the Procrustes mean shape back onto the mean tangent coordinate estimates in order to obtain means and standard errors for the shapes in their original space. We therefore produce a mean shape at any particular time point by $X\beta^* + \mu$, where μ is the vector (of length 27 in this case) containing the coordinates, stacked by dimension, of the mean shape across all individuals, β^* is the subset of parameters of θ^* containing only fixed effects and X is a design matrix corresponding to the particular group and time point of interest.

The covariance matrix for the mean tangent coordinates is estimated as $\hat{\Sigma} = XV(\hat{\beta}^*)X'$, where $V(\hat{\beta}^*)$ is the submatrix of $\text{var}(\hat{\theta}^*)$ corresponding only to the fixed effects and X is the design matrix for the particular group and time point of interest. For each landmark, the mean subvectors and covariance submatrices are extracted for each combination of x and y or x and z coordinates and these are used to produce bivariate confidence intervals for each landmark using the identity $(u - \mu)'\Sigma_{rs}^{-1}(u - \mu) \sim \chi_2^2$, where the random 2-vector u is distributed as $N(\mu, \Sigma_{rs})$ (Mardia et al., 1979). In this context Σ_{rs} is the (2×2) submatrix of Σ that corresponds to the coordinates of interest and μ is the corresponding mean. The boundaries of the confidence regions are therefore constructed by drawing the appropriate contour of a bivariate chi-squared density function with mean given by the estimate of interest and with covariance matrix Σ_{rs}. These confidence intervals are displayed in both the frontal and profile views for the two groups at each time point in Figure 4.6.

The figure shows clearly the differences between the groups and the trends over time, indicating that the landmarks around the upper lip are particularly affected, with asymmetry and a rather flat Cupid's bow persisting in the cleft group until 24 months. The presence of the 24 month control curves as a reference allows deviations at earlier time points for either group to be clearly identified. For instance, there is a narrowing of the nasal base for both groups as they grow from 3 to 24 months, and the upper face becomes longer relative to the

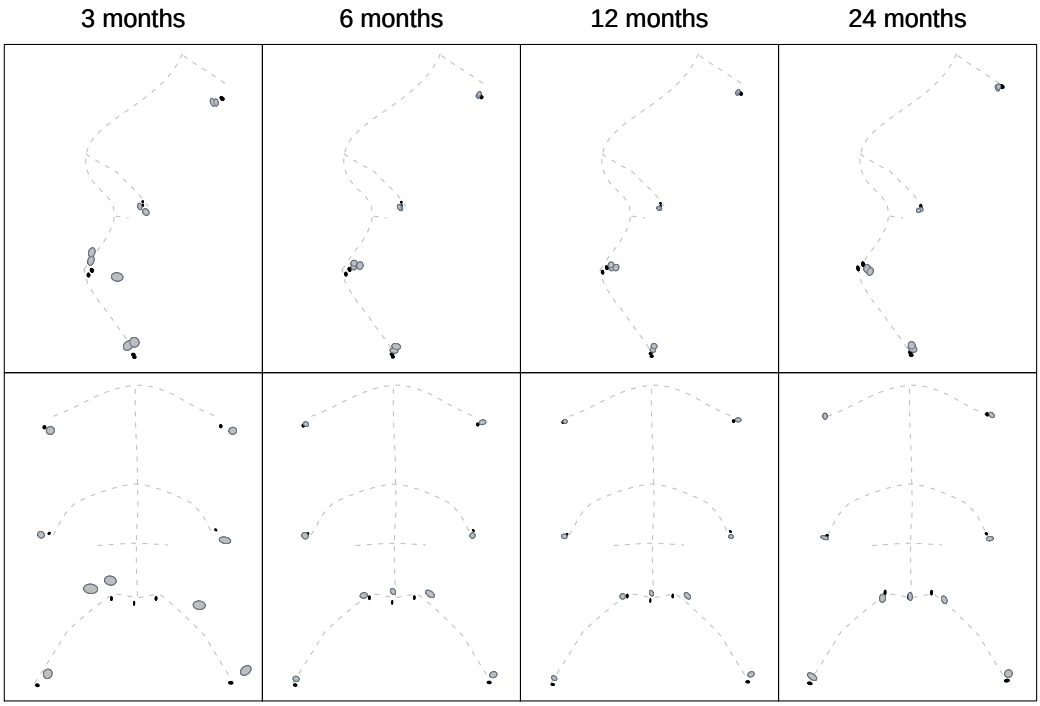

Figure 4.6: Under the independence model, 95% bivariate confidence regions for cleft (grey) and control (black) groups for each landmark in the lower face at each time point (upper: profile, lower: frontal view). The dashed lines give reference curves which are the same at each time point and fit most closely to the 24 month control landmarks.

lower.

The 95% bivariate confidence intervals are considerably larger for the cleft group than the controls, reflecting both the smaller group size and larger amount of variation amongst the clefts. The confidence intervals are particularly large for the cleft group at three months, which is to be expected since this is the point at which there is most variation between individuals.

Model checks

We consider various model checks to assess the fit of the model. Figure 4.7 gives the fitted mean trends over time with 95% confidence intervals for both groups, superimposed onto the empirical means for each tangent coordinate. These plots show that the fitted means

are reasonably close to their empirical equivalents, which indicates a good mean fit for the model. Perhaps most importantly, these plots indicate that for the most part a linear effect of time from 6 to 24 months (and from 3 to 24 months in the control group) is a reasonable fit, although there are some deviations from the general trend at 12 months.

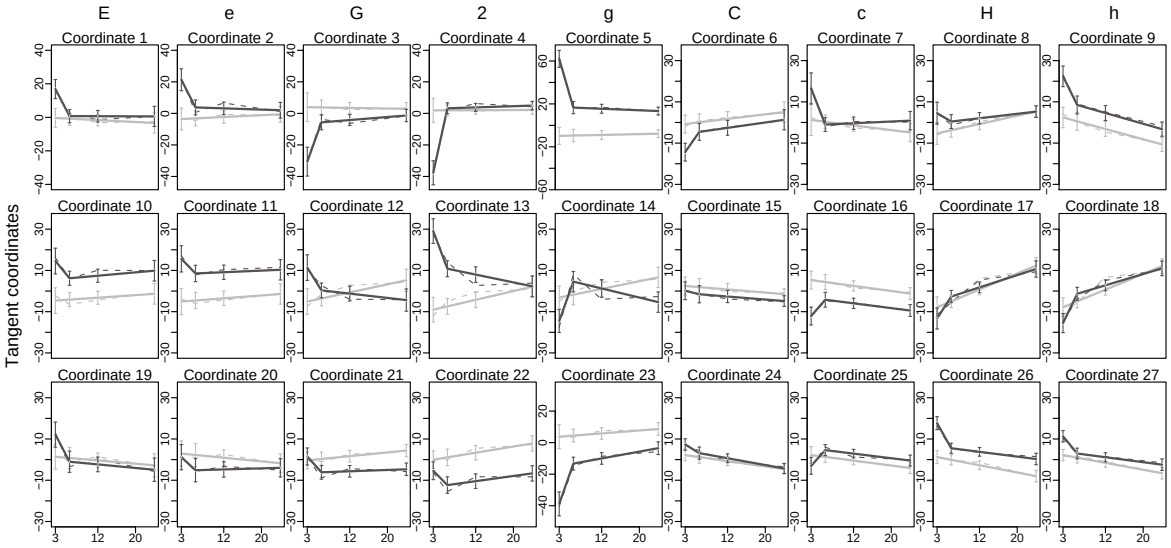

Figure 4.7: Fitted mean trends over time (solid) and 95% confidence intervals from the independence model, superimposed onto the empirical means (dashed) for each of the tangent coordinates and stratified into control (grey) and cleft (black) groups.

For a more intuitive interpretation of the model fit, we can superimpose the fitted means and 95% bivariate confidence intervals for the landmarks onto their empirical equivalents in shape space, in both profile and frontal views. These are displayed in Figure 4.8 at each time point and stratified by group and show a good mean fit of the model to the data at 3, 6 and 24 months. The 12 month confidence intervals do not show such a good fit to the empirical means, some of these being on the edge of intervals, so perhaps this is an indication that the linear trend is not entirely appropriate. This coincides with the deviations from the linear trend at 12 months in the tangent coordinates. However, these deviations may also be related to differential levels of missing data across time (cleft group: 40%, 42%, 37%, 33%; and control group: 18%, 8%, 9%, 10%; both at 3, 6, 12 and 24 months, respectively), and the fact that the linear mixed effects model takes into account both individual and group trends when estimating the mean (Diggle et al., 2002). The high levels of missing data should not be an issue in terms of ignorability since there is no basis to expect that the reason a cleft

patient drops out is related to facial shape, apart from perhaps the very extreme cases. For the sake of parsimony and the reasonably good fit of the linear trend, we continue with the linear trend for now, although this may not be appropriate for a final model.

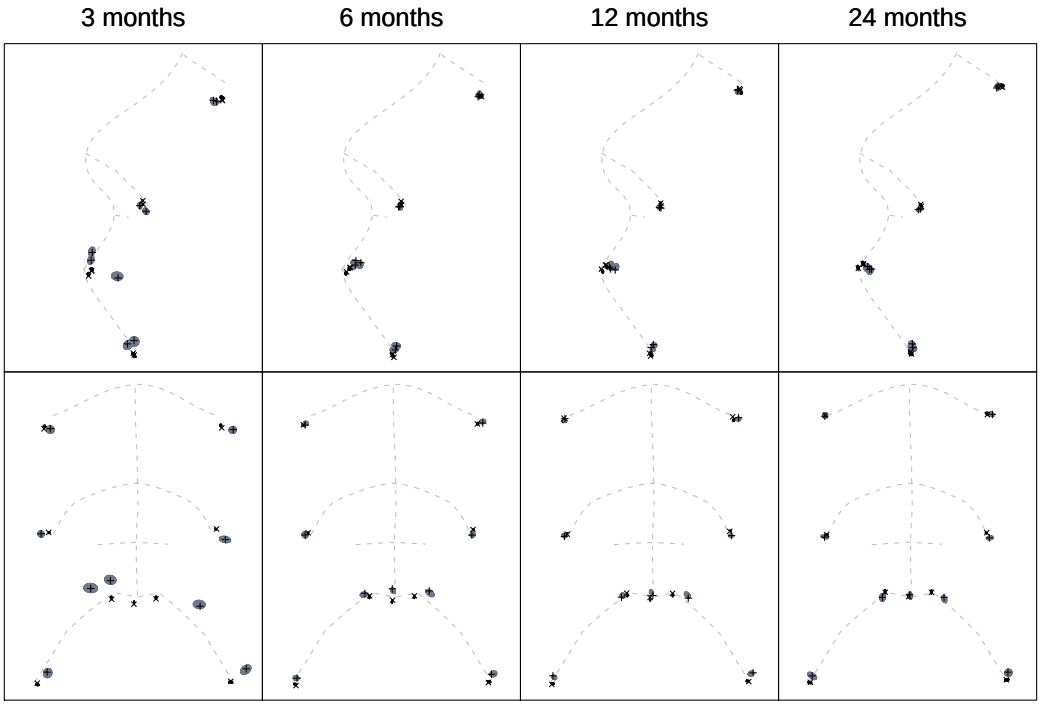

Figure 4.8: Fitted means and 95% bivariate confidence intervals from the independence model, superimposed onto the empirical means for each landmark in profile (upper) and frontal (lower) views. The empirical means for the cleft group are given by vertical crosses (on grey confidence intervals) and those for the controls by diagonal crosses (on black).

Finally, we consider the fit of the variance structure of the model to the data. The standardised residual for individual i may be obtained by (Verbeke & Molenberghs, 2000)

$$R_i = V_i^{-1/2}(y_i - X_i\beta), \tag{4.4}$$

where y_i is the response vector for individual i, across times and coordinates, X_i is the corresponding design matrix and β is the vector of fixed effects parameters. The covariance structure is modelled by the matrix $V_i = Z_i D Z_i' + \Sigma$ where D is the $km \times km$ covariance matrix of the random effects vector, $b_i = (b_{i1}, \ldots, b_{im})$, Z_i is the $n_i \times km$ coefficient matrix for the random effects and n_i is the number of observations over time on individual i, σ_s^2 is the

mean residual error variance across all pairs containing coordinate s and Σ is the constructed diagonal matrix containing these for all s (repeated across time points). The inverse square root matrix, $V_i^{-1/2}$, is calculated using the spectral decomposition, $V_i = \Gamma_i \Lambda_i \Gamma_i'$, where Γ_i is the matrix with the eigenvectors of V_i as its columns and Λ_i is a diagonal matrix of the corresponding eigenvalues, λ_{ij}, for $j = 1, \ldots, n_i$ (since V_i is $n_i \times n_i$). The inverse square root matrix, such that $V_i^{-1} = V_i^{-1/2} V_i^{-1/2}$ is

$$V_i^{-1/2} = \Gamma_i \Lambda^{-1/2} \Gamma_i',$$

where $\Lambda^{-1/2}$ is the diagonal matrix of $\lambda_{ij}^{-1/2}$ (Mardia et al., 1979).

The model only contains random intercepts and since for each coordinate they are assumed to be independent, D is a diagonal matrix containing only the variances, τ_s^2, of the random effect for each of the $s = 1, \ldots, km$ coordinates. Column s of the matrix Z_i has ones as its $(1 : n_{is} + (s-1)n_{is})$ entries (where $1 : n_{is}$ represents the sequence of integers from 1 to n_{is} inclusive) and is zero otherwise, and n_{is} is the number of measurements taken over time on coordinate s from individual i ($n_{is} \leq 4$ with equality only if there is no missing data on that individual).

Thus V_i is block-diagonal with the $n_{is} \times n_{is}$ blocks representing the covariance within coordinates across time points. The blocks are compound symmetric, with $\tau_s^2 + \sigma_s^2$ on the diagonal and τ_s^2 elsewhere, implying that the correlation is the same between any two time points for the same coordinate. Everywhere else is zero, since it is assumed that there is no correlation between coordinates.

Figure 4.9 gives boxplots of the standardised residuals for each tangent coordinate, calculated at each time point for each group. The plots of residuals indicate that the random error variance structure of the model may, in some cases, be inappropriate for the data. Although most of them are comparable, several coordinates have distributions of residuals that differ between the two groups and/or across time points. For the most part, where there are discrepancies, they involve higher variation in the residuals for the three month cleft individuals and it is perhaps not surprising that the model is less able to explain all the variation in this subgroup. Another useful model check is to plot the estimated absolute values of

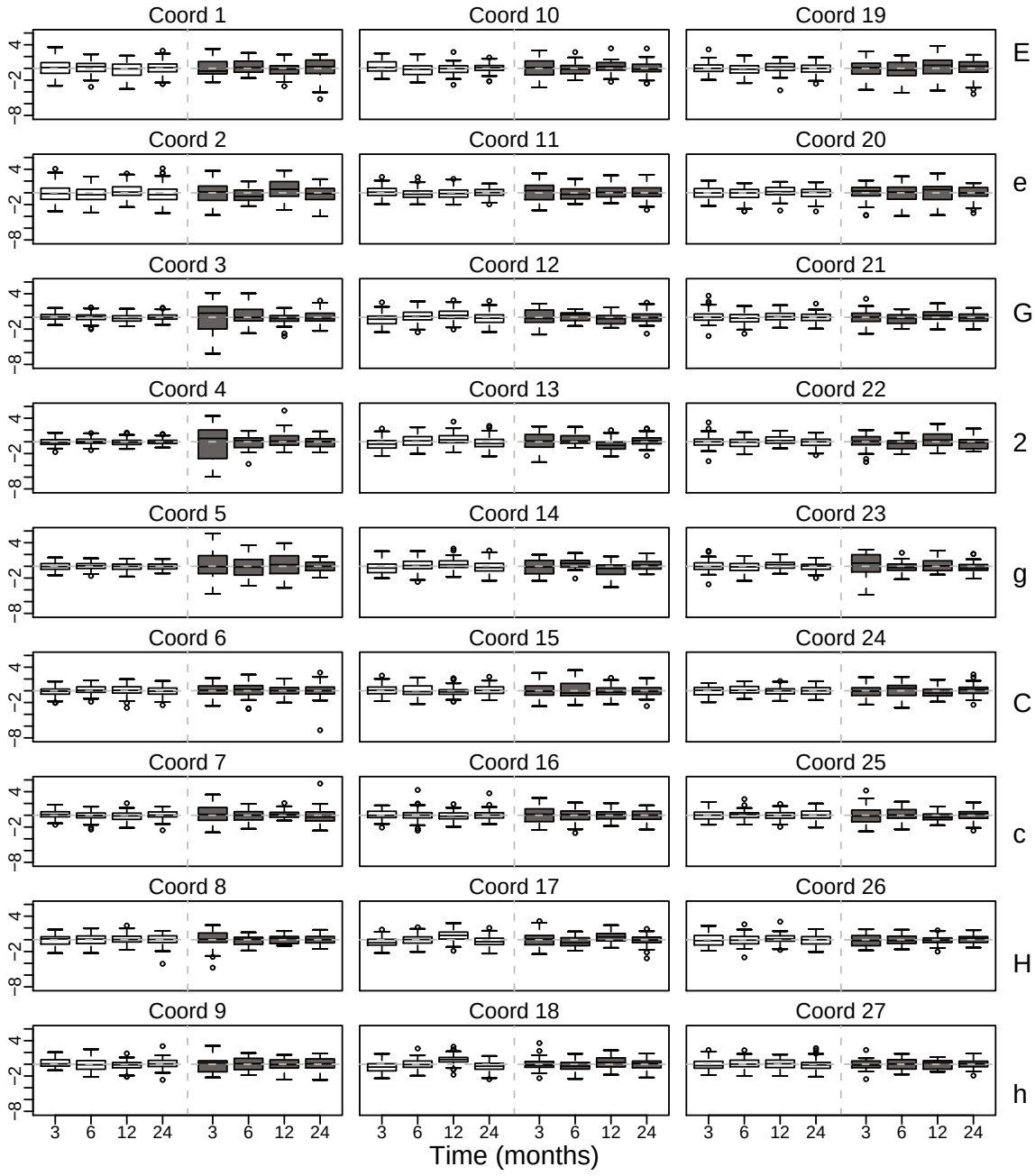

Figure 4.9: Boxplots of standardised residuals from the independence model, stratified by time point and into control (unfilled, left four boxes) and cleft (grey fill, right four boxes) groups

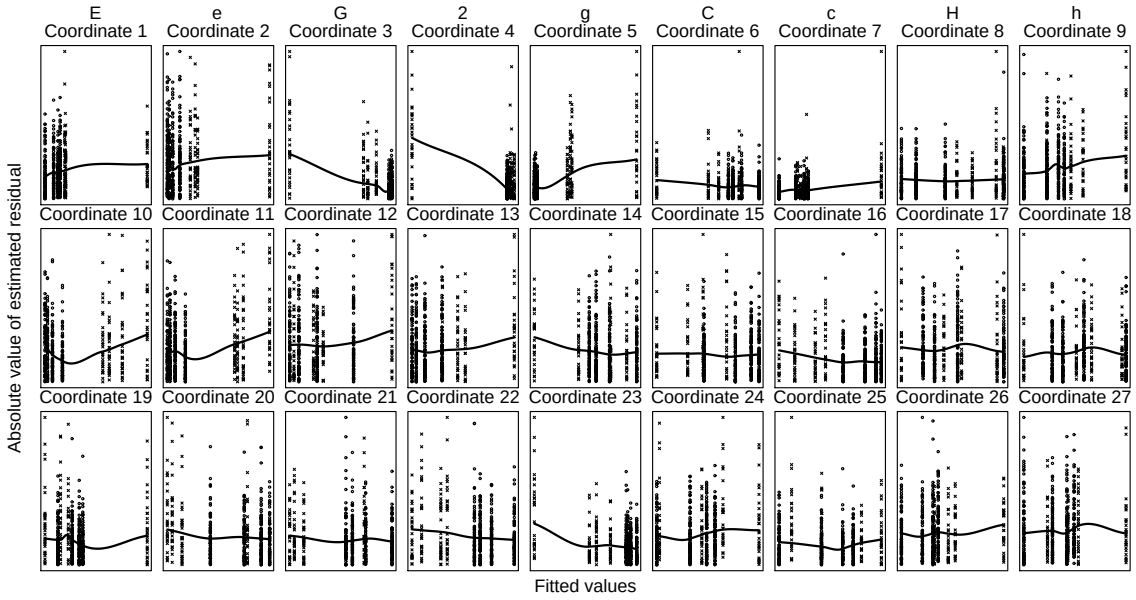

Figure 4.10: Plots of the absolute values of the (unstandardised) residuals versus the fitted values for each coordinate, under the independent random effects model, with a loess smooth line fitted

the unstandardised residuals, $|\epsilon_i| = |y_i - X_i\hat{\beta}|$, against the (group) population fitted values, $\hat{y} = X_i\hat{\beta}$, for each coordinate. Faraway (1997) recommends using the absolute values for the residuals rather than their actual values when checking constant variance across the fitted values. The plots are displayed for each coordinate in Figure 4.10, with smooth fitted lines which should be constant across the range of the fitted values if there is constant variance. The fitted values have been taken as the population rather than individual values, since we are primarily interested in the mean effects. The figure shows that the variance is far from constant as the fitted values change for many of the coordinates. Some coordinates, such as 3, 4 and 5, are particularly affected, with a large change in variance for a particular subset which in all three cases corresponds to the cleft group at 3 months.

The model checks have indicated that the independence model, while providing a reasonable mean fit, does not adequately describe the variance. This chapter therefore proceeds by fitting models with less simplistic covariance structures in order to obtain a potentially more appropriate model and to compare the fit against this reference model.

4.4.2 Correlated random effects

Different points on an individual's face are likely to be highly correlated with one another, particularly those that are close together. Therefore it is possible that the independence model is omitting an important aspect of the variation in assuming that the random effects are uncorrelated. Allowing correlation seems, intuitively, to be more appropriate than assuming independence. We proceed by fitting a model which is identical to (4.2), except it allows the random effects to be correlated:

$$\mathbf{b}_{ip} = (b_{ir}, b_{is}) \sim N(\mathbf{0}, D), \text{ where } D = \begin{pmatrix} \tau_r^2 & \tau_{rs} \\ \tau_{rs} & \tau_s^2 \end{pmatrix}.$$

The results are transformed back into shape space, as under the independence model, and bivariate confidence intervals in both the frontal and profile views are displayed in Figure 4.11.

The figure shows few differences when compared with the results of the independence model, in terms of mean or variance of landmark positions. It appears, therefore, that despite the rather more intuitive formulation of this model, there is actually little effect on the results. It may be that this model is as unrealistic as the independence model, since while the latter estimates none of the correlations between random effects for different coordinates, the former estimates all of them. In particular, it may not make sense to estimate the correlations between coordinates in different dimensions, and perhaps a model with some spatial structure may be more suitable; for example allowing points close together in the same dimension to be correlated whilst those far apart are assumed independent, or allowing correlation to decrease with distance.

Model checks

Since the mean fit of the model with correlated random effects is very similar to the independence model, we will display here only the fitted trends and 95% bivariate confidence regions in Figure 4.12 and omit the shape space plots. In general the mean fits for the individual coordinates are similar to those under the independence model but in this case

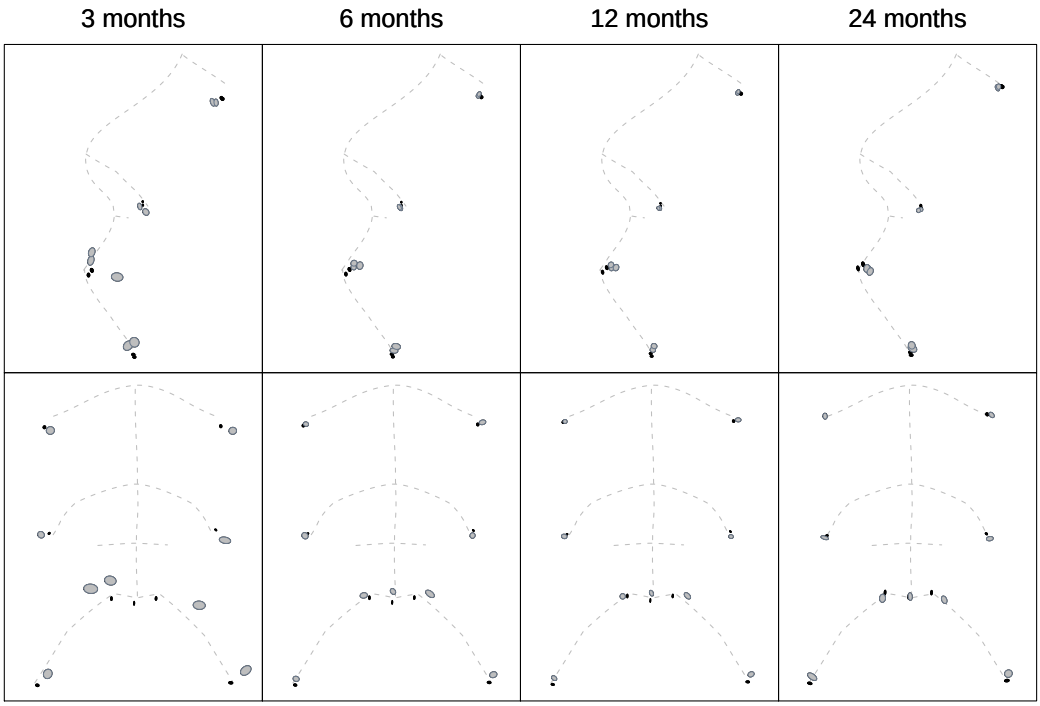

Figure 4.11: Under the correlated random effects model, 95% bivariate confidence regions for cleft (grey) and control (black) groups for each landmark in the lower face at each time point (upper: profile, lower: frontal view). The dashed lines give reference curves which are the same at each time point and fit most closely to the 24 month control landmarks.

several coordinates show very slightly larger bivariate confidence regions, particularly at three months. This indicates that the random effects covariance structure is having an effect on the standard errors of the fixed effects parameters, albeit slight, and that this may be a more appropriate structure than independence. The mean structure may not be entirely appropriate for the clefts, since for a small selection of coordinates, the average position at 12 months is not particularly well estimated by the linear trend. It may be worthwhile fitting a saturated mean model instead, but since the discrepancies are small and only for a small number of coordinates the linear trend will continue to be fitted for these preliminary investigations.

Theoretically, the standardised residuals under the correlated model may be estimated from (4.4). In this case the covariance matrix D is no longer diagonal, and contains on its off-diagonals the covariances τ_{rs} between random intercepts b_{ir} and b_{is}, for $r, s = 1, \ldots, M$,

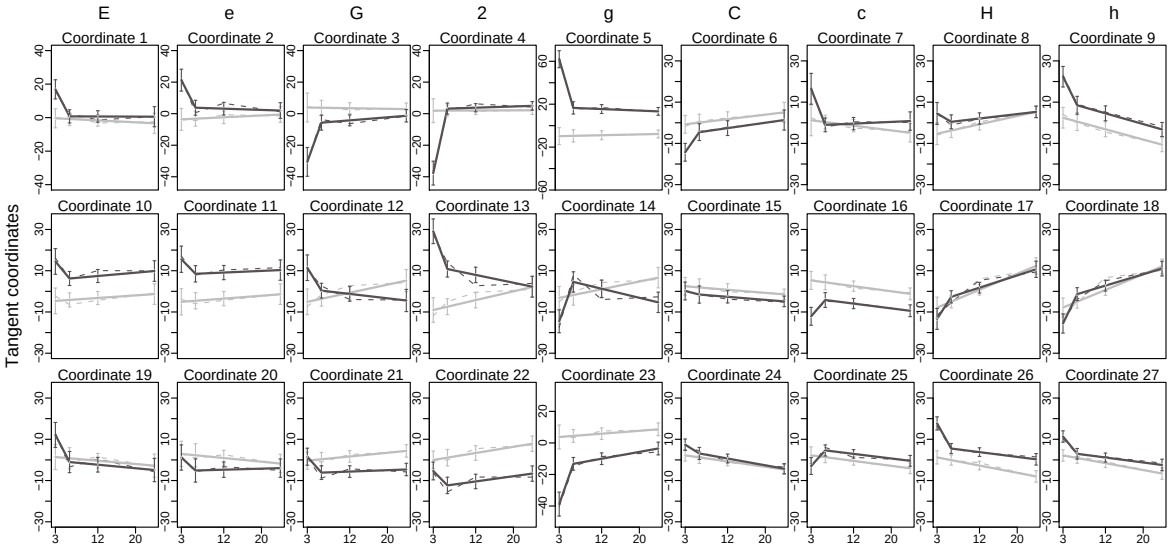

Figure 4.12: Fitted mean trends over time (solid) and 95% confidence intervals from the correlated random effects model, superimposed onto the empirical means (dashed) for each of the tangent coordinates and stratified into control (grey) and cleft (black) groups.

where $M = km$:

$$D = \begin{pmatrix} \tau_1^2 & \tau_{12} & \cdots & \tau_{1M} \\ \tau_{21} & \tau_2^2 & \cdots & \tau_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \tau_{M1} & \tau_{M2} & \cdots & \tau_M^2 \end{pmatrix}.$$

Since the variances in the diagonal of D have been averaged and the covariances each come from different models, there are no restrictions on the estimation of the parameters. Furthermore, the estimated covariances may not always accurately reflect the estimates that would be obtained under the fully joint model since each pairwise model considers only the covariance between the random effects for a particular pair of coordinates and ignores any correlations that either may have with others in the dataset. Despite the model estimating correlations between all pairwise combinations and therefore emulating a fully correlated model, effectively each correlation is estimated under the assumption of independence from any other coordinates. The fully joint model, in contrast, will take into account all of the coordinate correlations at the same time. The result, for this application, is that when all of the variances and covariances are slotted into their respective positions in D, the matrix is not necessarily positive definite and, by definition, is therefore not a real covariance matrix.

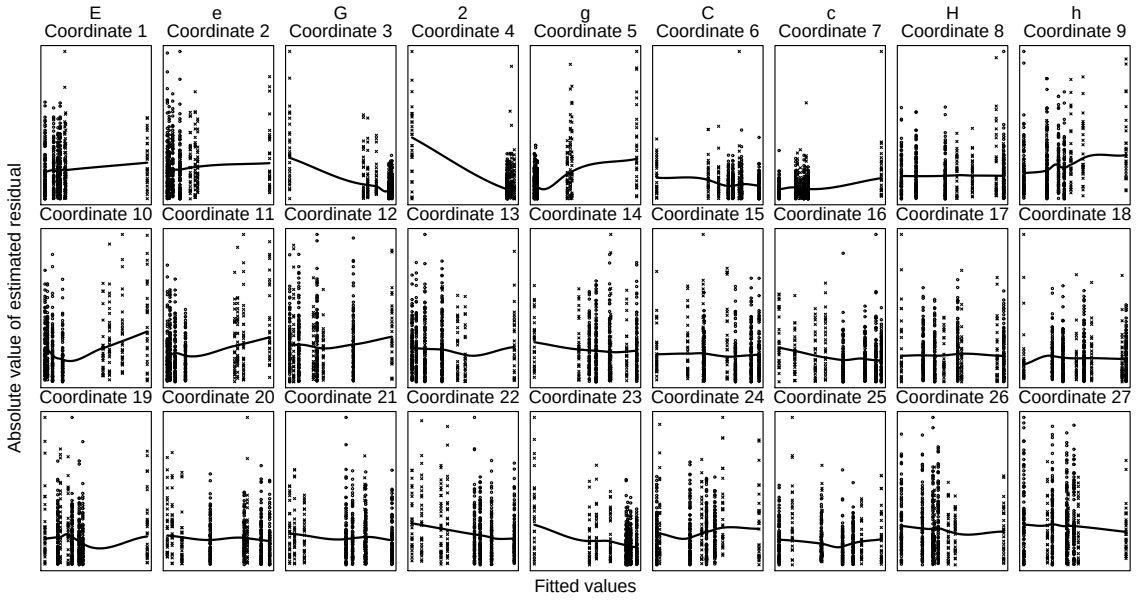

Figure 4.13: Plots of the absolute values of the (unstandardised) residuals versus the fitted values for each coordinate, under the correlated random effects model, with a loess smooth line fitted

This issue has no real effect on the model results since interest lies only in the fixed effects, but it does prevent ascertainment of the standardised residuals. Instead the unstandardised residuals are considered against the fitted values, in Figure 4.13. The plots show an improvement in terms of constant variance for some coordinates, and a worsening for others, but generally they suggest a poor fit of the covariance structure to the data.

An alternative is to plot the estimated correlations of the random effects from the pairwise models against the average distance between the coordinates within each dimension. Although the correlations are also calculated between coordinates within different dimensions, there is no clear way of defining a distance between these so they are displayed unordered. The Euclidean distance between two coordinates in the same dimension is calculated in shape space as the absolute difference between the positions of those coordinates for each individual, and the mean of those distances is then obtained.

The correlations are calculated using the estimated variances and covariances of the random effects within each pairwise model (so do not use the averaged estimate for the variances). The resulting within-dimension plots are displayed in Figure 4.14 for each of the x, y and

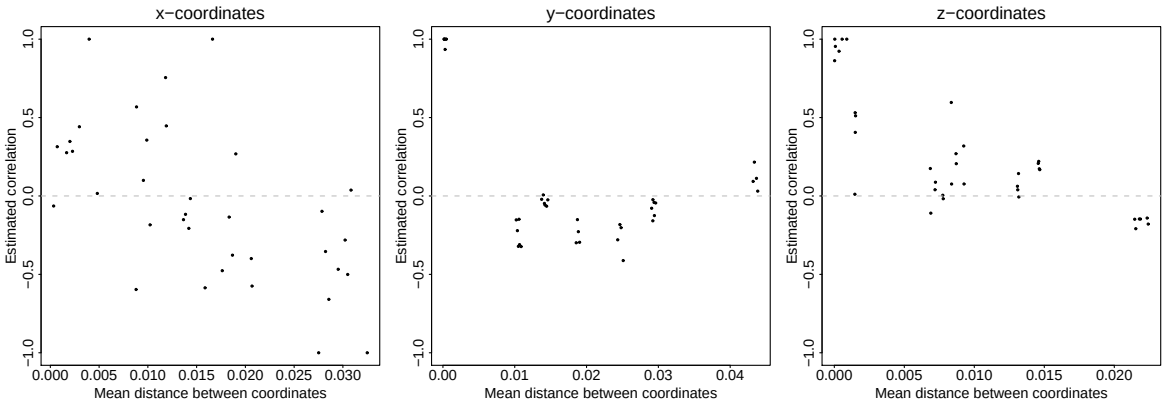

Figure 4.14: Estimated correlations between the random effects for pairs of coordinates, versus the mean distance between them in shape space, for each of the three dimensions

z dimensions and show that, for the x-coordinates, there is a clear relationship between increasing distance and decreasing correlation and that this reflects the symmetry of the face, with approximately half of the correlations being above zero and (the further apart) half below. There is a similar inverse relationship for the z-coordinates, although since there is no particular symmetry in this axis, the correlations are generally positive or around zero. The relationships for these two dimensions suggests that there is a need for allowing covariance between the random effects and, intuitively, it makes sense that the corresponding correlations should decrease as the distance between coordinates increases, or become negative for paired landmarks. For the y-coordinates, however, there is very little relationship, with a small selection of correlations close to one and the rest around zero and this may indicate a poorer fit of the model for these coordinates, or that allowing correlation between their random effects is unnecessary.

The unordered correlations between random effects for coordinates in different dimensions are displayed in Figure 4.15. For the most part, the correlations are below 0.5 in absolute value, but there are several, particularly between the x and z dimensions, that have higher correlations. Intuitively, it seems that for the most part there should be little relationship between points in different dimensions, but this model is estimating some reasonably high correlations. This may be due in part to the cleft defect, since it is reasonable to assume that, for example, a very asymmetric nose may also be rather flat.

One major issue that is highlighted in the plots of Figures 4.15 and 4.14 is that there are

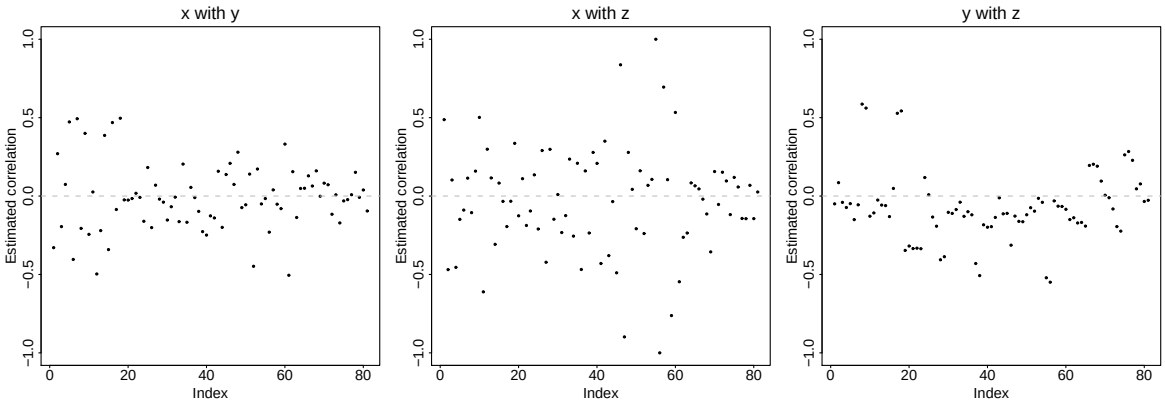

Figure 4.15: Estimated correlations between the random effects for pairs of coordinates in different dimensions (in no particular order)

some correlations that are estimated as exactly one or negative one. This may be due to model misspecification, resulting in the maximum of the log-likelihood being on the boundary of the parameter space, and can also result from small variances (Verbeke & Molenberghs, 2000), although this should have been dealt with to some extent by multiplying all of the tangent coordinates by 10000. Some refinement of the covariance structure may therefore be necessary, to allow correlations between the random effects but to ensure that the parameter estimates converge to values within the parameter space.

It appears, then, that allowing the random effects to be correlated has very little effect on the mean fit of the model in comparison with the independence model. However, Figure 4.14 suggests that the random effects correlations should be estimated, but with due care in order to obtain sensible parameter estimates.

4.4.3 Different random error variances for each group

We proceed by fitting a model which allows for different random error variances across the two groups but which has independent random effects, since allowing correlations between random effects in this situation causes failure of convergence of the model. This may not be a particularly realistic model since the analyses of the previous section indicated that correlations were necessary, but it is useful for illustration purposes. The model is identical

to the setup in (4.2), except for the distribution of the random error term:

$$\boldsymbol{\epsilon}_{ip}(t) \sim N(\mathbf{0}, \sigma_p^2 \delta_{pg_i}^2 I),$$

where $\delta_0 = 1$ for identifiability, and δ_1 therefore corresponds to a scale parameter for the ratio of the cleft to control group random error variance (Pinheiro & Bates, 2000). The variance matrix, D, of the vector of random effects, $\mathbf{b}_{ip}$, is diagonal, as under the independence model, with the same structure as (4.3).

Since the random error variance is not coordinate specific, one estimate of the scale parameter δ_{p1} is obtained from each bivariate model. This estimate is effectively averaged over the pair of coordinates within any particular model and we could average across the scale parameters obtained from all the pairwise models to obtain one estimate of δ_1. This may not be entirely realistic, since we may expect the scale parameter to vary from one coordinate to another. However, unless interest actually lies in the interpretation of the scale parameter, the pairwise modelling framework results in there being no difference between a model in which there is one aggregate estimate of the scale parameter, and one in which the scale parameter is estimated for each pairwise combination of coordinates. Since there is no particular interest in this application in the actual value of the scale parameter and we wish only to account for its potential effect, it is irrelevant here whether we aggregate the estimates or not (except for model comparison purposes, when some aggregation will be carried out). A more realistic framework, however, may be one in which the scale parameter is estimated separately for each coordinate, since Figure 4.9 indicates that for some coordinates the distributions of the cleft and control residuals differ, whereas for others there is little difference.

The parameter estimates from this model are transformed back into shape space in the usual way and the 95% bivariate confidence intervals for the profile and frontal views of the landmarks are presented in Figure 4.16. The results from this model are very similar to the previous models considered and it would therefore appear that allowing different random error variances has little effect on either the estimated average position of the landmarks, nor in their uncertainty.

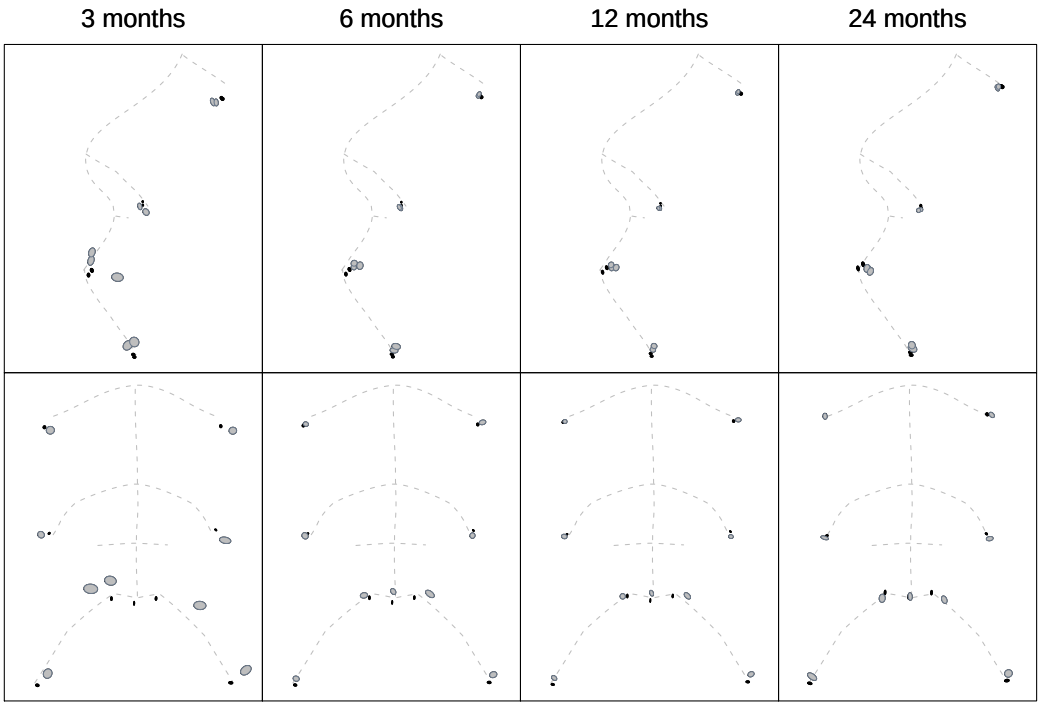

Figure 4.16: Under the model allowing different random error variances for the two groups, 95% bivariate confidence regions for cleft (grey) and control (black) groups for each landmark in the lower face at each time point (upper: profile, lower: frontal view). The dashed lines give reference curves which are the same at each time point and fit most closely to the 24 month control landmarks.

Model checks

Figure 4.17 shows boxplots of the scale parameter estimates, where each boxplot contains all of the scale estimates in which the particular coordinate is included. Most of the estimates range between about one and two, but there are a considerable number of them over two, up to about 3.5. This suggests that there is a difference between the groups in their random error variances, at least for some pairs. It is odd that there are no scale parameter estimates below or equal to one as there are coordinates for which there is either very little difference in variance between the groups or the control group has slightly higher variance (from Figure 4.5). This may arise because the scale parameter is estimated for a pair rather than each coordinate separately, but one would still expect a selection of estimates to be less than or equal to one. It is unclear, therefore, why this is the case, but it seems rather unrealistic.

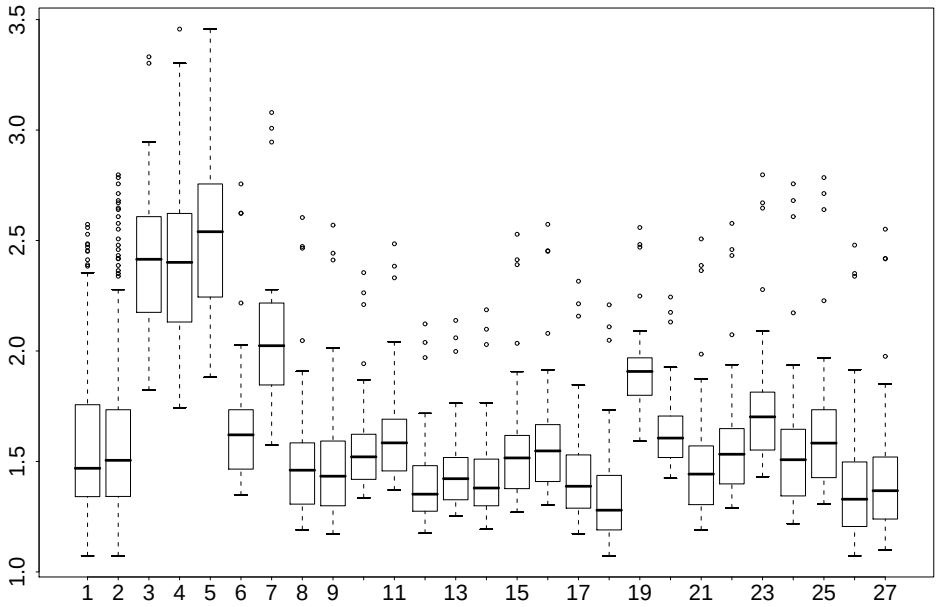

Figure 4.17: Boxplots of the scale parameter estimates for each coordinate (each from a bivariate model with another coordinate), representing the ratio of cleft to control group random error variances

The mean fit of the model is very similar to previous models, so we include only the fitted mean trends over time with 95% confidence intervals in Figure 4.18, and the standardised residuals. The fitted mean trends show little difference in either mean fit or uncertainty when compared to either the correlated or independent random effects model.

The residual for individual i at time point t and coordinate s is obtained here using (4.4) since the random effects for the different coordinates are independent. Once again the covariance matrix, V_i, is block-diagonal, but in this case the blocks, V_{is}, for coordinate $s = 1, \ldots, km$ are defined as

$$V_{is} = Z_{is}\tau_s Z_{is}' + \sigma_s^2 \delta_{sg_i} I_{n_{is}}, \tag{4.5}$$

where Z_{is} are the $n_{is} \times 1$ columns of Z_i (as defined under the independence model) corresponding to coordinate s and τ_s is the scalar variance of the random effect b_{is}. If $g_i = 1$, $\hat{\delta}_{sg_i}$ is the mean of the estimates of the scale parameter that correspond to coordinate s, obtained from the pairwise models which contain that particular coordinate; and otherwise it is equal

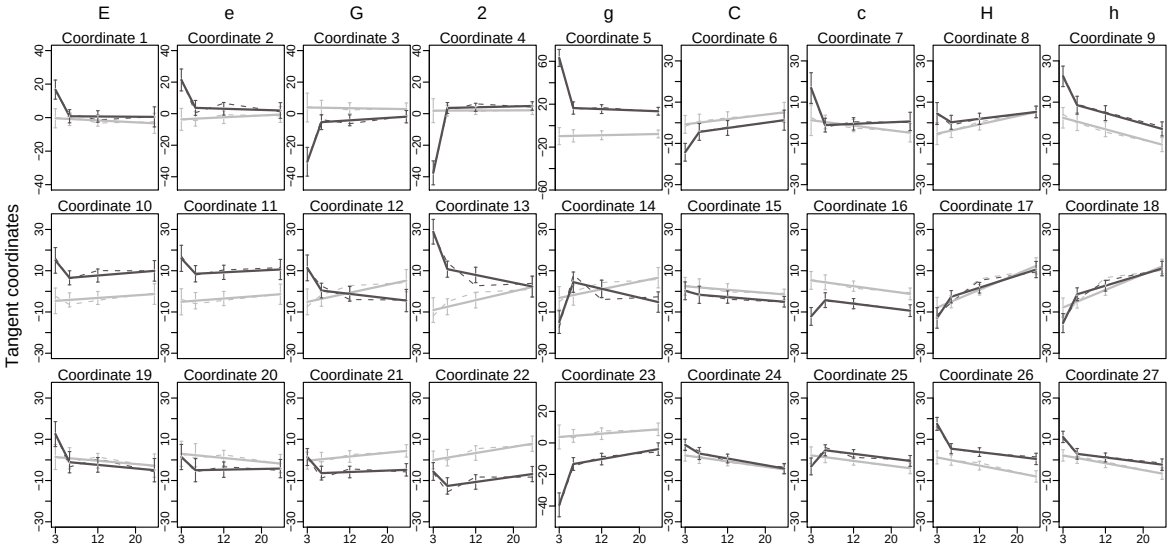

Figure 4.18: Fitted mean trends over time (solid) and 95% confidence intervals from the model with different random error variances for each group, superimposed onto the empirical means (dashed) for each of the tangent coordinates and stratified into control (grey) and cleft (black) groups.

to one. The estimate of σ_s^2 is likewise averaged over pairwise models containing coordinate s. This method of averaging seems a more appropriate way of estimating the residuals than by taking the mean across all of the scale parameter estimates, particularly since there was a reasonably wide range of estimates, as displayed in Figure 4.17. The blocks V_{is} now have $\tau_s^2 + \sigma_s^2\delta_{s1}$ on the diagonal if individual i is in the cleft group and $\tau_s^2 + \sigma_s^2$ otherwise. The off-diagonal elements of the blocks in both cases are τ_s^2.

The standardised residuals are displayed in Figure 4.19. The distributions of residuals show a much more even spread across time points and coordinates than those obtained from the independence model. There are still some discrepancies (for example, coordinates 3 and 4) but they are markedly reduced when compared to those of Figure 4.9. The unstandardised residuals versus the fitted values are displayed in Figure 4.20 and show some flattening out of the variance of the residuals as the fitted values change, in comparison to the independence model. A substantially higher number of coordinates now exhibit reasonably constant variance, especially for the y- and z-dimensions. However, there are still some major deviations from this, particularly amongst the x-coordinates 3, 4 and 5.

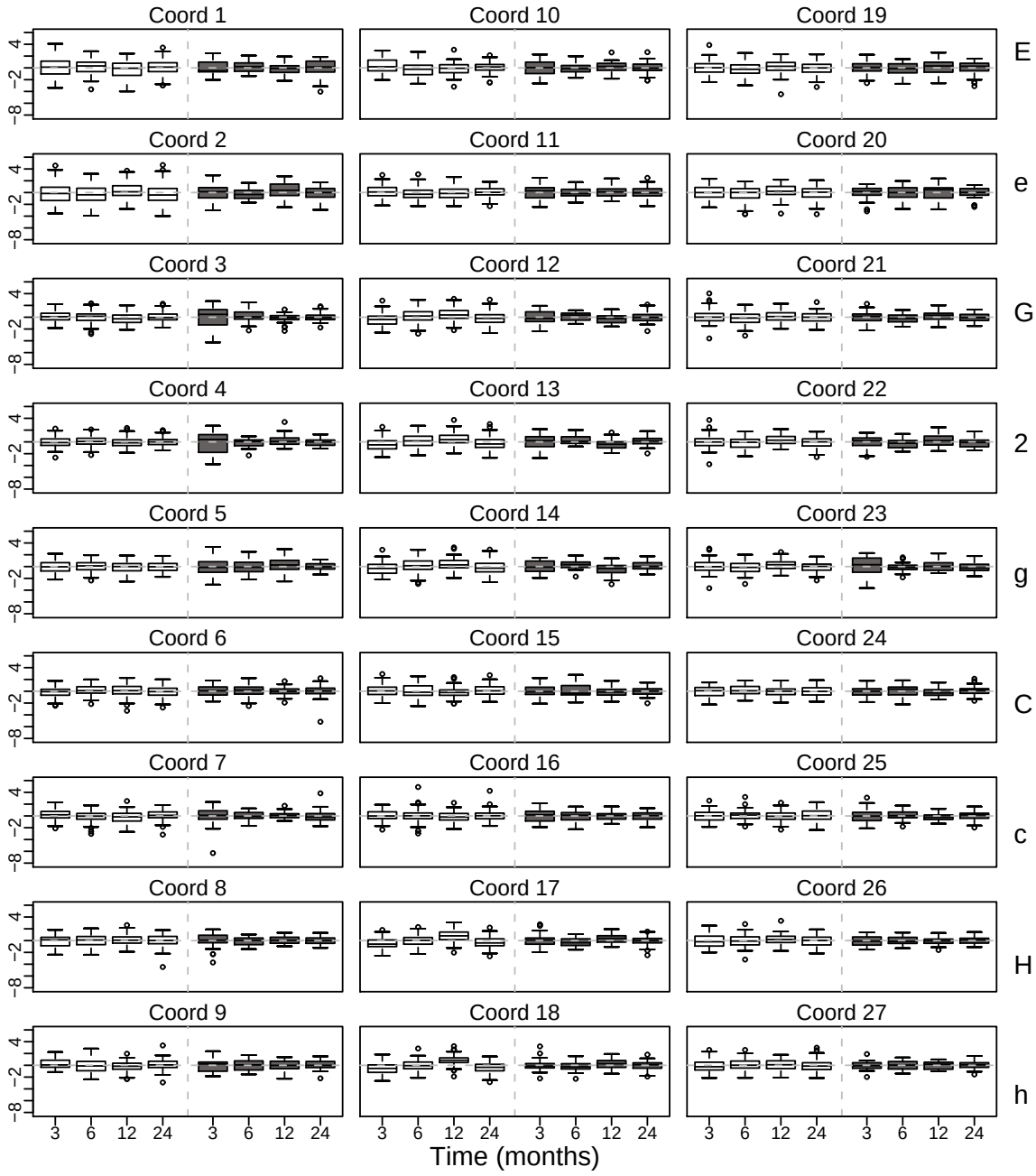

Figure 4.19: Boxplots of standardised residuals from the model with different random error variances for each group, stratified by time point and into control (unfilled, left four boxes) and cleft (grey fill, right four boxes) groups

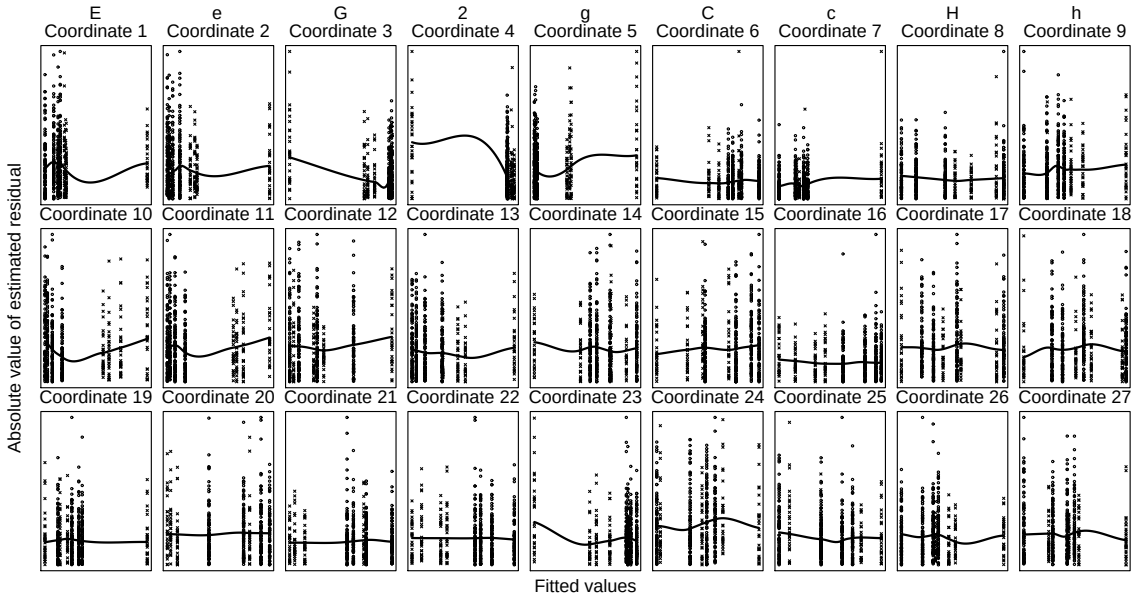

Figure 4.20: Plots of the absolute values of the (unstandardised) residuals versus the fitted values for each coordinate, under the different random error variances model, with a loess smooth line fitted

This analysis suggests, then, that although there are no substantial differences between the results comparing the cleft and control group means and confidence intervals in shape space, this variance structure is somewhat more appropriate than the independence model. However, the lack of convergence of the model when the random effects are allowed to be correlated is a problem, as the analyses of the previous section suggested that these were necessary. This model is also restrictive in that the scale parameter between the random error variances must be pair-specific rather than coordinate specific, and the covariance structure is still not entirely appropriate.

4.4.4 Different random effects variances for each group

An alternative formulation for different group variances is considered here, allowing for different random effects variances across the two groups. This is coordinate specific because the variance of the random intercepts already depends on the coordinates. The random error variance is also allowed to be coordinate rather than pair specific in this case. We therefore

fit the following model:

$$\mathbf{y}_{ip}(t) = \boldsymbol{\beta}_{0p} + \mathbf{b}_{ipg_i} + \boldsymbol{\beta}_{1p} g_i + \boldsymbol{\beta}_{2p} t + \boldsymbol{\beta}_{3p} g_i \cdot t + \boldsymbol{\beta}_{4p} g_i \cdot S(t) + \boldsymbol{\epsilon}_{ip}(t), \tag{4.6}$$

where, for an individual i who is in the control group (when $g_i = 0$), their random effect vector is distributed as

$$\mathbf{b}_{ip0} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{r0}^2 & 0 \\ 0 & \tau_{s0}^2 \end{pmatrix}\right).$$

For a cleft group subject ($g_i = 1$),

$$\mathbf{b}_{ip1} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{r1}^2 & 0 \\ 0 & \tau_{s1}^2 \end{pmatrix}\right).$$

Thus there is a different random intercept b_{irg_i} for each coordinate r, which is also allowed to differ across groups using the group indicator g_i. The random error vector, $\boldsymbol{\epsilon}_{ip}(t) = (\epsilon_{ir}(t), \epsilon_{is}(t))$, is distributed as $N(0, \Sigma_p)$, where Σ_p is a diagonal matrix with the variances σ_r^2 and σ_s^2 as its entries. All b_{irg_i} are assumed independent, since allowing otherwise results in some random effects variances being estimated as exactly zero, a sign that there is a problem with the estimation of these variances (this will be discussed later).

Bivariate 95% confidence intervals for the profile and frontal view of each of the landmarks obtained from this model fit are displayed in Figure 4.21 and show close similarities to the results of the previous models in both the mean landmark positions and the extent of their variation.

Model checks

Once again, the mean fit of the model is presented in the form of the fitted mean trends and 95% confidence intervals superimposed onto the empirical means, which are displayed in Figure 4.22. The fit is very similar to those under the correlated random effects and different random error models.

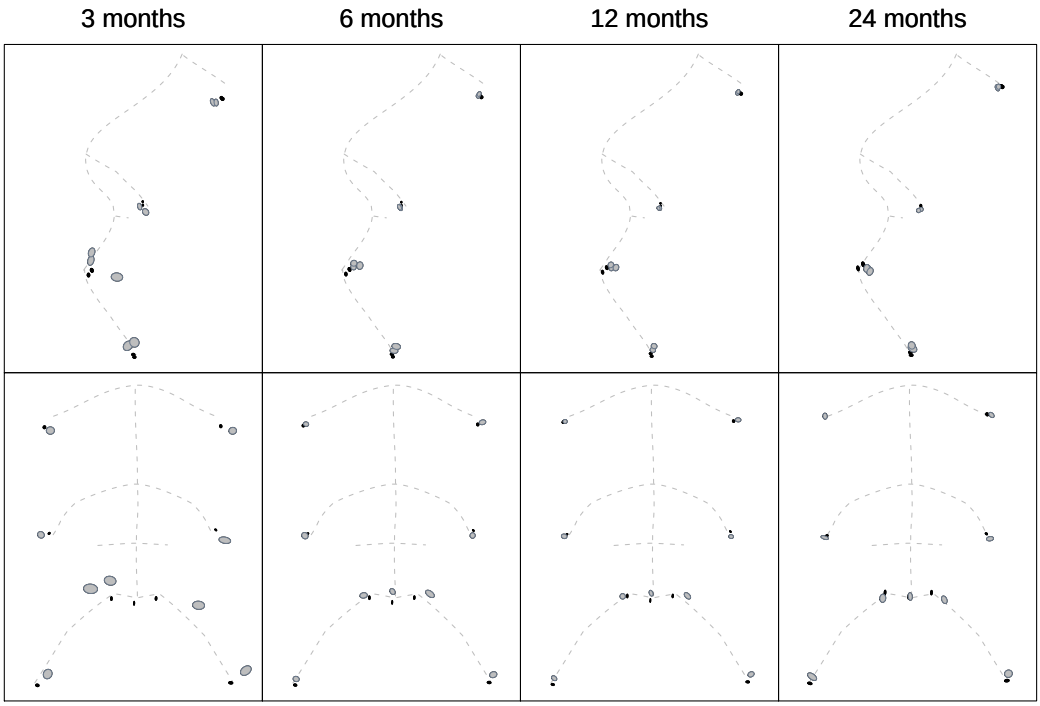

Figure 4.21: Under the model allowing different random effects variances for each group, 95% bivariate confidence regions for cleft (grey) and control (black) groups for each landmark in the lower face at each time point (upper: profile, lower: frontal view). The dashed lines give reference curves which are the same at each time point and fit most closely to the 24 month control landmarks.

The standardised residuals are estimated again from (4.4) since the random effects are again assumed independent. Under this model, however, the block-diagonal covariance matrix is defined as

$$V_i = Z_i D_{g_i} Z_i' + \Sigma,$$

where D_{g_i} has the variances τ_{sg_i} on the diagonal ($s = 1, \ldots, km$) and is zero otherwise, and Σ is a diagonal matrix containing the random error variances, σ_s^2, as its entries. The blocks V_{is} in this case therefore contain $\tau_{s0}^2 + \sigma_s^2$ on the diagonal and τ_{s0}^2 otherwise for those in the control group, and $\tau_{s1}^2 + \sigma_s^2$ on the diagonal with τ_{s1}^2 otherwise for the cleft subjects.

The boxplots with the standardised residuals for this model are displayed in Figure 4.23. The extent of the variability amongst the residuals is slightly reduced in comparison to the independence model, but is generally larger than the different random errors model,

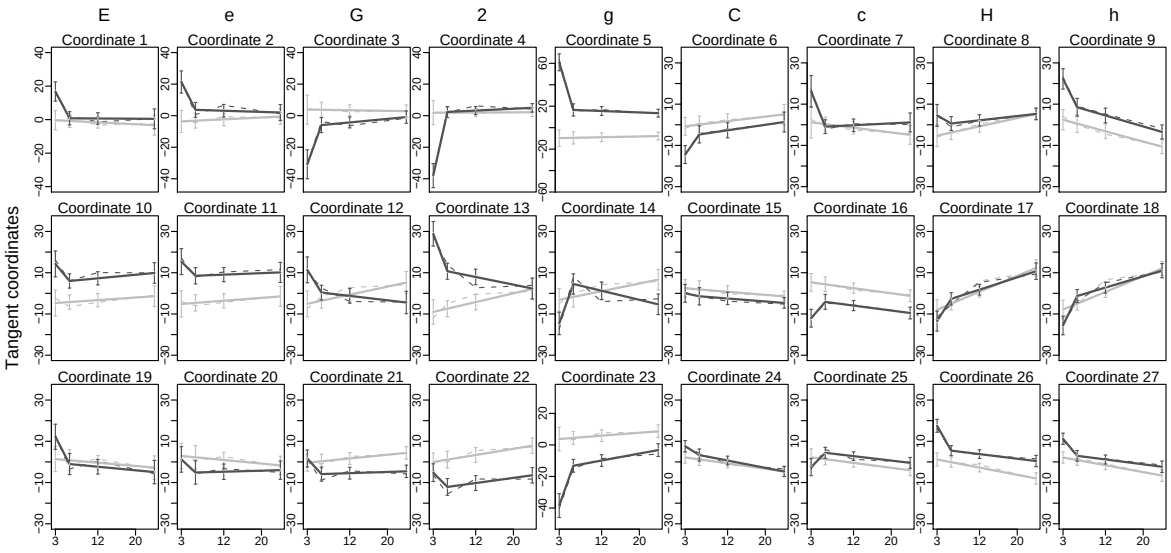

Figure 4.22: Fitted mean trends over time (solid) and 95% confidence intervals from the model with different random effects variances for each group, superimposed onto the empirical means (dashed) for each of the tangent coordinates and stratified into control (grey) and cleft (black) groups.

indicating that that model explains more of the variation than the current model. There remain some differences here in the distributions of residuals between time points and groups, particularly with respect to the 3 month time point amongst the cleft patients. Figure 4.24 contains the plots of residuals versus fitted values for each coordinate. The plots show similar patterns to those exhibited by the different random error variances model, in that there is an improvement in terms of constant variance over the independence model but there are still some substantial deviations.

A final check for this model is to compare the estimated random effects variances for each group with the unstratified random effects variance from the models where the estimate is not stratified by group. This is displayed in Figure 4.25, where the random effects standard errors have been given for each coordinate under the independence, correlated random effects and different random error variances models, along with the separate cleft and control group random effect standard error (SE) estimates from the current model. A number of observations may be made from this plot: that the unstratified SEs are always between the stratified equivalents and that the former tend to be closer to the control group SE estimates (because this group has a larger sample size); that the cleft group tends to have higher random effects

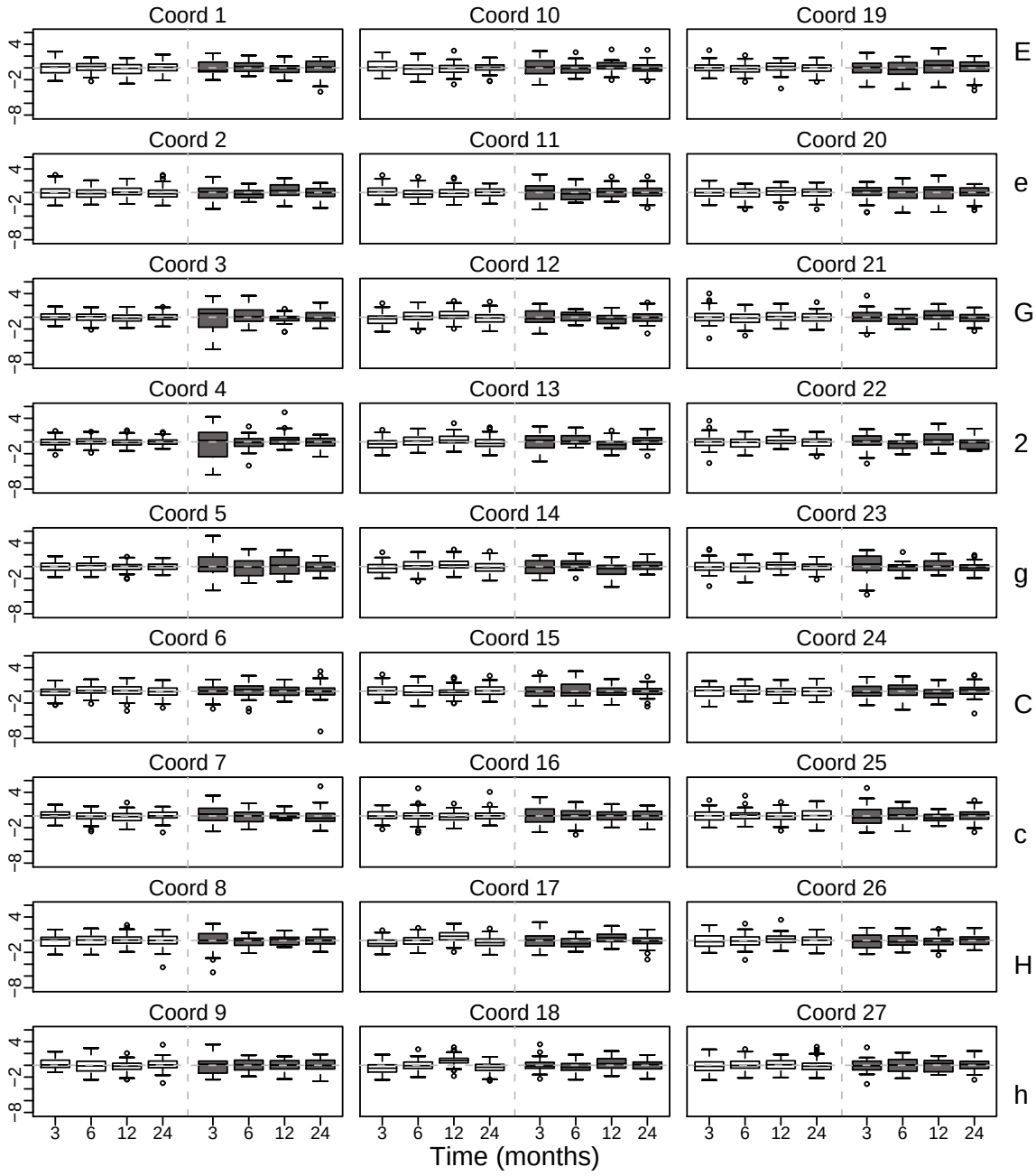

Figure 4.23: Boxplots of standardised residuals from the model with different random effects variances for each group, stratified by time point and into control (unfilled, left four boxes) and cleft (grey fill, right four boxes) groups

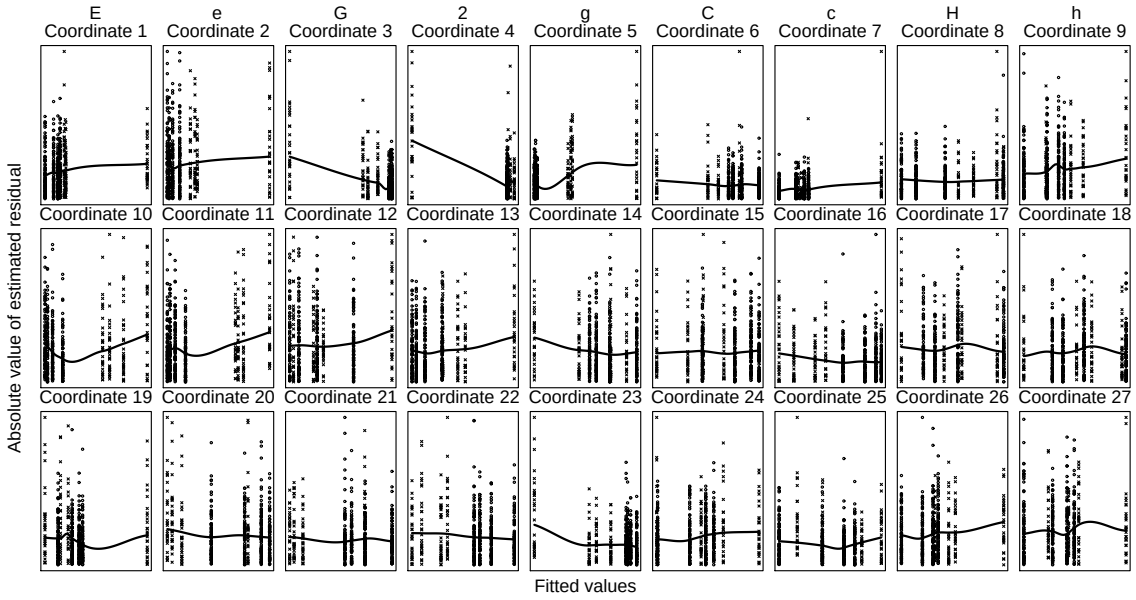

Figure 4.24: Plots of the absolute values of the (unstandardised) residuals versus the fitted values for each coordinate, under the different random effects variances model, with a loess smooth line fitted

SE estimates than the control group, except for a handful of coordinates that correspond to the coordinates in Figure 4.5 which demonstrated higher variance for the controls than the cleft group; and that the random effects SE estimates from the independence model are very similar to those from the correlated random effects model and generally higher than those from the different random error variances model. The first two observations are to be expected and indicate suitable model fit (of the current model) and the latter point shows that allowing the random error variances to be different also has the effect of reducing the random effects variances. This, again, is to be expected since explaining more of the variation by the random error is bound to reduce the amount that the random effects must explain.

Another observation from Figure 4.25 is that coordinate 4, corresponding to the x-coordinate of the midline point "2" in Figure 4.1, has an extremely low random effects SE estimate (0.0006) for the control group, and that this results in a ratio of the cleft to control group random effects SEs of over 13000. It is clear from Figure 4.25 that the control group SE is considerably lower than all other estimates, whilst the cleft group SE is of a similar order to other coordinates. The midline coordinates appear to vary very little amongst the controls,

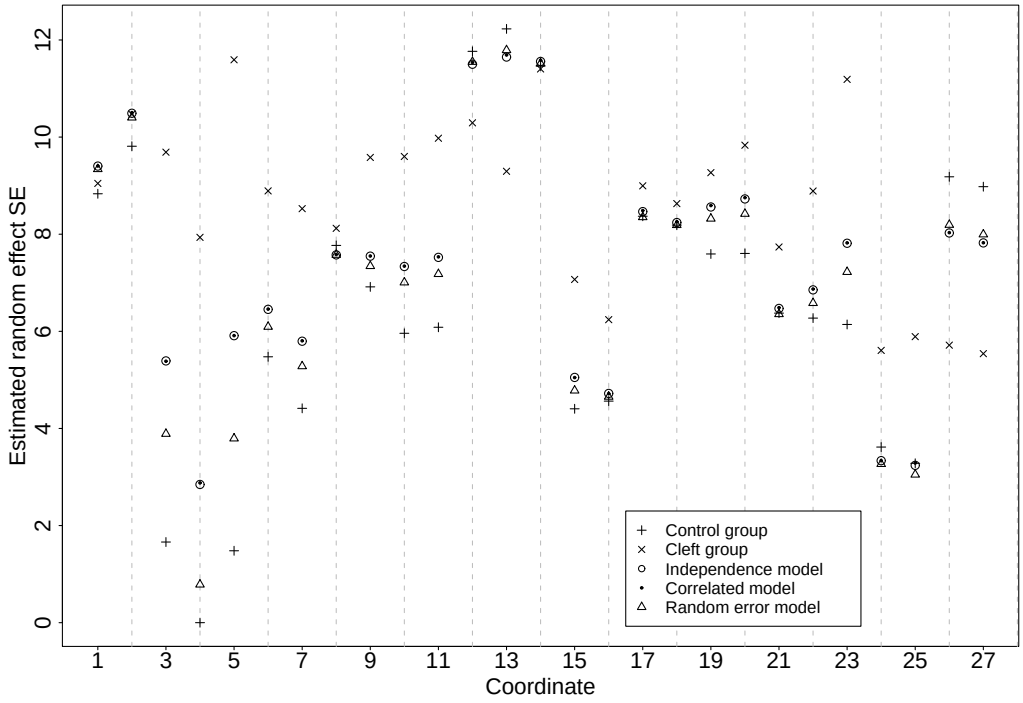

Figure 4.25: Random effects standard error estimates (by coordinate) for the group stratified random intercepts under the current model with a separate estimate for the cleft (plusses) and control (crosses) groups); and unstratified random effects standard error estimates from the independence (large, empty circles), correlated random effects (triangles) and different random error variances (small, filled circles) models

at least in the x dimension, whereas clearly there is substantial variation between the cleft individuals in the position of these landmarks and this may explain the difference in the variance of the random effect. However, it seems unlikely that the coordinate would have as little variance amongst the controls as the estimate suggests.

4.5 Discussion

The pairwise approach itself is generally very easy to fit to data such as these, using the `lmepair` function, but it is not without its problems regarding convergence of models, spurious zero estimates of random effects variances and correlations being estimated as exactly one or negative one. The set of landmarks chosen for the analyses of this chapter was not

optimal in terms of studying the cleft defect, and this was a result of prior problems with the random effects variance parameter estimation for the x-coordinates of several of the midline landmarks (not included in this example), particularly in terms of variances being estimated as exactly zero (and corresponding individual random effects also all being estimated as zero). Hence, for the purposes of the illustration of the models, their comparisons and appropriateness of fit, a set of landmarks was chosen that caused fewer such difficulties. Even for this set of landmarks, however, there were still problems allowing correlated random effects for the models with different random effects or error variances across the two groups. In the next chapter, there will be further discussion of these issues and some potential methods of avoiding them will be suggested.

Computer memory is also an issue, such that there is an upper limit of about ten landmarks (in three dimensions) which may be included in the analysis, but with a machine with greater memory it is likely that this could be increased. This issue will also be addressed in the following chapter, in an attempt to structure the model so that a larger number of coordinates may be included in the analysis.

Section	Model	Number of parameters in θ	Number of parameters in θ^*
4.4.1	Independence	2808	513
4.4.2	Correlated random effects	3159	864
4.4.3	Different random error variances per group	3159	864
4.4.4	Different random effects variances per group	3861	216

Table 4.1: A summary of the models fitted in this chapter, with their Section number and the number of parameters in θ and θ^*

Various model covariance structures have been considered in this chapter for the application of the pairwise approach to the cleft-lip and palate data and these are summarised in Table 4.1. The reason for the fourth model having fewer parameters is that the random error variance is coordinate-specific in that model, whereas it is pair-specific in the other models. Therefore in the fourth model it may be averaged across bivariate models while it must remain pair-specific in the others. The independence model was the simplest considered, with each individual having one random intercept for each coordinate, all of which

are uncorrelated. A simpler model could allow each individual a single random intercept, but to average this across all coordinates is rather unrealistic since some coordinates have substantially higher variation around the mean, particularly at three months, than others due to the nature of the cleft defect. The independence model, therefore, is the reference to which the other considered models are compared. This model showed a good mean fit to the data and a reasonable variance fit, although there were some differences in the distributions of the residuals across groups and time points.

Allowing the coordinate-specific random effects to be correlated involved estimating substantially more parameters than in the independence model, but justified the use of the pairwise approach, since the fully joint model could be fitted for independent random effects. The mean fits of the models were very similar, but the estimated correlations between random effects for different coordinates were quite wide ranging, suggesting that these are necessary in the model. Although, in general, it seems unlikely that there should be many high correlations between random effects for coordinates in different dimensions, several of these were reasonably high and in this application that may not be entirely unrealistic.

An issue with the correlated random effects model was the inability to produce residuals because of the covariance matrix, V_i, not being positive definite in some cases. This is evidently not a problem with the model, however; more that in collecting together all of the different covariance estimates from the different pairwise models, we are in no way guaranteed to form a positive-definite covariance matrix. As a model check, the unstandardised residuals were appropriate and so this is not a disadvantage.

The residuals for the model with different random error variances for each group showed a more uniform distribution across coordinates and time than the model with different independent random effects variances, although both were an improvement on the independence model. The model with different random error variances has the disadvantage that the scale parameter giving the ratio of cleft to control random error variance is pair specific rather than coordinate specific. This is not particularly appropriate, especially in the case where one coordinate in a pair has similar variance across groups and the other has substantially different, but since the scale parameter for a coordinate is estimated for each pair in which it is included, it may average out to a reasonable level. In any case, the scale parameter is

not of particular interest in this analysis so it is only an issue if it affects other parameter estimates.

There were several occasions under this model of correlations being estimated as exactly one or negative one. As previously discussed, this may be due to model misspecification and, if so, is problematic because of the necessity of the same specified model for each pairwise combination of coordinates. The next chapter will consider different methods of GPA and the use of principal components analysis to attempt to overcome this problem as it may be related to Procrustes aligned tangent coordinates being too alike for a subset of individuals, while being very different for others.

Whether different random errors or random effects (or both) are assumed, it does seem necessary from Figures 4.25 and 4.19 to allow different variances for the two groups, in particular for coordinates 3, 4 and 5 (the points g, G and 2 around the upper lip in Figure 4.1). These are the points that are most affected by the cleft and so have the highest variance, especially at three months. Under the pairwise approach, however, this may cause problems with over-parameterisation of models for coordinates which do not have such large differences in variances between the groups, such as coordinates 1 and 2 (E and e in Figure 4.1). This does not appear to be the root of the convergence and model-misspecification problems, however, as these appear under the correlated random effects model where there is no stratification by group. The correlations which equal one occur, for example, in the pairwise model with coordinates 3 and 4, which both have high variance.

Another issue is whether compound symmetry across times is a realistic variance structure. This assumes that the correlation between an individual's observations (on a particular coordinate) at any two time points is the same. At first glance this does not appear to be realistic, but there are many factors that go into facial growth. Points may start to change in a different way as a child gets older; for instance a landmark that was moving upwards may start to move outwards instead. Furthermore, the change in tangent coordinates over time is slight and so it is unlikely that there would actually be a substantial difference in the correlations between different time points. Initial exploration of this issue by fitting individual pairwise models and considering the random slope has shown that the variance for this random effect tends to be extremely small compared to the corresponding random intercept.

A related issue is that Figure 4.5 also suggests heterogeneous errors over time for a small selection of coordinates. However, for most coordinates the errors are homogeneous over time and for those that are not, the differences between time points are slight. Therefore, in the interests of parsimony (particularly in light of the various convergence and other computational problems), random intercepts only with homogeneous errors seem a reasonable assumption.

Chapter 5

Further developments of the pairwise approach to shape data

5.1 Introduction

In Chapter 4, various problems were encountered in applying the pairwise approach to shape data, specifically comparing facial shapes between cleft-lip and palate children with controls. There were some failures of model convergence, and a number of models where either the random effects standard errors (SEs) were estimated as exactly zero (none presented in Chapter 4), or the correlations between them were equal or very close to one in absolute value. A set of landmarks was chosen to illustrate the models in Chapter 4 that minimised such problems, but some still occurred. In particular, most of the midline landmarks were omitted, despite some of these, such as the base and tip of the nose, being of particular importance in an analysis such as this. The x-coordinates of these midline landmarks were particularly problematic, causing convergence problems with `lme` and zero estimated random effects SEs with `lmer`. The latter often occurred when the two coordinates included in a pairwise model had very different levels of variation, particularly under the independence model or any models that included correlated random effects. If the responses from one coordinate in a particular bivariate model were much less variable than those from the other coordinate in the model, `lmer` would tend to estimate the random effects for the former as

exactly zero and therefore the variance likewise. Generally, models with independent random effects whose variances were different across groups or where the random error variance was allowed to differ for the two groups, the problem arose less often.

Take, for example, the x-coordinate of the midpoint between the eyes (point 3 in Figure 4.1) and the y-coordinate of the midpoint of the upper lip (point 2 in Figure 4.1). Point x_3 is not very variable across individuals in comparison to point y_2, which is highly variable, amongst the cleft patients at least. Figure 5.1 shows the responses for each of the two tangent coordinates from each individual, stratified by group. Since both axes have the same range, it is clear that there is considerably higher variation amongst individuals for coordinate y_2 than for x_3, especially for the controls, and very little correlation between the two. Table 5.1 shows that if `lmer` was employed, fitting the independence model to the responses from these two coordinates resulted in the random effect SE for coordinate x_3 being estimated as exactly zero, while under the correlated model it was estimated as effectively zero. Using `lme`, the independence model gave a slightly larger value of the SE, but still it is effectively zero when compared with the SE of the random effect for y_2. The correlated model failed to converge using `lme`.

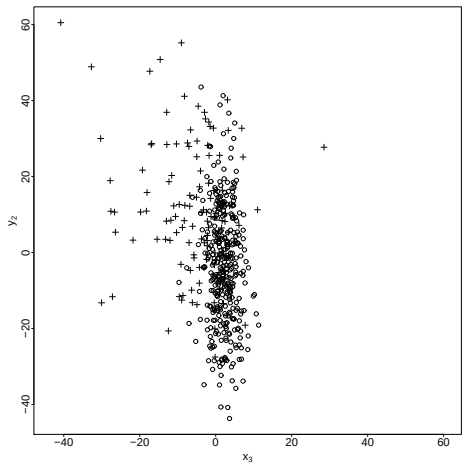

Figure 5.1: Responses for tangent coordinates x_3 (x-coordinate of the midpoint between the eyes) and y_2 (y-coordinate of the midpoint of the upper lip). The cleft subjects are marked by crosses and the controls by circles

The issue is that there clearly are differences in the variance of the responses between the

Model	SE coordinate x_3	SE coordinate y_2
Independence model `lmer`	0.000	11.94
Correlated model `lmer`	5.3×10^{-10}	11.94
Independence model `lme`	6.4×10^{-4}	11.94
Correlated model `lme`	-	-

Table 5.1: Random effects standard error (SE) estimates from a bivariate model containing the responses from coordinates x_3 and y_2, in the form of (4.2). The correlated model under `lme` failed to converge

two coordinates, and while there is less variance for coordinate x_3, it is hard to believe that it is purely measurement error. The standard error for this coordinate is not estimated as zero under every pairwise model and this only seems to occur when there is a large discrepancy in the variance between two coordinates. This results in a bias towards zero of the averaged SE for this coordinate, and this is the case for a handful of other coordinates, often also x-coordinates on the upper midline.

The models with different random effects and random error variances for each group do not seem to suffer such problems, provided the random effects are assumed independent, although for the former there may still be issues with one very small variance estimated for the control group as illustrated in Figure 4.25. It would be preferable to allow correlated random effects when fitting either of these models and the aim of this chapter is to refine the application of the pairwise approach to enable this. The models fitted will generally allow different variances for the two groups and the various sections will consider a subset Procrustes alignment method, analysis of the landmarks of different dimensions separately and the use of principal components analysis.

5.2 Subset GPA and more landmarks

In some other analyses that have been carried out (but not presented), there appeared to be a large difference between the groups in the y-coordinate of the midpoint between the eyes, in that the cleft group were considerably higher in the y direction than the controls. The clefts were also substantially higher in the points at the corners of the mouth, while the tip of the nose was very close in all dimensions to the controls. We believed that there was

likely to be a bigger difference between the groups in the points around the nose and upper lip than those in the upper or lower face and so this seemed an unlikely result. A related issue is that of zero-estimated random effects SEs and that these may arise from individuals' points being forced by the GPA to be very close together. It is possible that the GPA might be giving more weight to the point at the end of the nose (say), either because it protrudes considerably in the z direction, or because there is a large number of points in that area (see Figure 5.2; point 5 represents the end of the nose). If this is the case, then it may be that artificially low levels of variation are occurring in points such as the end of the nose, because the registration is forcing them to be very similar. Also of interest is the potential effect of altering the registration, since the entire analysis is highly dependent upon it, and of adding in extra points and whether this would have an effect on the points already there.

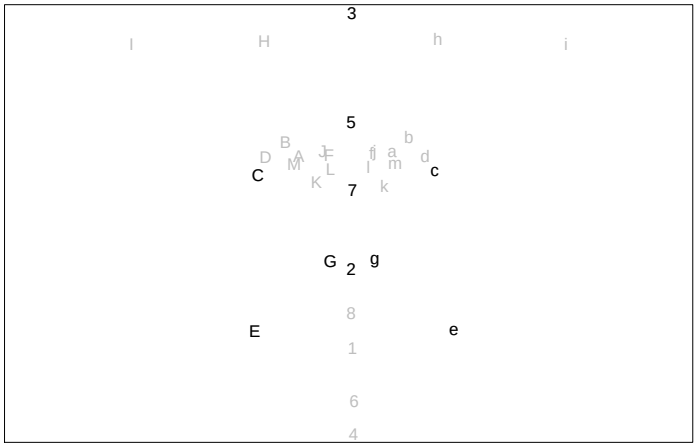

Figure 5.2: Landmarks used to describe facial shape, with lower case letters marking left-side landmarks and upper case marking the equivalent landmarks on the right side of the face. The numbered landmarks are placed along the midline of the face and the black points included in the analysis set following subset GPA.

Therefore we used a subset GPA on the data, matching only on stable points that were reasonably unaffected by the cleft but using the resulting rotation, scale and location parameters to transform the entire configuration and subsequently selecting an entirely different set of points for the analysis. The three-dimensional subset configurations, Y_i, of the full configurations X_i, the former containing only stable points, may be aligned using GPA to form Y_i^P, for $i = 1, \ldots, n$ (where n is the total number of configurations obtained across all N

individuals). For each individual configuration, the means (or centres) of the aligned and unaligned coordinates in the j^{th} dimension ($j = 1, \ldots, m$) may be then calculated as

$$\bar{Y}_{ij}^P = \frac{\sum_{q=1}^k Y_{ijq}^P}{k} \quad \text{and} \quad \bar{Y}_{ij} = \frac{\sum_{q=1}^k Y_{ijq}}{k}$$

respectively, where Y_{ijq} represents the $(j,q)^{th}$ entry of Y_i for the q^{th} landmark of the j^{th} dimension and similarly for Y_i^P. Both means are subtracted from the corresponding configurations and the centroid sizes calculated as (Dryden & Mardia, 1998)

$$S(Y_i) = \sqrt{\sum_{j=1}^m \sum_{q=1}^k (Y_{ijq} - \bar{Y}_{ij})^2}$$

and likewise for $S(Y_i^P)$. The scale is calculated for each individual as the ratio of the size of their aligned to unaligned configurations, $\beta_i = S(Y_i^P)/S(Y_i)$. Ordinary Procrustes analysis (OPA) is carried out to match the scaled but unaligned configurations onto their corresponding aligned points such that

$$D_{OPA}^2(Y_i^P, \beta_i Y_i) = ||\beta_i Y_i - Y_i^P \Gamma_i - \mathbf{1}_k \gamma'||^2$$

is minimised. The unaligned mean vector, $\bar{Y}_i$ (containing the m means across dimensions), is then subtracted from each row, X_{iq}, of X_i, the entire configuration of landmarks, to centre it (on the mean, or centre, of the subset configuration). The estimated rotation matrix, $\hat{\Gamma}_i$, and scale, β_i, are used to rotate and scale the configuration respectively. Finally the mean vector, $\bar{Y}_i^P$, of the original aligned subset configuration is added on to each row, to shift the location of the resulting configuration to the aligned centre of the subset. The resulting aligned configuration is

$$X_i^P = \beta_i (X_i - \bar{Y}_i)\hat{\Gamma}_i + \bar{Y}_i^P.$$

This has the same form as the full (ordinary) Procrustes fit defined in equation (2.1) in Chapter 2. The procedure is carried out for each individual configuration.

The stable landmarks were chosen for this analysis as 4, 6, 1, 8, I, H, h, i in Figure 5.2 and were thought to be those least affected by the cleft defect. Note that this set is completely

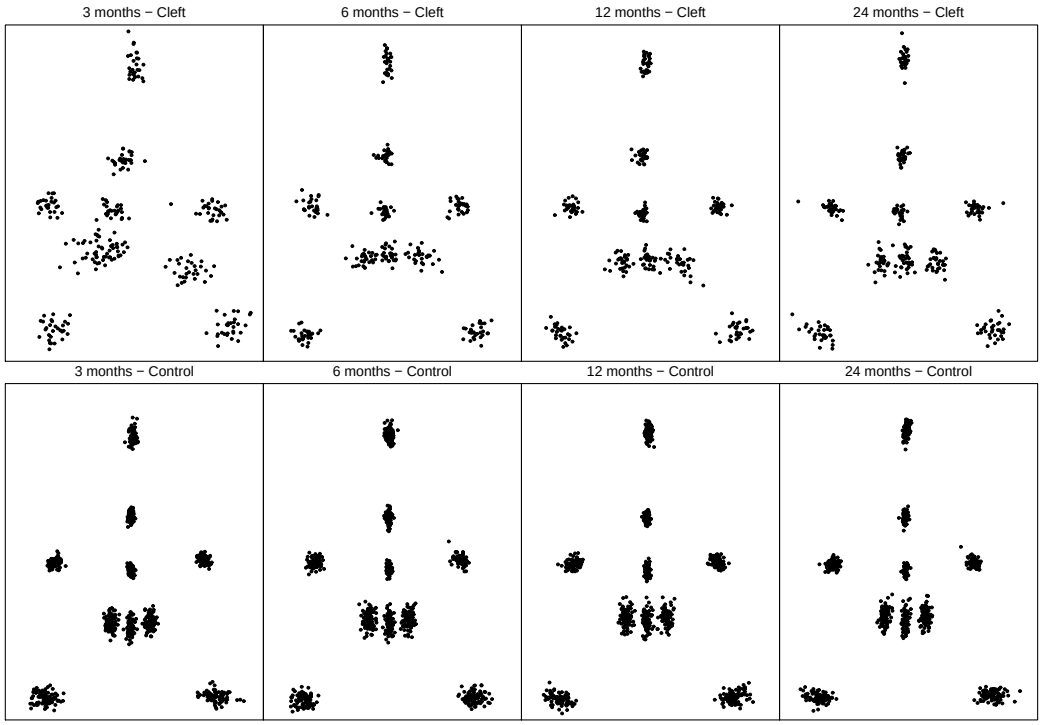

Figure 5.3: Procrustes aligned individual landmarks in the frontal view at 3, 6, 12 and 24 months, with cleft subjects included on the top row and controls on the bottom

separate from the one intended for analysis. Following the subset GPA, 10 landmarks were chosen, which were thought to be more suitable in tracking the effect of the cleft and providing a slightly more detailed description of the face. These are marked in black on Figure 5.2 and the individual aligned configurations are displayed in the frontal view in Figure 5.3, showing once again a greater degree of variation in the cleft group than amongst the controls. The entire set of landmarks (minus the stable landmarks used for the GPA) could not be included because at the stage of forming the J and K matrices in order to calculate the variance of $\hat{\theta}^*$ in (3.4), the computation collapsed as `R` was unable to deal with matrices of that size (26 landmarks in 3 dimensions = 78 coordinates; $78 * 77/2 = 3003$ models; under the model with different random effects variance for each group there are 7 parameters per coordinate within each model plus residual error variance = 15 parameters per model; so each of 3003 J_{pp} matrices are of size (15×15); and the resulting J matrix is of size (45045×45045) and similarly for the K matrix). The set of 10 landmarks highlighted in Figure 5.2 was close to the largest that could be successfully analysed using the pairwise approach, without

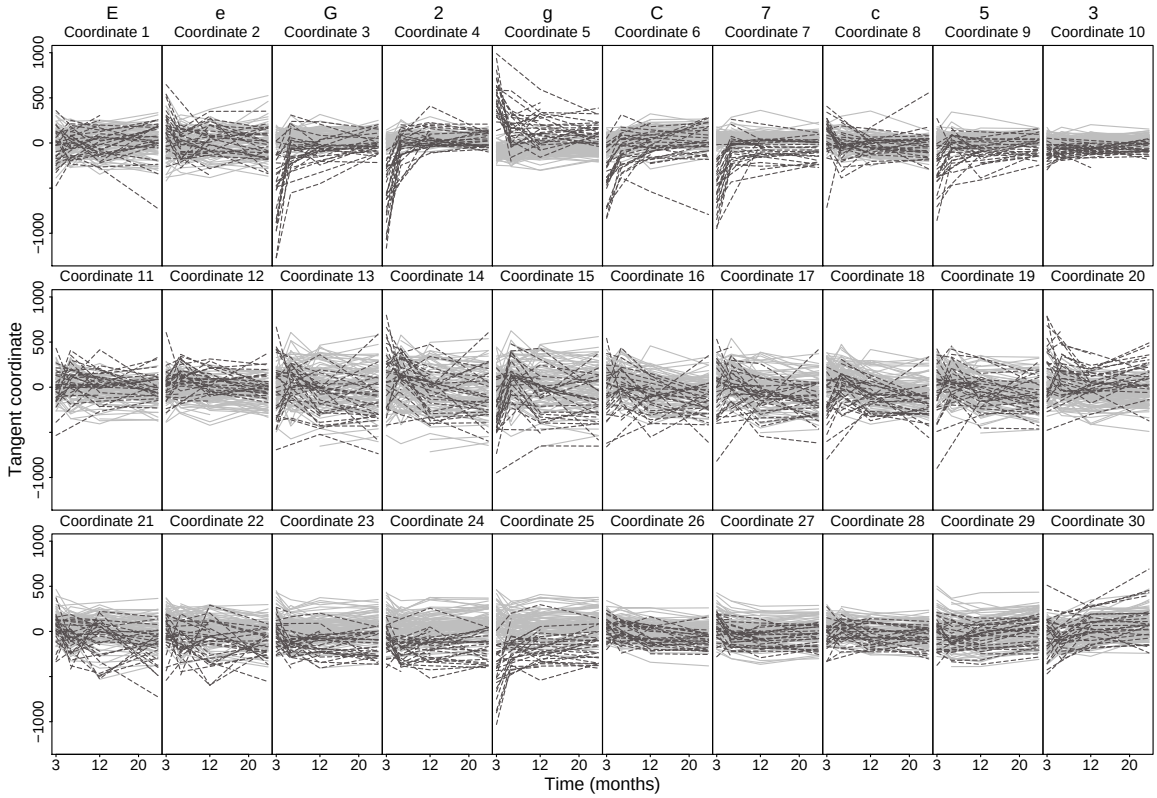

Figure 5.4: Traces over time of the tangent coordinates for each individual, stratified into clefts (dashed black lines) and controls (solid grey lines). The rows represent the x, y and z dimensions and the columns represent the landmarks (with the symbols from Figure 5.2 for identification)

convergence and other problems.

The approximate tangent coordinates (Procrustes residuals) were calculated by subtracting the Procrustes mean of the aligned configuration containing the 10 landmarks marked on Figure 5.2 from the corresponding individual configurations. The tangent coordinate trends are displayed in Figure 5.4, and their group means with 95% confidence intervals in Figure 5.5. For this analysis, the tangent coordinates have been multiplied by 100000, since there seemed to be less chance of zero-estimated random effects SEs than if multiplied by 10000, as in the previous chapter (it made little difference there). There are seven coordinates that are common to the analyses in Chapter 4 and the current analysis. Once again, there is considerable variation between both coordinates and groups and so the model fitted has

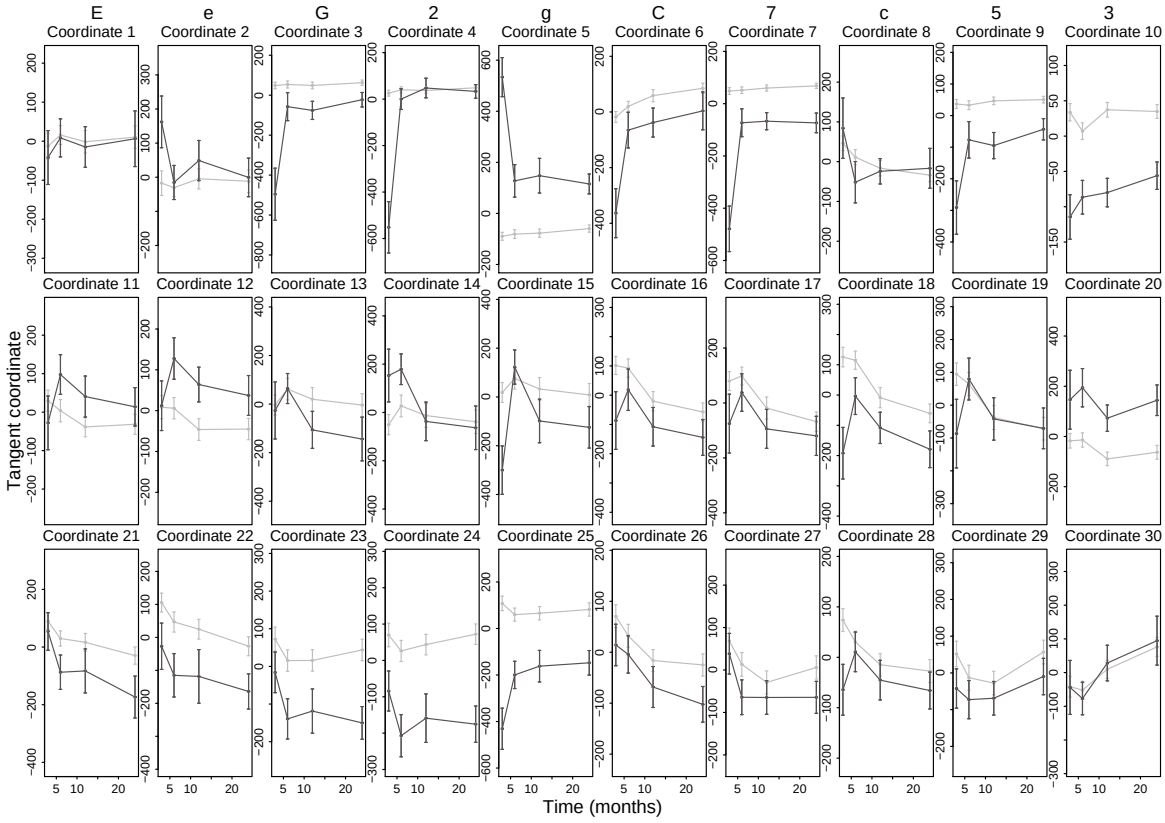

Figure 5.5: Mean trends over time of the tangent coordinates for the clefts (black line) and controls (grey line), with 95% confidence intervals

different random intercepts for each group and coordinate:

$$\mathbf{y}_{ip}(t) = \boldsymbol{\beta}_{0p} + \mathbf{b}_{ipg_i} + \boldsymbol{\beta}_{1p} g_i + \boldsymbol{\beta}_{2p} t + \boldsymbol{\beta}_{3p} g_i \cdot t + \boldsymbol{\beta}_{4p} g_i \cdot S(t) + \boldsymbol{\epsilon}_{ip}(t), \tag{5.1}$$

where, for an individual i who is in the control group (when $g_i = 0$), their random effect vector is distributed as

$$\mathbf{b}_{ip0} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{r0}^2 & \tau_{rs0} \\ \tau_{rs0} & \tau_{s0}^2 \end{pmatrix}\right).$$

For a cleft group subject ($g_i = 1$),

$$\mathbf{b}_{ip1} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{r1}^2 & \tau_{rs1} \\ \tau_{rs1} & \tau_{s1}^2 \end{pmatrix}\right).$$

The random error vector $\boldsymbol{\epsilon}_{ip}(t) = (\epsilon_{ir}(t), \epsilon_{is}(t))$ is distributed as $N(0, \Sigma_p)$, where Σ_p is a diagonal matrix with the variances σ_r^2 and σ_s^2 as its entries. The overall covariance matrix for all outcomes is $V_i = Z_i D_{g_i} Z_i' + \Sigma$, where D_0 and D_1 are defined as the covariance matrices of b_{ip0} and b_{ip1} above and Σ is the diagonal covariance matrix with the variances σ_r^2 for $r = 1, \ldots, km$ as its entries. The matrix V_i has, on the diagonal, compound symmetric blocks of size $(n_i \times n_i)$ which contain $\tau_{rg_i}^2 + \sigma_r^2$ on the diagonal and $\tau_{rg_i}^2$ elsewhere (recall the n_i is the number of observations on individual i over time). On the off-diagonal V_i has $(n_i \times n_i)$ compound symmetric blocks which contain (for every entry) the covariances, τ_{rsg_i}, between the random effects for coordinates r and s. A similar model was introduced in Chapter 4 in (4.6), but there were problems with model convergence when allowing the random intercepts to be correlated and so there they were assumed independent and V_i was block-diagonal. As a result of using the subset GPA on a different set of coordinates to the one analysed, the correlations may now be estimated without convergence failure or random effects SEs being estimated as zero. There were still problems with fitting the model with different random error variances for each group and that is not displayed here.

The 95% bivariate confidence regions resulting from the fitted model are displayed along with reference curves in Figure 5.6. There are some distinct differences between these results and those of the previous chapter, in both variance and mean position of the landmarks (for those that are in both analyses). Generally, there is higher variation in the positions of the cleft landmarks in this analysis, and slightly higher variation for the controls. It seems, intuitively, that having more landmarks in the analysis should cause generally less variation because there is more information on each coordinate (due to being included in more pairwise models). It is likely that the higher variance instead arises from the subset GPA allowing for more variation between individuals in the Procrustes registered positions of those landmarks that are not included in the matching process. Furthermore, the landmarks that are matched upon are reasonably stable and do not get unduly affected by the cleft defect, so it is more realistic to assume that they should be comparable across individuals than should be the landmarks that may be highly varying in one group and very similar in another.

There are also substantial differences in the mean positions of the landmarks that are included in both analyses. The primary difference is that, as expected, there is higher correspondence

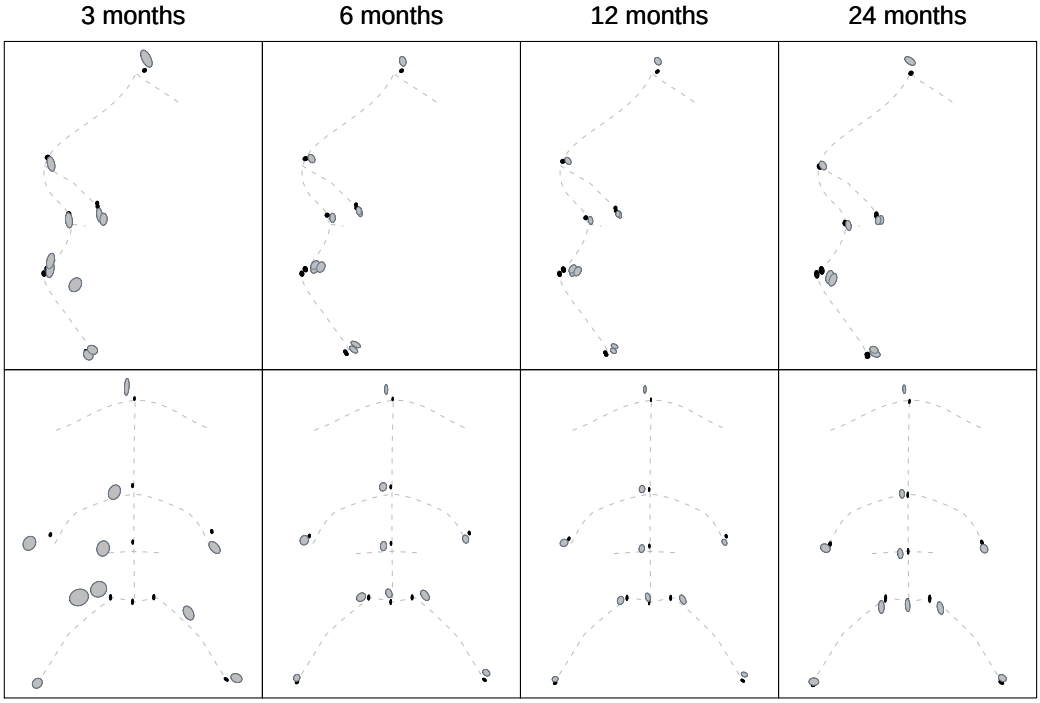

Figure 5.6: Under the model allowing different (correlated) random effects variances for each group after subset GPA, 95% bivariate confidence regions for the cleft (grey) and control (black) groups for each landmark at each time point (upper: profile, lower: frontal view)

between the clefts and controls in some of the more stable landmarks, such as the corners of the lips, and less in the unstable ones around the upper lip and base of the nose. In the previous analyses, the upper lip, whilst being rather flatter than that of the controls, was nonetheless in a similar position relative to the rest of the face. Here the upper lip shows a larger extent of persistent flatness across the entire Cupid's bow, in both the profile and frontal views and, while the corners of the lips are comparable between the groups at 24 months, the upper lip is considerably lower for the clefts than for the controls. Other persistent effects include the midpoint between the eyes remaining higher amongst the cleft group up until 24 months, and residual asymmetry in the positions of the base and end of the nose.

Intuitively, the results from this analysis seem more realistic, in that they show larger differences in the expected areas of the upper lip and base of nose, and fewer differences around the corners of the lips, where the cleft would not be expected to have such a large effect.

However, the discrepancies between the analyses raise concerns about using Procrustes registration. Clearly it is necessary in analysing data such as these to align the landmark configurations, but these analyses show that the results are highly dependent on the choice of landmarks included in the registration process. Careful consideration should be given to the inclusion of landmarks in the Procrustes matching.

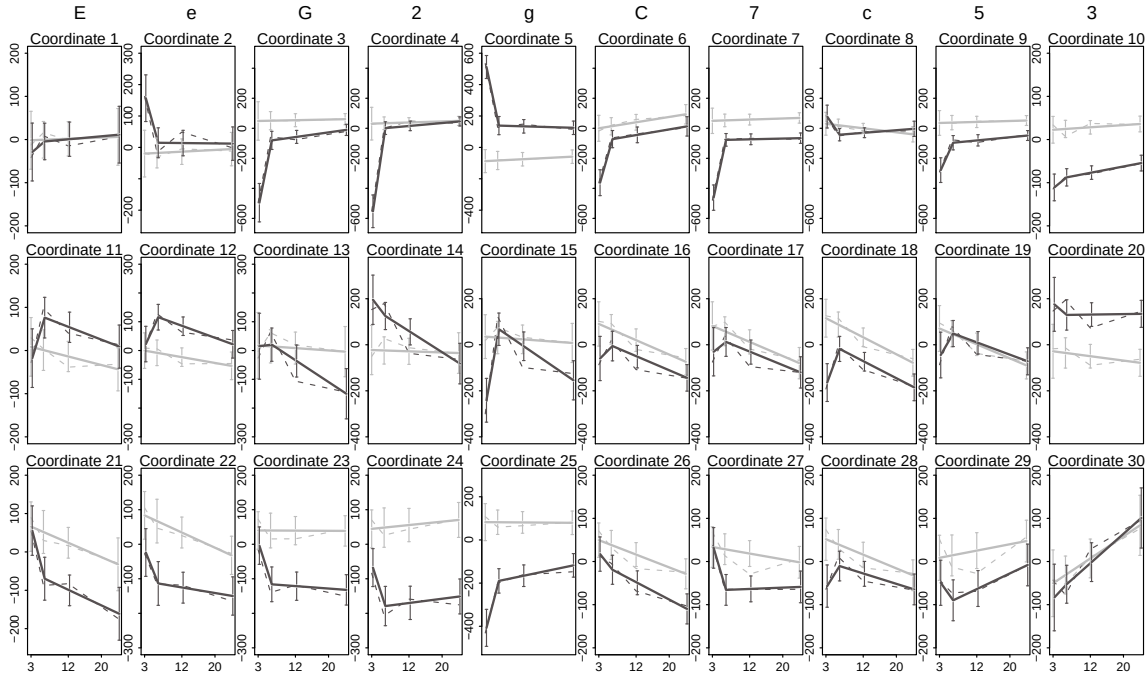

Figure 5.7: Fitted mean trends over time (solid) and 95% confidence intervals from the model with different random effects variances for each group after subset GPA, superimposed onto the empirical means (dashed) for each of the tangent coordinates and stratified into control (grey) and cleft (black) groups.

As a model check the fitted mean tangent coordinate trends over time with confidence intervals are superimposed onto their corresponding empirical means and displayed in Figure 5.7. There are some coordinates for which there is a change at 12 months, particularly for the cleft group, which the linear trend does not fit very well, but for the most part it is a good fit and the linear trend from six months appears to be reasonably appropriate.

Just as in the case of the correlated random effects model in Chapter 4, the covariance matrix $V_i = Z_i D_{g_i} Z_i' + \Sigma$ here is not positive definite for some cases and so the standardised residuals cannot be plotted. Figure 5.8 gives the absolute values of the unstandardised residuals versus

the fitted values for each coordinate. Most of the y- and z-coordinates display reasonably constant variance for the residuals, with only a handful showing small consistent changes across the fitted values. Several of the x-coordinates, however, show a consistent increase or decrease in variance which is an indication that the model is still not adequately explaining the covariance in the data. The large variance of the cleft group at three months is not entirely explained by the model for some coordinates, despite the mean part being modelled adequately using the surgery effect.

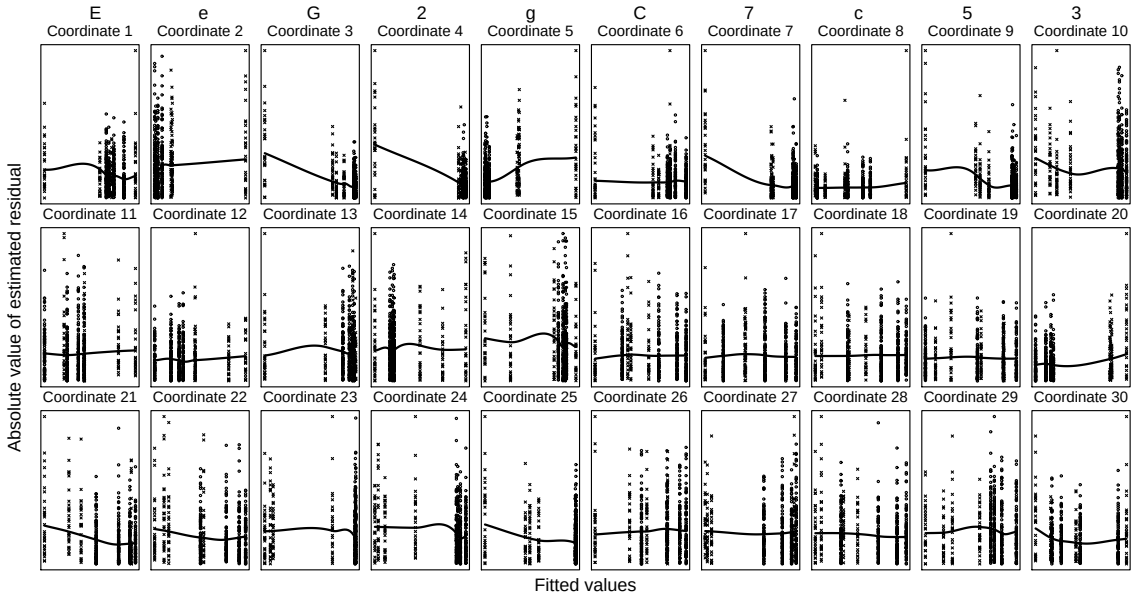

Figure 5.8: Plots of the absolute values of the (unstandardised) residuals versus the fitted values for each coordinate, under the different random effects variances model, with a loess smooth line fitted

The estimated random effects SEs for the cleft and control groups are displayed in Figure 5.9 along with the aggregated estimates from the correlated random effects model (not otherwise presented) that did not stratify for group. As expected, the cleft group random effect SE is generally higher than the corresponding control estimate, and the aggregated estimate from the correlated model is in-between, slightly closer to the control value than the cleft.

The estimated correlations between random intercepts for different coordinates in the same dimension are displayed in Figure 5.10, against the distance between the coordinates in shape space; and also for coordinates in different dimensions (in no particular order) in Figure 5.11.

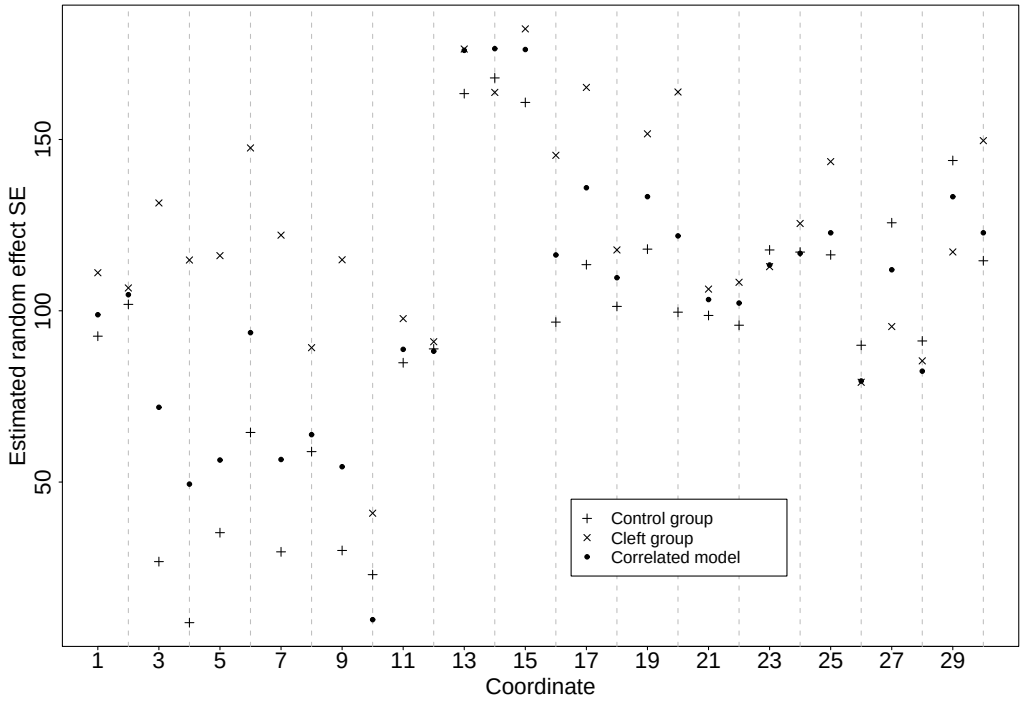

Figure 5.9: Random effects standard error estimates (by coordinate) for the group stratified random intercepts under the current model with a separate estimate for the cleft (crosses) and control (plusses) groups); and corresponding standard error estimates from the model with correlated random effects that are not stratified by group (black filled circles)

Figure 5.10 shows clear relationships between the estimated correlations and distances, for all three dimensions, although there are subsets of each that are equal (or very close) to one. Most of the estimated correlations in Figure 5.11 are between -0.5 and 0.5, but there are several of these too that are equal to or very close to -1 or 1. The majority are for the control group and many involve coordinate 4, which corresponds to the x-coordinate of landmark 2 in Figure 5.2 (and which is why there is some regularity in the spacing of some of these correlations). This corresponds to the very small random effects variance estimated for the control group for this coordinate, displayed in Figure 5.9, and reflects the effect of the pairwise model attempting to estimate two random effects SEs that are very different, as described in Section 5.1. Therefore, while this model appears to fit quite well and give reasonable results, there still may be a problem with model-misspecification or possibly over-parameterisation for some coordinates.

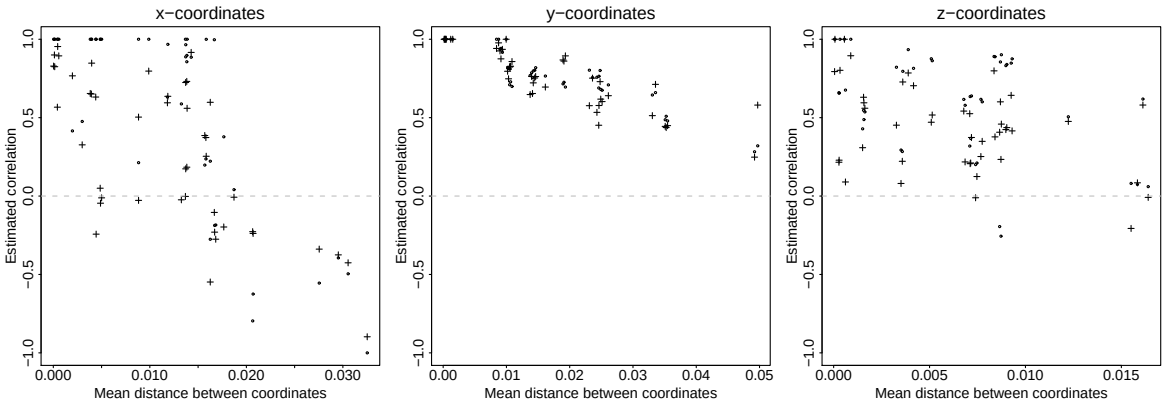

Figure 5.10: Estimated correlations between the random effects for pairs of coordinates, stratified by group and versus the group mean distance between them in shape space, for each of the three dimensions (crosses - cleft group; circles - control group)

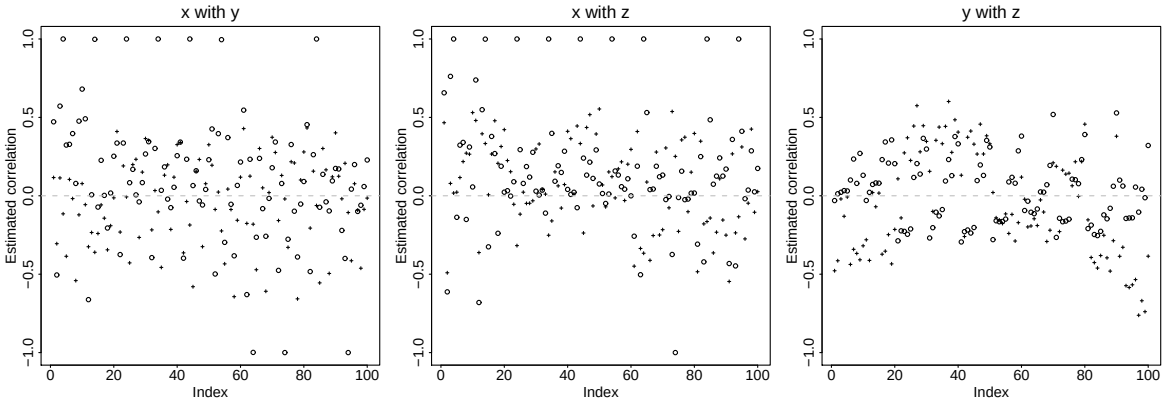

Figure 5.11: Estimated correlations between the random effects for pairs of coordinates in different dimensions and stratified by group (in no particular order; crosses - cleft group; circles - control group)

5.3 Dimension separation

In Chapter 4 a model was fitted that had correlated random effects (but no stratification by group), and assumed an unstructured covariance matrix of the random effects for different coordinates. At the other extreme, the independence model assumed that the random effects for different coordinates were uncorrelated. This latter situation seems particularly unrealistic, since one would expect that there might be some relationship between different coordinates within the same face. However, the correlated model with a general covariance structure requires the estimation of a large number of covariance parameters, some of

which may be very close to zero. In particular, it seems rather unintuitive to expect high correlations between the different dimensions of different landmarks.

Figure 5.12 gives plots of all the pairwise combinations of the 30 coordinates included in the subset GPA of the previous section. It is clear that the coordinates within each dimension are much more highly correlated with each other than those across dimensions. There are few between-dimension combinations that have any notable relationship, particularly when considered relative to the within-dimension combinations, some of which have very strong correlation. It does not seem unreasonable, therefore, to consider a model that allows an unstructured covariance matrix for the random effects for the coordinates within each dimension, whilst assuming independence between dimensions. Using this approach, the pairwise models may be applied to the set of coordinates within each dimension separately and thus substantially more coordinates may be included in the analysis, since the memory-imposed restriction on the number of outcomes here is within dimension.

Once again subset GPA is used to align the images, with the same set of landmarks being used for this matching as in Section 5.2. The pairwise models were fitted to the coordinates of the set of 22 landmarks highlighted in Figure 5.13. It is theoretically possible to include more landmarks (up to around 30), but there were some occurrences of zero-estimated SEs when this was attempted, particularly in the midline landmarks from the lower lip down to the chin. This may be as a result of these coordinates being included in the subset GPA and, to avoid such issues, these landmarks were excluded from the analysis. There is still a small overlap in the subset GPA, since landmarks H and h are included both in the alignment and the analysis. These are neither particularly influential nor variable landmarks, but they are important in both stages since there are few other landmarks available in that area of the face. Since the majority of landmarks are not included in both stages, it is unlikely that this is a problem.

The individual aligned configurations are displayed in the frontal view in Figure 5.14 at each time point and for the subjects of both groups. There is considerable overlap in some landmarks, particularly around the nostrils, and there are different degrees of variation between landmarks and groups.

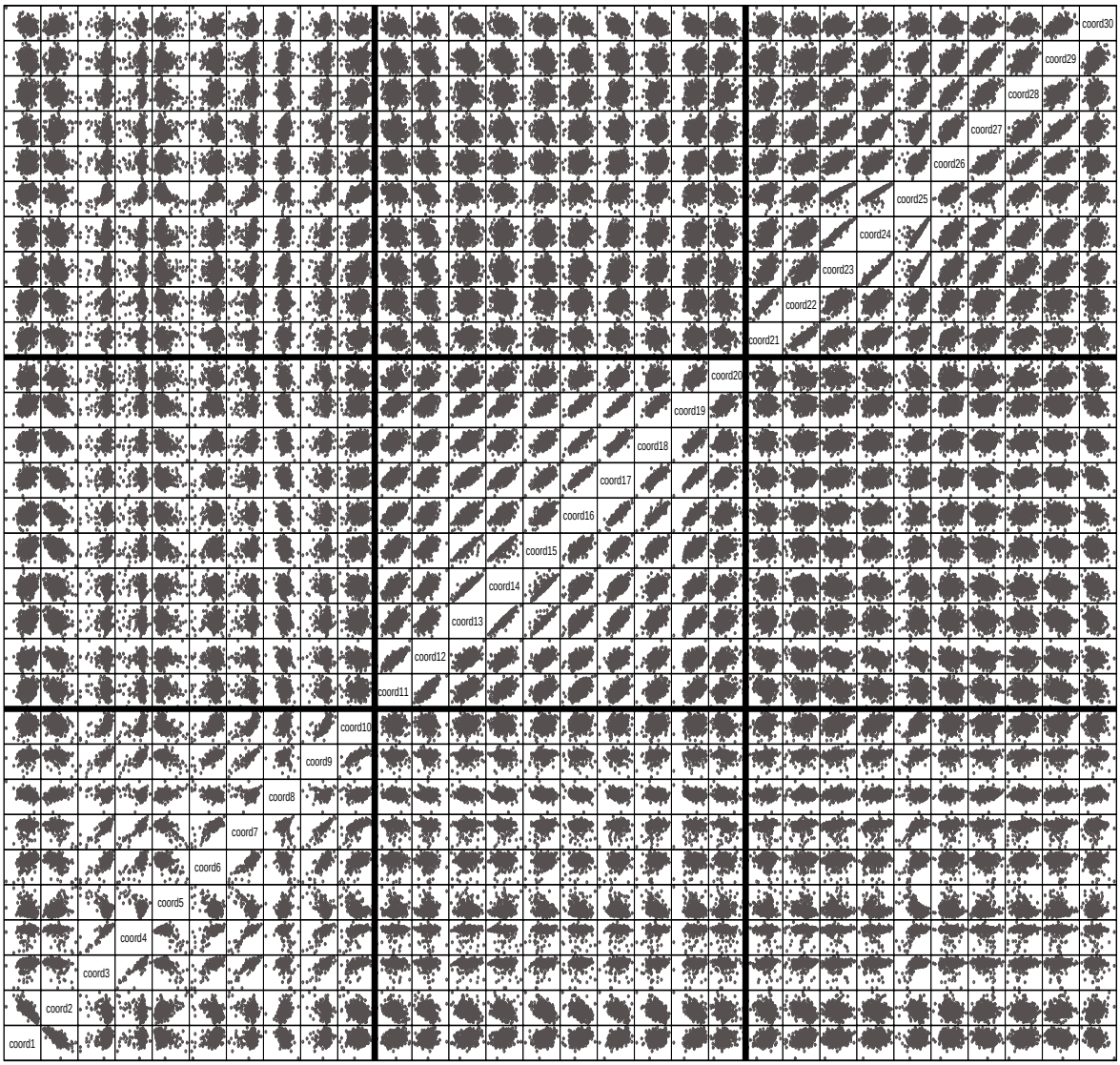

Figure 5.12: Scatterplot of all the pairwise combinations of coordinate responses. Coordinates 1 to 10 are $x-$, 11 to 20 are $y-$ and 21 to 30 are z-coordinates and the thick lines indicate the dimension divisions.

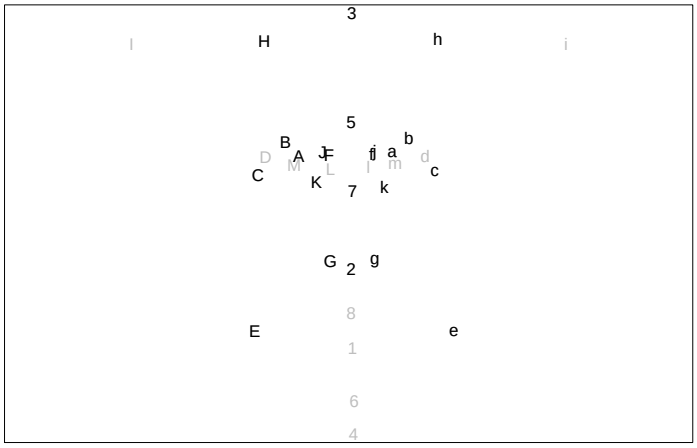

Figure 5.13: Landmarks used to describe facial shape, with lower case letters marking left-side landmarks and upper case marking the equivalent landmarks on the right side of the face. The numbered landmarks are placed along the midline of the face and the black points are those included in the analysis

The model fitted to the dimension separated data has different random intercepts for each group, which are correlated across coordinates. For pair $p = (r, s)$, where $r = 1, \ldots, k-1$ and $s = r + 1, \ldots, k$, and k is the number of coordinates in each dimension (i.e. the number of landmarks), the bivariate model is specified as in (5.1), except the random error variances are pair-specific rather than coordinate-specific because that model with different error variances for each coordinate did not converge. The other difference with the model in (5.1), is that the random intercepts for coordinates in different dimensions are here independent.

The results of the analysis are presented in terms of mean positions with bivariate confidence regions in Figure 5.15. This analysis shows a substantially more detailed view of the differences between the cleft and control groups at each time point, particularly around the nose. The cleft group clearly have a higher bridge of the nose, which persists out to 24 months, and, at 3 months, slightly wider set eyes than the controls on average. At 24 months there remains some asymmetry in the cleft group, and the nostrils are rather wider and flatter, in general, than those of the controls. The upper lip, consistently with previous analyses, is both flatter in the Cupid's bow and protrudes less for the clefts than the controls and this persists out to 24 months.

It is rather more difficult when using so many landmarks to study plots checking the model

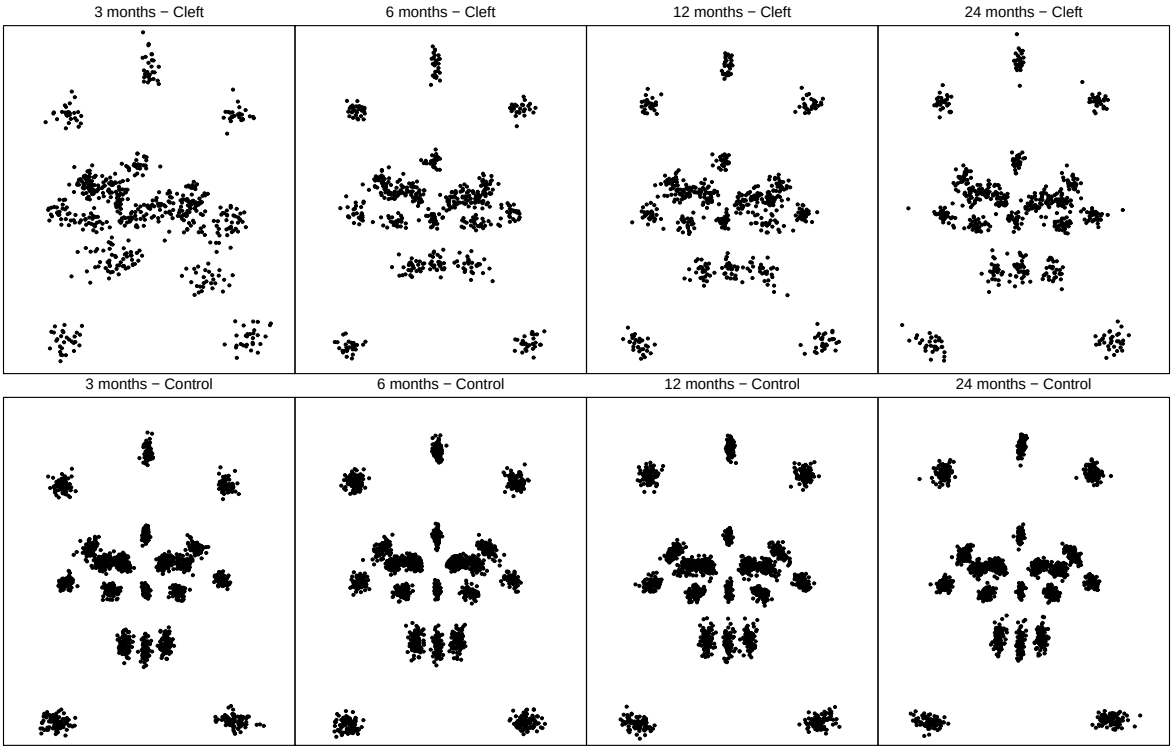

Figure 5.14: Procrustes aligned individual landmarks in the frontal view at 3, 6, 12 and 24 months, with cleft subjects included on the top row and controls on the bottom

fit. It is likely, in any case, that the fitted tangent coordinate trends will follow a similar pattern to those of previous examples, since there is considerable coordinate overlap. When Figure 5.15 is compared to Figure 5.6 in Section 5.2, the positions and, for the most part, the variability of the coincident landmarks are extremely similar so it is likely that the model fits are also comparable. We concentrate instead on the covariance structure.

Figure 5.16 shows the plots of residuals versus fitted values with, again, most of the y- and z-coordinates showing reasonably constant variance but some deviations remaining for the x-coordinates. Figure 5.17 gives the correlations between different coordinates versus their distance apart in shape space. By definition this may only be within-dimension, since the assumption is that coordinates in different dimensions are independent. There are many correlations estimated as equal or close to one in all three plots. This may not be entirely unreasonable in the case of very close landmarks, particularly those around the nose that

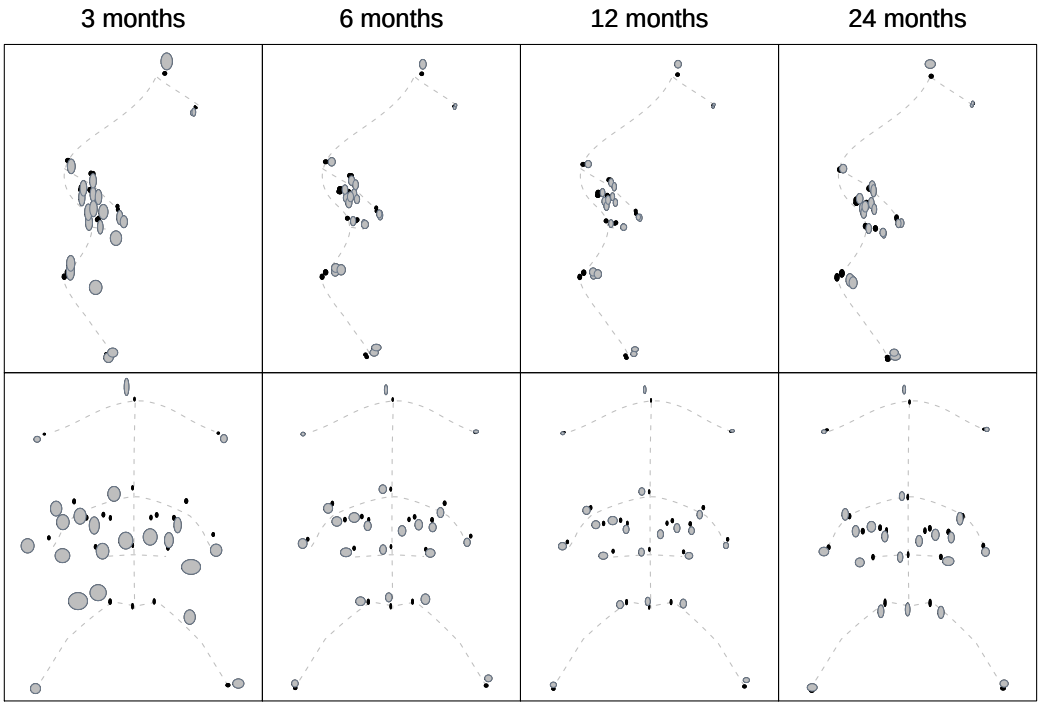

Figure 5.15: Under the dimension separated model allowing different (correlated) random effects variances for each group, 95% bivariate confidence regions for cleft (grey) and control (black) groups for each landmark at each time point (upper: profile, lower: frontal view)

may indeed follow very similar patterns within an individual. It is a source of concern, however, that this issue remains in this analysis.

The inverse relationship of the estimated correlations for the x-coordinates with distance remains, albeit with fewer coordinates that are far apart relative to those that are close. The symmetry of the face is once again apparent, by the strong negative correlations for the furthest apart coordinates of this dimension, these generally being on opposite sides of the face. There is also an inverse relationship between the coordinate correlations and distance for the y-coordinates, although the vast majority of these are strongly positive and only a handful around zero at high distance. It could be argued that there is a similar, though less pronounced, inverse relationship for the z-coordinates though there are still some high correlations at large distances. Although most correlations are positive (reflecting the lack of symmetry in this dimension of the face), three are estimated as close to, or exactly, negative one. These seem likely to be spurious since there are so few and at mid-distance.

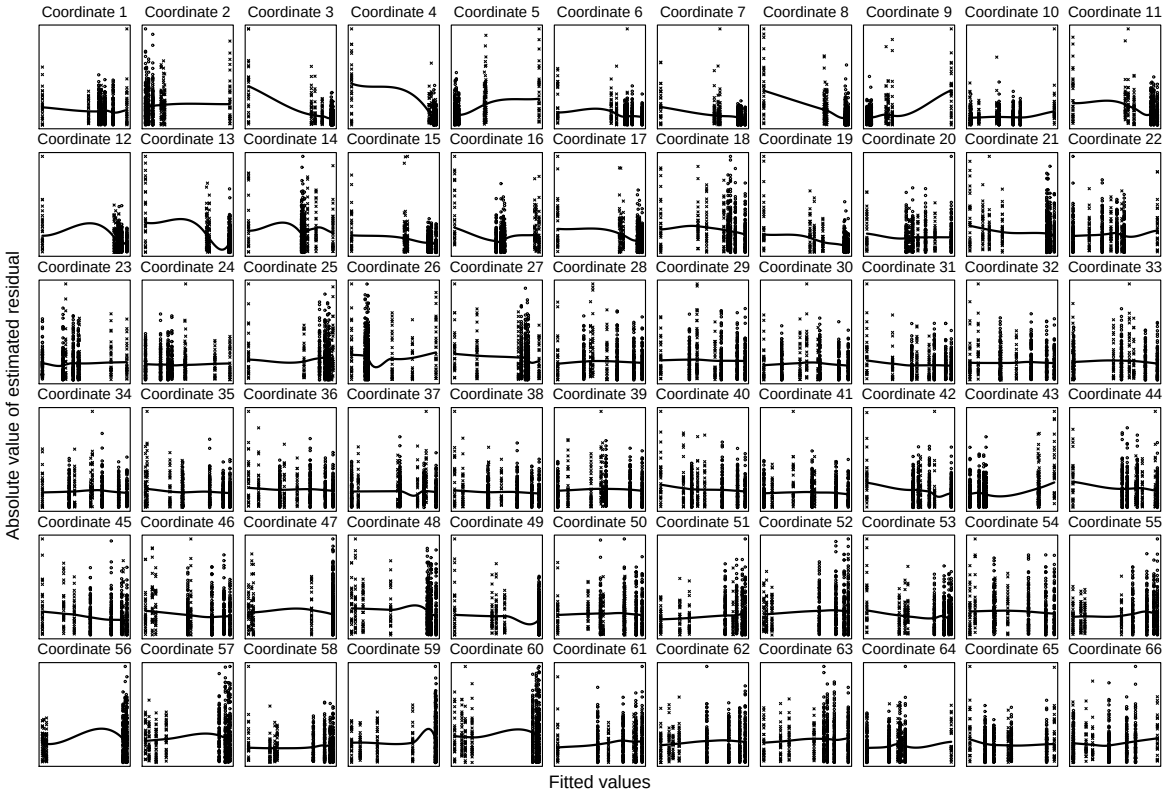

Figure 5.16: Plots of the absolute values of the (unstandardised) residuals versus the fitted values for each coordinate, under the dimension separated model, with a loess smooth line fitted. The x-coordinates are in the first two rows, the y in the middle two and the z in the bottom two rows.

Finally, the random effects SE estimates are displayed in Figure 5.18 for each group, along with the corresponding SEs resulting from the sub-model with no stratification by group for the random effects. There are clear differences between the group random effects SEs for some of the coordinates, whilst others are very similar. The estimated SEs from the model with unstratified random effects are all between their stratified equivalents and generally slightly closer to the control group SEs, reflecting the larger sample size in this group.

5.4 Effect of estimated correlations equal to one

A common issue with the analyses of this and Chapter 4, is that some of the correlations between random effects for different coordinates have been estimated as exactly, or very close

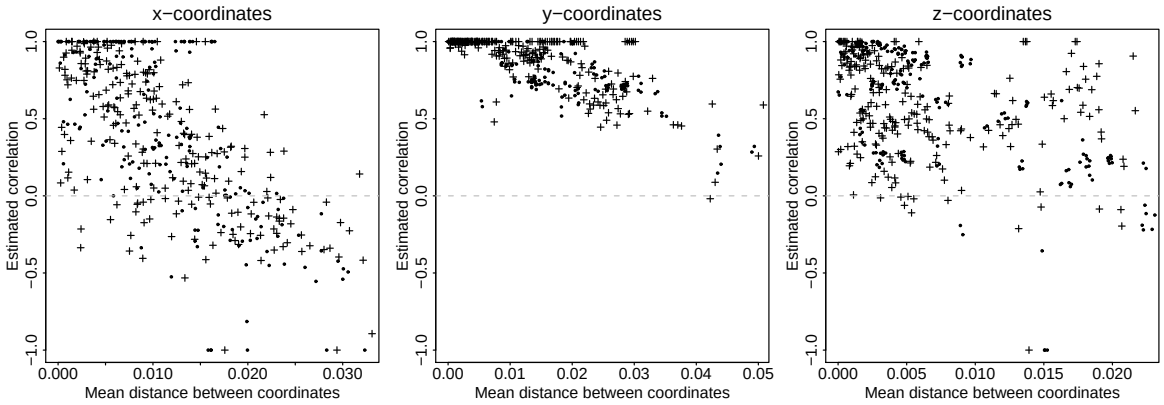

Figure 5.17: Estimated correlations between the random effects for pairs of coordinates, stratified by group and versus the group mean distance between them in shape space, for each of the three dimensions (crosses - cleft group; circles - control group)

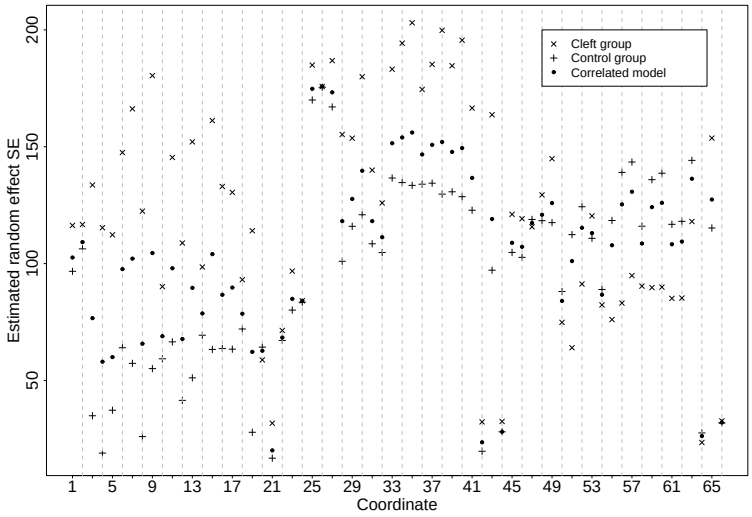

Figure 5.18: Random effects standard error estimates (by coordinate) for the group stratified random intercepts under the current model with a separate estimate for the cleft (crosses) and control (plusses) groups); and unstratified random effects standard error estimates from the model that does not stratify the random effects by group (small, filled circles)

to, one in absolute value. At the same time, apparently valid results are produced and it is unclear how much of an effect this problem may have on such results. There is no explicit interest in the estimated correlations, but the fixed effects estimates and standard errors do depend on them. It is also unclear as to whether the high estimates are arising from underlying truly high correlations, or whether they are spurious.

To these ends a simulation study was carried out. Data on four coordinates from 200 individuals measured at three time points were simulated from the model with different random effects SEs in each of two groups, similar to that described in (5.1), but here there is no surgery effect. Four scenarios were proposed: (1) some of the true standard errors are close to zero and one of the correlations is equal to one; (2) identical to (1) except that none of the correlations are equal to one; (3) all of the true standard errors are away from zero and one of the correlations is equal to one (same correlation structure as (1)); (4) identical to (3) except that none of the correlations are equal to one (same correlation structure as (2)). The true values of the fixed effects are identical under all scenarios and are displayed in Table 5.2.

Parameter	True values (coordinates $1-4$)
Group 0 intercept ("coord$j$")	(-4.0,2.0,-3.0,2.0)
Group 0 slope ("coord$j$.time")	(3.0,0.0,3.5,0.0)
Group difference at baseline ("coord$j$.group")	(0.05,0.05,0.05,0.05)
Group difference in slope ("coord$j$.group.time")	(0.575, 0.575, 0.575, 0.575)

Table 5.2: True values of the fixed effects parameters in the simulations under all scenarios

For each of the four scenarios, the correlation matrices for the random intercepts for individuals from each of the two groups are displayed below with their corresponding SEs.

(1)
$$\begin{pmatrix} 1 & 0.3 & 0.1 & 0.1 \\ 0.3 & 1 & 0.2 & 0.2 \\ 0.1 & 0.2 & 1 & 1 \\ 0.1 & 0.2 & 1 & 1 \end{pmatrix}$$
Group 0 SE: $(5, 5, 2, 1)$
Group 1 SE: $(10, 5, 4, 3)$

(2)
$$\begin{pmatrix} 1 & 0.3 & 0.1 & 0.1 \\ 0.3 & 1 & 0.2 & 0.2 \\ 0.1 & 0.2 & 1 & 0.4 \\ 0.1 & 0.2 & 0.4 & 1 \end{pmatrix}$$
Group 0 SE: $(5, 5, 2, 1)$
Group 1 SE: $(10, 5, 4, 3)$

(3)
$$\begin{pmatrix} 1 & 0.3 & 0.1 & 0.1 \\ 0.3 & 1 & 0.2 & 0.2 \\ 0.1 & 0.2 & 1 & 1 \\ 0.1 & 0.2 & 1 & 1 \end{pmatrix}$$
Group 0 SE: $(3, 5, 8, 7)$
Group 1 SE: $(6, 15, 8, 7)$

(4)
$$\begin{pmatrix} 1 & 0.3 & 0.1 & 0.1 \\ 0.3 & 1 & 0.2 & 0.2 \\ 0.1 & 0.2 & 1 & 0.4 \\ 0.1 & 0.2 & 0.4 & 1 \end{pmatrix}$$
Group 0 SE: $(3, 5, 8, 7)$
Group 1 SE: $(6, 15, 8, 7)$

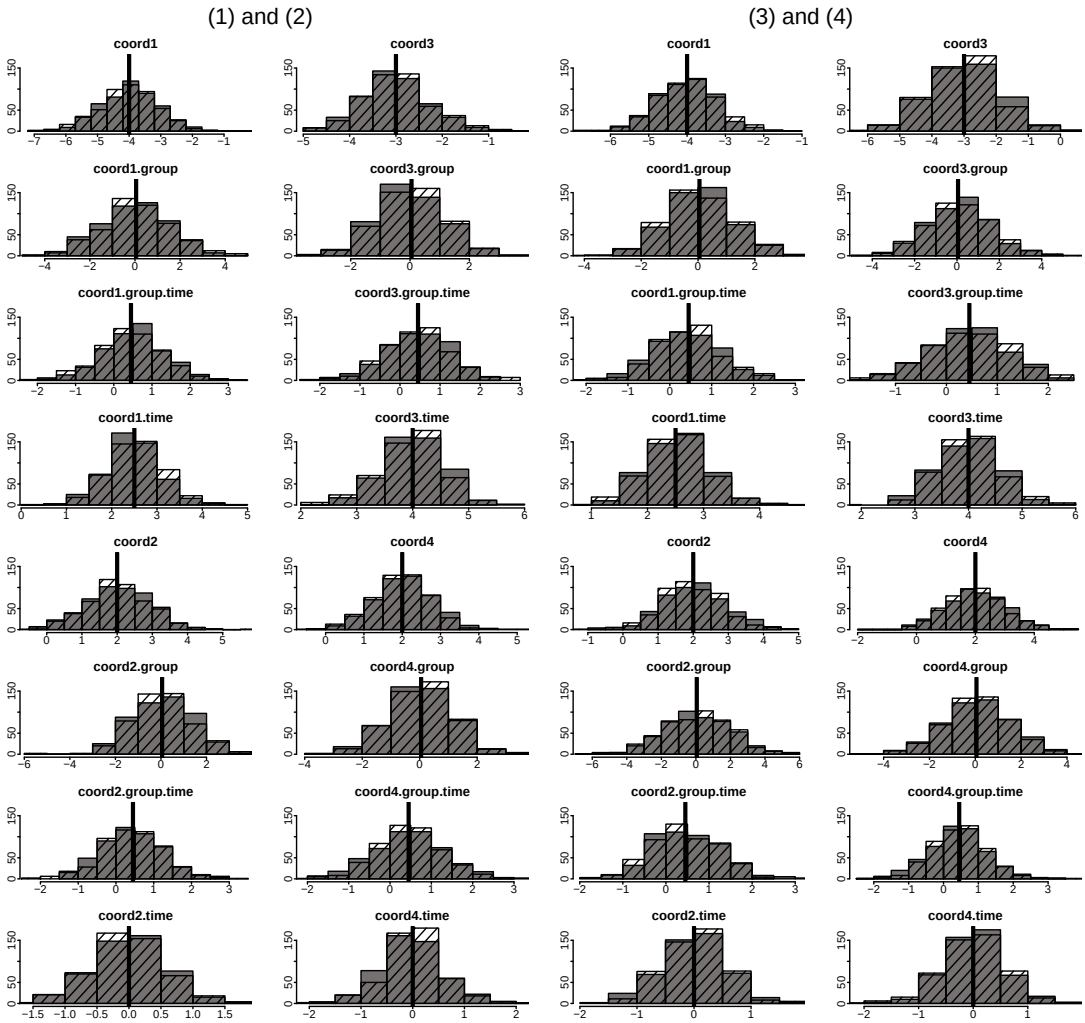

Figure 5.19: Histograms of the fixed effects estimates from 500 simulations of datasets under scenarios (1) (grey shading) and (2) (diagonal lines), left two columns; and (3) (grey shading) and (4) (diagonal lines), right two columns. The thick black lines represent the true parameter values.

Under each scenario, 500 datasets were simulated and the fully joint model as well as the pairwise models fitted to each. Both models give the same estimates for the fixed effects parameters but there are differences for the random effects. Figure 5.19 gives histograms of the fixed effects parameter estimates from all four scenarios, with the estimates from scenarios (1) and (2) superimposed onto one another (left two columns) and likewise for scenarios (3) and (4) (right two columns). There are few differences between the histograms under any of the four scenarios, except that the estimates under (3) and (4) are slightly more variable, but that is to be expected since the random effects SEs are higher. Though

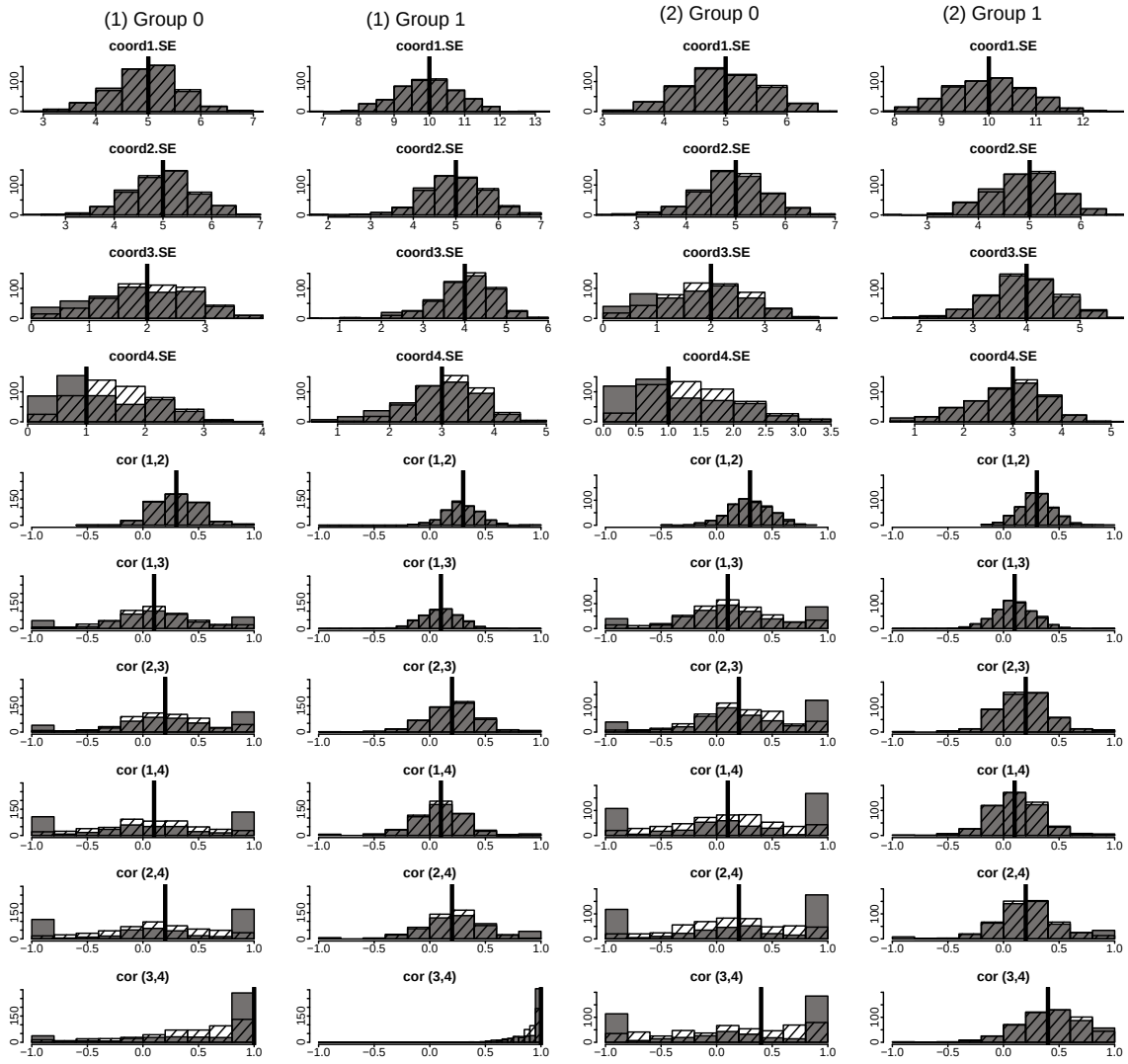

Figure 5.20: Histograms of the random effects standard error and correlation estimates from 500 simulations of datasets under scenarios (1) (left two columns) and (2) (right two columns), under the pairwise approach (grey shading) and the fully joint model (diagonal lines). Each parameter is group-specific and the thick black lines represent the true parameter values. The random effect SE for coordinate j is represented by coord$j$.SE and cor(i,j) is the correlation between the random intercepts for coordinates i and j.

there are some slight differences between the estimates from scenarios (1) and (2), these are minimal and likewise for (3) and (4).

The random effects parameter estimates are displayed in Figure 5.20 for scenarios (1) and (2) and in Figure 5.21 for (3) and (4). In the case of Figure 5.20 the fully joint and pairwise model estimates are superimposed and the left two columns represent the parameter estimates

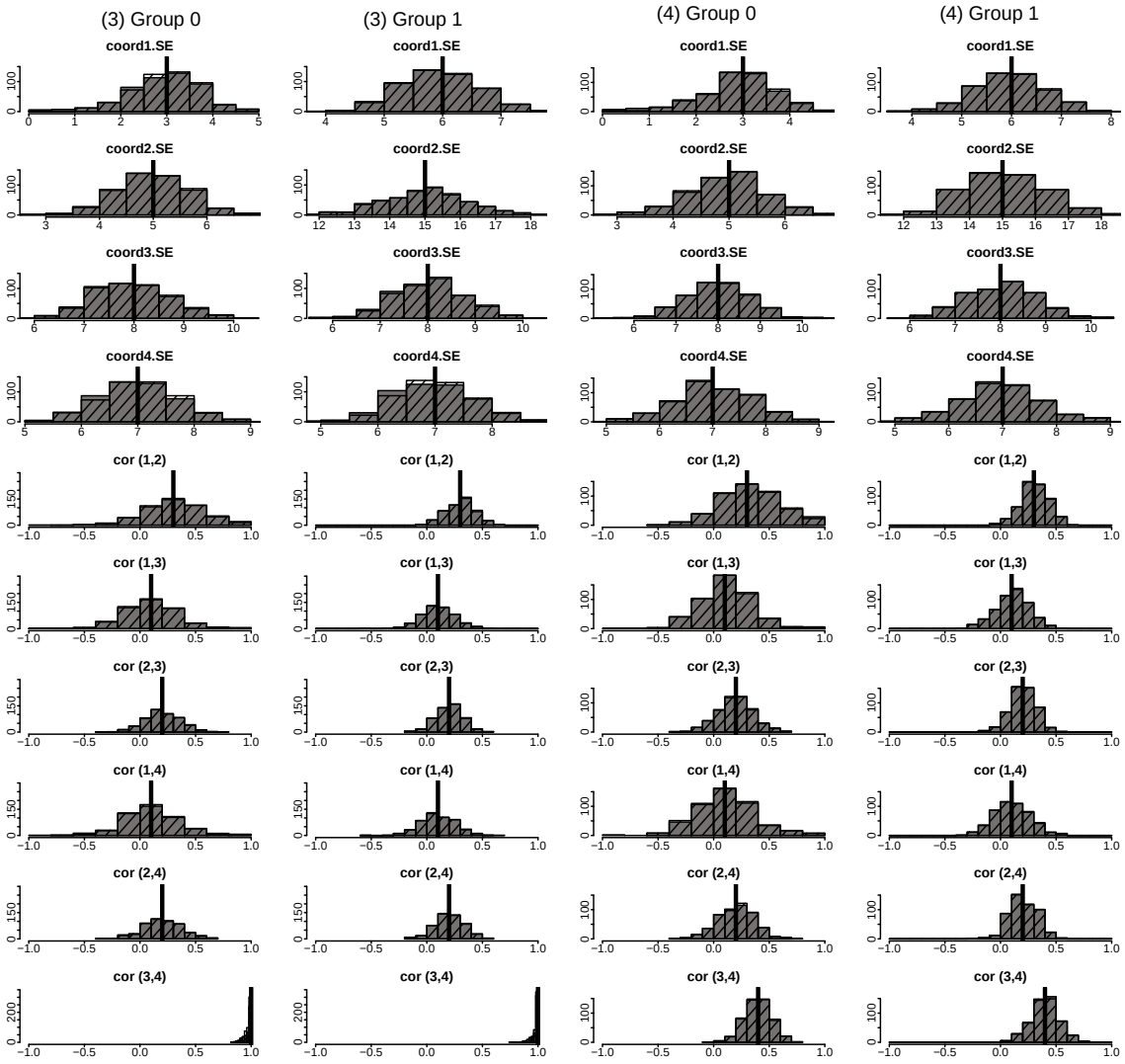

Figure 5.21: Histograms of the random effects standard error and correlation estimates from 500 simulations of datasets under scenarios (3) and (4), under the pairwise approach (grey shading) and the fully joint model (diagonal lines). Each parameter is group-specific and the thick black lines represent the true parameter values. The random effect SE for coordinate j is represented by coord$j$.SE and cor(i,j) is the correlation between the random intercepts for coordinates i and j

under scenario (1), while the right two columns come from (2), and likewise in Figure 5.21 for scenarios (3) and (4).

Figure 5.20 shows pertinent differences between the parameter estimates from the fully joint and pairwise models, under both scenarios (1) and (2). Most of the random effects SEs have very similar distributions, with the exception of coordinate 4, which has a true value of one

for group 0. Under both scenarios, the pairwise model estimates the random effects SE for this coordinate as close to zero (although never exactly zero) more often than the fully joint model and also has a more skewed distribution. In both cases, however, both the median and mean of the distributions are closer to the true value under the pairwise model. The random effects SE for coordinate 4 amongst the group 1 subjects, with a true value of 3, is better estimated and has a more symmetric distribution under both scenarios.

Further discrepancies between the different model estimates and the true values occur for the correlations between random effects. In particular, under both scenarios (1) and (2), the correlations between the random effects for group 0 for coordinates 1 and 3, 2 and 3, 1 and 4, 2 and 4, and 3 and 4, are estimated as either -1 or 1 more often by the pairwise approach than by the fully joint model, when in fact their true values are close to zero. The corresponding correlations for group 1, which have the same true values, are estimated quite well, including the correlation for coordinates 3 and 4. The group 0 correlation for these coordinates, under scenario (1), is correctly estimated as being most likely to be close to 1, but has a long left tail stretching all the way to -1 for both models. The corresponding correlation under scenario (2) is again estimated poorly. The fully joint model, while still showing a large degree of variation in its estimates of the correlations, does do slightly better on average and this may be because it estimates all correlations concurrently, whereas the pairwise model ignores the correlations of each coordinate in a pair with any that are excluded from that pair.

The observations of the previous paragraphs suggest that the correlations may be poorly estimated when the random effects SEs for one or both of the coordinates involved are close to zero, i.e. the boundary of the parameter space, rather than through any aspect of the correlations themselves. This makes sense, since the definition of the correlation has the standard errors as its denominator so if they are particularly small then the correlation will be large, and if the true SEs are close to the parameter space boundary then the model may struggle to estimate them. The correlations between the random effects for different coordinates for group 1 are all estimated well, as are those for both groups between coordinates 1 and 2, which all have SEs that are away from zero. As corroboration of this inference, the distributions of the parameter estimates under scenarios (3) and (4), displayed in Figure 5.21

and where the true values for the SEs are all away from zero, show that the parameters are being well estimated despite one true correlation being equal to one.

These simulations suggest that the correlations being estimated as equal to or very close to one in the cleft data analysis, are not necessarily arising from truly high correlations and may in fact be spurious. However, on the positive side, it also seems that there is little effect on the estimation of other parameters, in particular the fixed effects parameters. Table 5.3 gives the bias of the mean estimate from the 500 simulations for each fixed effects parameter, along with the 95% empirical coverage interval (ECIs; i.e. the 5^{th} and 95^{th} percentiles of the distribution), for each of the four scenarios (recall that these are identical for the pairwise and fully joint approaches). Similarly, the biases and ECIs for the random effects SE and correlation parameter estimates are displayed in Table 5.4 for scenarios (1) and (2) and Table 5.5 for (3) and (4). Although the biases for the fixed effects (the primary interest) do differ slightly and even have different signs across the different scenarios, they are all extremely small in comparison to the true parameter values and so it appears that the differing variance structure and poor estimation of variance parameters for some scenarios does not seem to unduly affect the mean values. With regard to the variance, the empirical coverage intervals show that there are a few fixed effects which are more variable under scenarios (3) and (4) than under (1) and (2), such as 'coord2.group,' 'coord3.group' and 'coord4.group,' although for the 'coord1.group' effect, the opposite is true. It may be, then, that the very small random effects SEs and high correlation knock-on effect is that the fixed effects SEs are slightly anti-conservative, although this is by no means certain since the ECIs for the other effects were very similar across scenarios. This should be taken into account in any formal model comparisons, but the more informal plots of bivariate confidence regions considered up to this point should not be unduly affected.

The biases of the random effects SE and correlation parameters in Tables 5.4 and 5.5 are concordant with the histograms of Figures 5.20 and 5.21, with little bias occurring for the SEs that are away from zero and more for those close to zero and the corresponding correlations. Since the fixed effects and their SEs are of primary interest in terms of comparing between groups and any potential prediction, the random effects parameter estimates are only relevant in the way they affect the mean parameters.

Parameter	Scenario (1)		Scenario (2)	
	Bias	95% ECI	Bias	95% ECI
coord1	0.003	(-5.52, -2.57)	-0.040	(-5.61, -2.61)
coord1.group	-0.058	(-2.51, 2.40)	0.033	(-2.56, 2.69)
coord1.group.time	0.072	(-0.86, 1.86)	-0.009	(-1.10, 1.82)
coord1.time	-0.039	(1.49, 3.46)	0.027	(1.57, 3.54)
coord2	0.033	(0.50, 3.45)	-0.001	(0.45, 3.43)
coord2.group	0.016	(-1.99, 2.12)	0.020	(-1.98, 2.30)
coord2.group.time	-0.043	(-0.90, 1.75)	-0.013	(-1.02, 1.85)
coord2.time	0.002	(-0.99, 0.94)	0.006	(-0.97, 0.99)
coord3	-0.010	(-4.16, -1.70)	0.035	(-4.11, -1.63)
coord3.group	-0.050	(-1.71, 1.91)	0.008	(-1.83, 1.92)
coord3.group.time	-0.010	(-0.99, 1.73)	0.004	(-0.85, 1.88)
coord3.time	0.006	(3.07, 4.90)	-0.039	(2.92, 4.84)
coord4	0.049	(0.72, 3.30)	-0.045	(0.79, 3.13)
coord4.group	-0.064	(-1.88, 1.69)	-0.011	(-1.82, 1.76)
coord4.group.time	0.027	(-0.97, 1.98)	-0.006	(-0.78, 1.80)
coord4.time	-0.045	(-1.04, 1.03)	0.026	(-0.94, 0.91)
	Scenario (3)		Scenario (4)	
	Bias	95% ECI	Bias	95% ECI
coord1	0.011	(-5.30, -2.71)	0.020	(-5.38, -2.72)
coord1.group	-0.038	(-1.97, 2.15)	-0.093	(-2.13, 1.98)
coord1.group.time	0.035	(-0.85, 1.82)	0.040	(-0.99, 1.75)
coord1.time	0.016	(1.60, 3.46)	-0.014	(1.61, 3.41)
coord2	-0.002	(0.62, 3.35)	0.079	(0.68, 3.50)
coord2.group	0.049	(-3.10, 3.37)	-0.204	(-3.27, 2.88)
coord2.group.time	-0.022	(-0.92, 1.65)	0.003	(-0.82, 1.72)
coord2.time	0.031	(-0.89, 0.95)	-0.027	(-0.92, 0.85)
coord3	0.032	(-4.93, -1.13)	0.035	(-4.71, -1.37)
coord3.group	-0.022	(-2.63, 2.55)	-0.038	(-2.52, 2.44)
coord3.group.time	0.031	(-0.81, 1.69)	0.018	(-0.78, 1.79)
coord3.time	-0.010	(3.07, 4.92)	-0.020	(3.10, 4.86)
coord4	0.074	(0.37, 3.70)	0.064	(0.28, 3.70)
coord4.group	-0.031	(-2.34, 2.39)	-0.067	(-2.34, 2.37)
coord4.group.time	-0.010	(-0.91, 1.76)	0.019	(-0.94, 1.80)
coord4.time	-0.006	(-0.89, 0.86)	-0.029	(-0.96, 0.96)

Table 5.3: Biases for the fixed effects means and 95% empirical coverage intervals from 500 simulated datasets for all four scenarios

		Scenario (1)		Scenario (2)	
Parameter	Model	Bias	95% ECI	Bias	95% ECI
coord1.SE.group0	Pairwise	-0.050	(3.79, 5.94)	-0.041	(3.89, 6.05)
	Fully joint	-0.029	(3.86, 5.94)	-0.027	(3.94, 6.07)
coord1.SE.group1	Pairwise	-0.058	(8.37, 11.43)	-0.001	(8.59, 11.36)
	Fully joint	-0.050	(8.39, 11.44)	0.004	(8.62, 11.37)
coord2.SE.group0	Pairwise	-0.040	(3.76, 6.10)	-0.074	(3.78, 6.04)
	Fully joint	-0.003	(3.86, 6.12)	-0.054	(3.86, 6.06)
coord2.SE.group1	Pairwise	-0.030	(3.83, 6.06)	-0.095	(3.81, 5.93)
	Fully joint	-0.011	(3.86, 6.08)	-0.086	(3.80, 5.96)
coord3.SE.group0	Pairwise	-0.137	(0.31, 3.25)	-0.226	(0.33, 3.16)
	Fully joint	0.058	(0.68, 3.27)	-0.037	(0.59, 3.19)
coord3.SE.group1	Pairwise	-0.023	(2.55, 5.00)	-0.058	(2.73, 5.03)
	Fully joint	0.044	(2.67, 5.06)	-0.039	(2.84, 5.04)
coord4.SE.group0	Pairwise	0.266	(0.30, 2.69)	0.151	(0.25, 2.57)
	Fully joint	0.552	(0.51, 2.74)	0.424	(0.45, 2.60)
coord4.SE.group1	Pairwise	-0.042	(1.55, 4.02)	-0.136	(1.37, 4.00)
	Fully joint	0.147	(2.10, 4.11)	-0.087	(1.59, 4.01)
coord1.coord2.cov.group0	Pairwise	0.000	(-0.02, 0.63)	-0.003	(-0.02, 0.6)
	Fully joint	-0.005	(-0.02, 0.61)	-0.007	(-0.02, 0.59)
coord1.coord3.cov.group0	Pairwise	0.009	(-1.00, 1.00)	0.064	(-1.00, 1.00)
	Fully joint	-0.003	(-0.54, 0.74)	0.021	(-0.66, 0.87)
coord1.coord4.cov.group0	Pairwise	-0.025	(-1.00, 1.00)	0.041	(-1.00, 1.00)
	Fully joint	-0.042	(-0.76, 0.82)	-0.002	(-0.76, 0.89)
coord2.coord3.cov.group0	Pairwise	0.053	(-1.00, 1.00)	0.073	(-1.00, 1.00)
	Fully joint	0.006	(-0.41, 0.94)	0.007	(-0.46, 0.89)
coord2.coord4.cov.group0	Pairwise	-0.047	(-1.00, 1.00)	-0.037	(-1.00, 1.00)
	Fully joint	-0.086	(-0.74, 0.85)	-0.066	(-0.74, 0.89)
coord3.coord4.cov.group0	Pairwise	-0.424	(-1.00, 1.00)	-0.224	(-1.00, 1.00)
	Fully joint	-0.582	(-0.58, 0.98)	-0.247	(-0.85, 0.96)
coord1.coord2.cov.group1	Pairwise	-0.011	(0.04, 0.56)	-0.004	(0.02, 0.55)
	Fully joint	-0.010	(0.04, 0.56)	-0.005	(0.02, 0.55)
coord1.coord3.cov.group1	Pairwise	-0.004	(-0.19, 0.37)	0.000	(-0.21, 0.39)
	Fully joint	-0.002	(-0.18, 0.36)	-0.002	(-0.22, 0.39)
coord1.coord4.cov.group1	Pairwise	-0.002	(-0.32, 0.49)	0.020	(-0.24, 0.51)
	Fully joint	-0.009	(-0.25, 0.4)	0.013	(-0.24, 0.47)
coord2.coord3.cov.group1	Pairwise	0.013	(-0.17, 0.59)	-0.016	(-0.18, 0.58)
	Fully joint	0.003	(-0.17, 0.56)	-0.020	(-0.17, 0.54)
coord2.coord4.cov.group1	Pairwise	0.057	(-0.27, 1.00)	0.041	(-0.21, 1.00)
	Fully joint	0.009	(-0.20, 0.61)	0.017	(-0.20, 0.66)
coord3.coord4.cov.group1	Pairwise	-0.085	(0.63, 1.00)	0.032	(-0.04, 1.00)
	Fully joint	-0.119	(0.62, 1.00)	0.020	(-0.04, 0.88)
residual error SE	Pairwise	-0.011	(7.75, 8.22)	-0.006	(7.77, 8.21)
	Fully joint	-0.035	(7.74, 8.18)	-0.017	(7.76, 8.20)

Table 5.4: Biases for the random effects standard errors and correlations with 95% empirical coverage intervals (ECI) from 500 simulated datasets under scenarios 1 and 2

		Scenario (3)		Scenario (4)	
Parameter	Model	Bias	95% ECI	Bias	95% ECI
coord1.SE.group0	Pairwise	-0.079	(1.52, 4.03)	-0.116	(1.31, 4.09)
	Fully joint	0.006	(1.74, 4.05)	-0.080	(1.54, 4.08)
coord1.SE.group1	Pairwise	-0.071	(4.85, 7.01)	-0.033	(4.82, 7.14)
	Fully joint	-0.052	(4.87, 7.01)	-0.030	(4.83, 7.10)
coord2.SE.group0	Pairwise	-0.027	(3.85, 6.02)	-0.039	(3.83, 6.08)
	Fully joint	-0.002	(3.88, 6.05)	-0.039	(3.84, 6.07)
coord2.SE.group1	Pairwise	-0.021	(13.09, 16.92)	0.026	(13.17, 17.08)
	Fully joint	-0.014	(13.1, 16.92)	0.025	(13.17, 17.08)
coord3.SE.group0	Pairwise	-0.070	(6.75, 9.15)	-0.016	(6.70, 9.27)
	Fully joint	-0.030	(6.78, 9.18)	-0.015	(6.70, 9.28)
coord3.SE.group1	Pairwise	0.026	(6.80, 9.27)	0.003	(6.70, 9.25)
	Fully joint	0.066	(6.92, 9.31)	0.004	(6.69, 9.24)
coord4.SE.group0	Pairwise	-0.021	(5.87, 8.13)	0.036	(5.84, 8.22)
	Fully joint	0.038	(5.94, 8.14)	0.036	(5.83, 8.23)
coord4.SE.group1	Pairwise	-0.053	(5.89, 8.12)	-0.032	(5.70, 8.29)
	Fully joint	0.009	(5.99, 8.18)	-0.033	(5.69, 8.28)
coord1.coord2.cov.group0	Pairwise	0.015	(-0.11, 0.78)	0.040	(-0.07, 0.84)
	Fully joint	0.000	(-0.11, 0.72)	0.035	(-0.08, 0.79)
coord1.coord3.cov.group0	Pairwise	-0.002	(-0.29, 0.51)	0.014	(-0.25, 0.52)
	Fully joint	-0.006	(-0.27, 0.48)	0.012	(-0.25, 0.49)
coord1.coord4.cov.group0	Pairwise	0.003	(-0.29, 0.55)	0.008	(-0.31, 0.59)
	Fully joint	0.001	(-0.26, 0.52)	0.012	(-0.31, 0.60)
coord2.coord3.cov.group0	Pairwise	0.002	(-0.06, 0.47)	-0.007	(-0.08, 0.45)
	Fully joint	-0.001	(-0.06, 0.46)	-0.007	(-0.08, 0.44)
coord2.coord4.cov.group0	Pairwise	0.005	(-0.10, 0.5)	0.000	(-0.08, 0.45)
	Fully joint	0.000	(-0.09, 0.47)	0.000	(-0.08, 0.45)
coord3.coord4.cov.group0	Pairwise	-0.025	(0.90, 1.00)	-0.002	(0.20, 0.59)
	Fully joint	-0.033	(0.90, 1.00)	-0.002	(0.20, 0.59)
coord1.coord2.cov.group1	Pairwise	0.003	(0.07, 0.52)	0.006	(0.09, 0.51)
	Fully joint	0.002	(0.07, 0.51)	0.006	(0.09, 0.51)
coord1.coord3.cov.group1	Pairwise	0.000	(-0.14, 0.33)	0.010	(-0.15, 0.34)
	Fully joint	-0.001	(-0.14, 0.32)	0.010	(-0.15, 0.34)
coord1.coord4.cov.group1	Pairwise	0.001	(-0.15, 0.36)	0.004	(-0.14, 0.37)
	Fully joint	-0.001	(-0.14, 0.35)	0.004	(-0.14, 0.38)
coord2.coord3.cov.group1	Pairwise	0.004	(0.00, 0.39)	0.002	(0.01, 0.38)
	Fully joint	0.002	(0.00, 0.39)	0.002	(0.01, 0.38)
coord2.coord4.cov.group1	Pairwise	0.005	(0.00, 0.42)	0.004	(0.02, 0.40)
	Fully joint	0.002	(0.00, 0.41)	0.004	(0.02, 0.40)
coord3.coord4.cov.group1	Pairwise	-0.021	(0.91, 1.00)	-0.002	(0.19, 0.61)
	Fully joint	-0.025	(0.91, 1.00)	-0.002	(0.19, 0.60)
residual error SE	Pairwise	-0.012	(7.77, 8.23)	0.002	(7.78, 8.23)
	Fully joint	-0.041	(7.75, 8.18)	0.003	(7.78, 8.23)

Table 5.5: Biases for the random effects standard errors and correlations with 95% empirical coverage intervals (ECI) from 500 simulated datasets under scenarios 3 and 4

5.5 Principal components analysis

One way of both reducing the dimension and potentially avoiding the problem of correlations being spuriously estimated as equal to one, is to analyse the principal components scores of the data. This also makes sense because many of the landmarks on the face are likely to be very highly correlated and so may be explained by a small number of principal components (PCs). In taking the first p of P scores that explain the majority of the variance, the dimension should be substantially reduced, allowing more landmarks to be included in the analysis, and we may avoid analysing outcomes for which there is an extremely small amount of variance, leading to spurious highly estimated correlations.

To avoid any of the problems arising by using the same landmarks for the alignment and analysis, subset GPA was employed using the landmarks 4, 6, 1, 8, I, H, h, i in Figure 5.2. The Procrustes mean of the resulting aligned configurations of the 22 landmarks displayed in black in Figure 5.13 was subtracted from the individual configurations to obtain the tangent coordinates. This is necessary since tangent space is linearised and hence within it multivariate analyses (such as principal components analysis) may be carried out (Dryden & Mardia, 1998).

The vectors of tangent coordinates are defined as v_i for individuals $i = 1, \ldots, n$, while the mean tangent coordinate vector is $\bar{v} = \frac{1}{n}\sum v_i$. The sample covariance matrix of the v_i may be calculated as

$$S_v = \frac{1}{n}\sum_{i=1}^{n}(v_i - \bar{v})(v_i - \bar{v})'$$

and its orthonormal eigenvectors denoted as γ_j with corresponding eigenvalues

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_P,$$

where $P = 3k - 7$, is the dimension of the shape space (for k the number of landmarks each in three dimensions) and as such the number of non-zero eigenvalues (and eigenvectors). The standardised PC scores, c_{ij}, for individual i and the j^{th} PC, γ_j, are calculated as (Dryden

& Mardia, 1998)

$$c_{ij} = \gamma_j'(v_i - \bar{v})/\lambda_j^{1/2}, \tag{5.2}$$

and these are used as the responses in the forthcoming model.

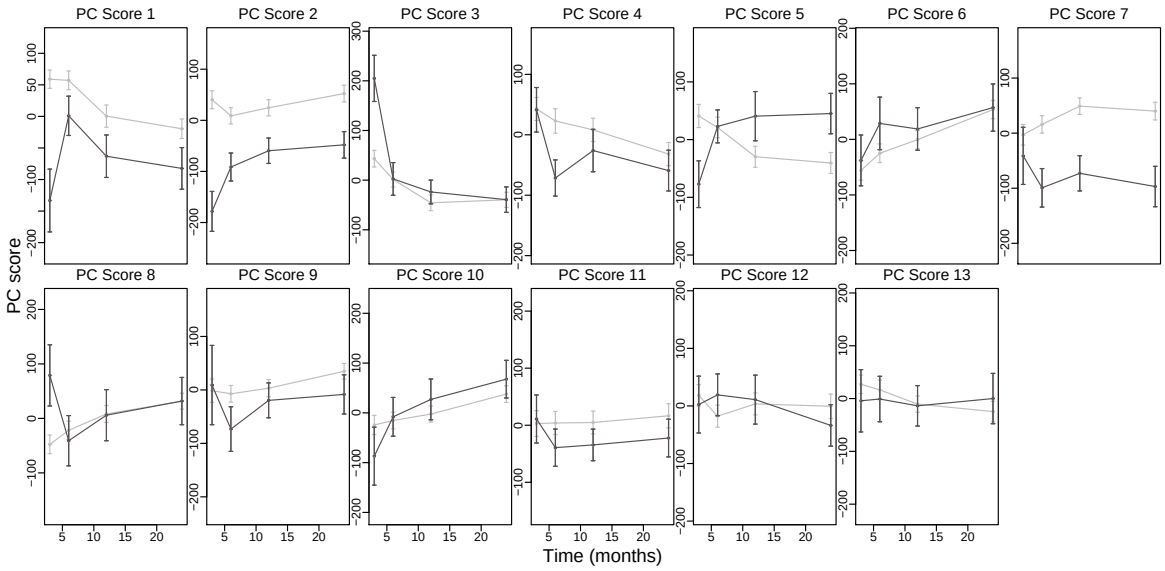

Figure 5.22: Mean trends over time of the principal components scores for the clefts (black line) and controls (grey line), with 95% confidence intervals

The dimension of the shape space, and correspondingly the number of non-zero eigenvalues, is in this case $P = 3 * 22 - 7 = 59$. It would be possible to include all 59 vectors of principal component scores, but this seems unnecessary since a small number of the eigenvectors explain a large amount of the variance in the data. Approximately 90% of the variance is explained by the first 13 PCs and it is these that are included in the analysis as the outcome variable. Figure 5.22 shows the means and 95% confidence intervals of the first $p = 13$ PCs and suggests that the assumption in previous models of a linear time trend for the controls and the clefts from six months is not particularly realistic here. There are still differences in the extent of the variance between the two groups and so this should be reflected in the model.

Figure 5.23 shows all of the PC scores plotted against one another. As would be expected, since the PCs are orthogonal to one another, there is little relationship between any of the

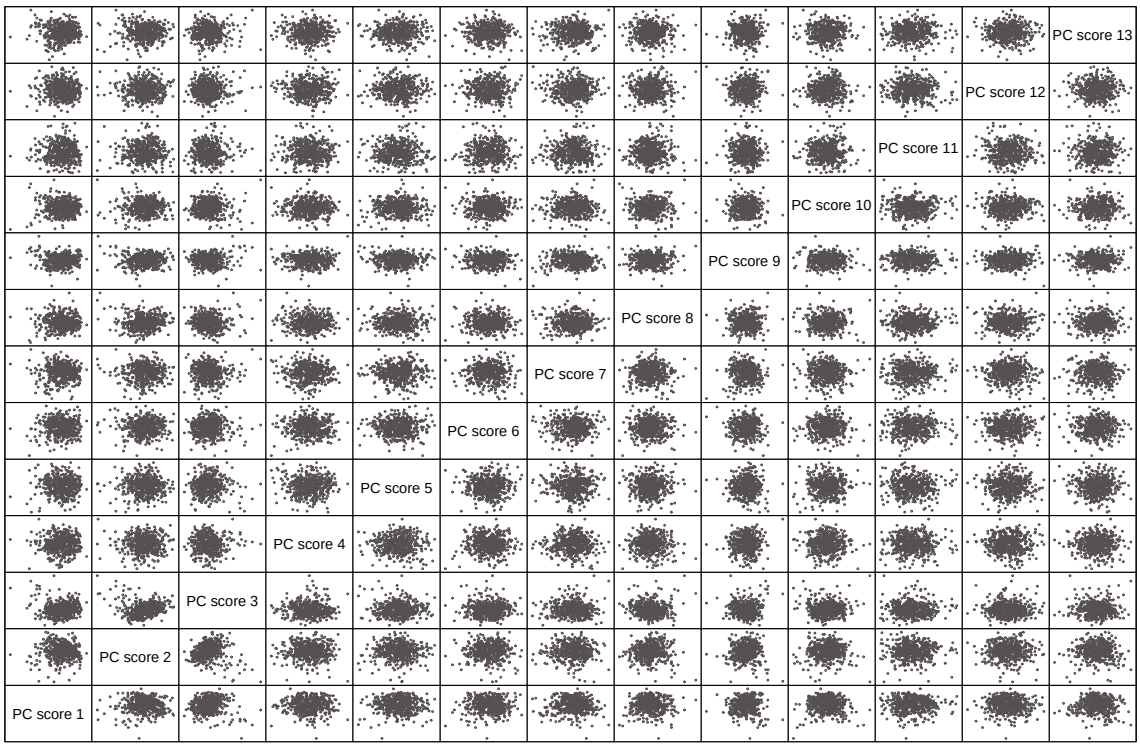

Figure 5.23: Scatterplot of all the pairwise combinations of principal component scores

scores and it may be, therefore, that independent random effects would be adequate. For completeness, however, and also because on a local level with a pair of PC scores depending on various covariates there may be some level of correlation, a similar compound symmetry variance structure to previous models is used, allowing the correlation of the random intercepts for each pairwise combination to be estimated.

The PC scores were all multiplied by 100 so that they were closer in order to unity. For an individual i and pair of PC scores $p = (r, s)$, where $r = 1, \ldots, k-1$ and $s = r+1, \ldots, k$, the model fitted to the response vector $\mathbf{c}_{ip}(t) = (c_{ir}(t), c_{is}(t))$ is

$$\begin{aligned} \mathbf{c}_{ip}(t) &= \boldsymbol{\beta}_{0p} + \mathbf{b}_{ipg_i} + \boldsymbol{\beta}_{1p} g_i + \boldsymbol{\beta}_{2p} t_6 + \boldsymbol{\beta}_{3p} t_{12} + \boldsymbol{\beta}_{4p} t_{24} \\ &\quad + \boldsymbol{\beta}_{5p} g_i \cdot t_6 + \boldsymbol{\beta}_{6p} g_i \cdot t_{12} + \boldsymbol{\beta}_{7p} g_i \cdot t_{24} + \boldsymbol{\epsilon}_{ip}(t), \end{aligned} \tag{5.3}$$

where the random effect vector, $\mathbf{b}_{ipg_i}$, is distributed as in (5.1) and the overall covariance matrix, V_i, has the structure described there. The mean part of the model is also similar to

that of (5.1), but here time is treated as discrete to avoid the assumption of a linear effect of trend in the PC scores. The variables t_6, t_{12} and t_{24} are indicators which take the value one if the time point is 6, 12 or 24 months, respectively, and zero otherwise. The random error, $\boldsymbol{\epsilon}_{ip}(t)$, is once again coordinate specific and so $\epsilon_{ir}(t)$ and $\epsilon_{is}(t)$ have variances σ_r^2 and σ_s^2, respectively.

In order to obtain the results in shape space, the estimated mean PC scores for each combination of covariates must be transformed back to tangent coordinates and then the Procrustes mean added on to these. For an individual i with j^{th} PC score c_{ij}, backtransformation to the vector of tangent coordinates is given by (Dryden & Mardia, 1998)

$$v_i = \bar{v} + \sum_{j=1}^{p} \lambda_j^{1/2} c_{ij} \gamma_j.$$

If $c_j(\hat{\beta})$ is the estimated mean j^{th} PC score for some linear combination of the fixed effects, $\hat{\beta}$, then the corresponding estimated mean vector of tangent coordinates may be estimated as

$$\hat{v} = \bar{v} + \sum_{j=1}^{p} \lambda_j^{1/2} c_j(\hat{\beta}) \gamma_j.$$

The sum can be written in matrix form as $\Gamma \Lambda^{1/2} c(\hat{\beta})$, where Γ is the matrix with the eigenvectors γ_j as its columns; Λ is a matrix with the eigenvalues λ_j on the diagonal and zeros otherwise (and so $\Lambda^{1/2}$ has $\lambda_j^{1/2}$ on the diagonal); and $c(\hat{\beta})$ is the vector of the $c_j(\hat{\beta})$. Therefore the variance of the estimated mean vector of tangent coordinates is

$$\text{var}(\hat{v}) = \Gamma \Lambda^{1/2} V(c(\hat{\beta}))(\Lambda^{1/2})' \Gamma',$$

and the covariance matrix, $V(c(\hat{\beta})) = X \text{var}(\hat{\beta}) X'$ for a particular design matrix X, of the estimated PC scores is obtained using the covariance matrix of the fixed effects estimates produced by the pairwise fitting of the models and subsequent aggregation.

The resulting means and 95% bivariate confidence regions are plotted in shape space in Figure 5.24. They are similar to those in Figure 5.15 in both mean position and variability. The confidence regions in this plot are slightly narrower and some have a tendency to be more ellipsoidal than the corresponding regions in Figure 5.15, which are reasonably spherical.

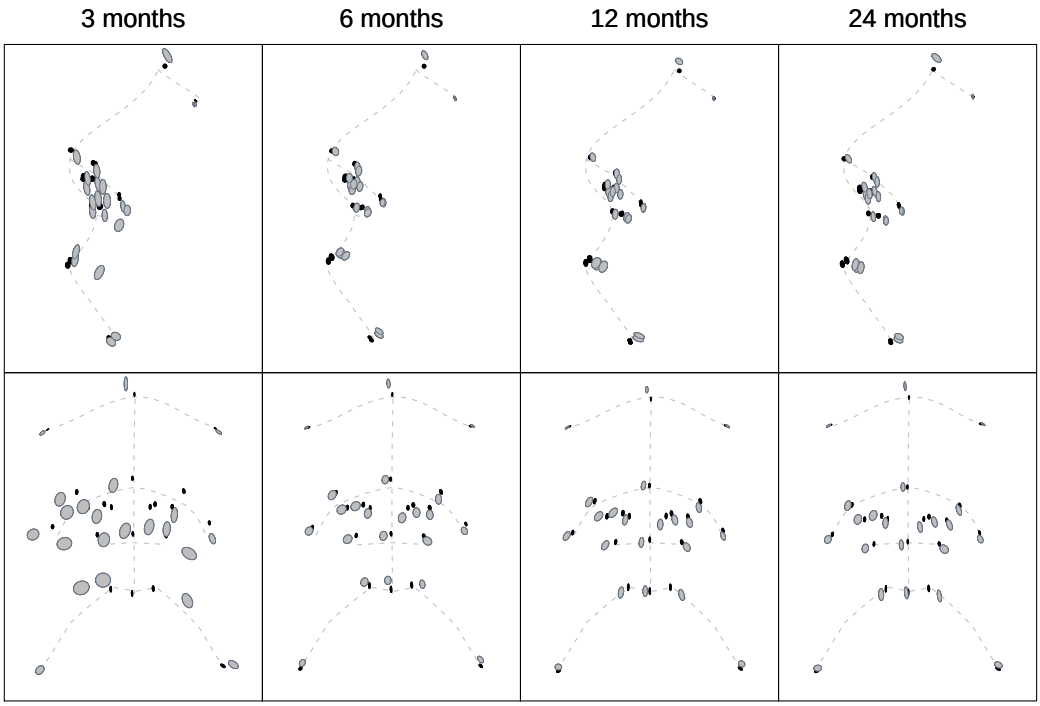

Figure 5.24: Bivariate 95% confidence regions from the model fitted to the principal component scores, for cleft (grey) and control (black) groups for each landmark at each time point (upper: profile, lower: frontal view)

Those that have this tendency generally point in towards the upper centre part of the face and this may be because the main source of variation in the data is the cleft points (particularly around the nose) starting out very wide and flat at three months, and coming in towards the control positions as time elapses. It could also be related to higher estimated correlations using this method, since this will lead to tilted ellipses, rather than vertical or horizontal, which would arise under independence. In any case, given that the first three PCs explain about 70% of the variance, it is not surprising that particular directions may be dominating.

Qualitatively, there are few differences between the results of the dimension separated analysis and this one. There are some small differences around the nostrils in the frontal view at 12 months, and this may reflect the use of discrete time rather than continuous. At 24 months, both figures show very similar means and levels of variation, with differences remaining between the clefts and controls in the clefts having a higher midpoint between the eyes, asymmetry of the end and base of the nose, wider and flatter nostrils and a rather flat

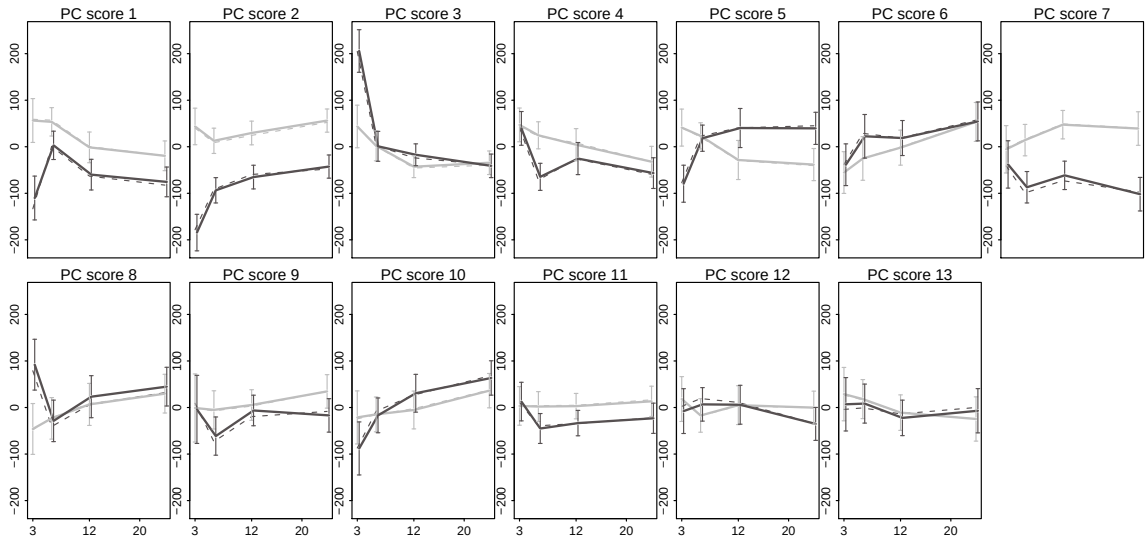

Figure 5.25: Fitted mean trends over time (solid) and 95% confidence intervals from the model fitted to the principal component scores, superimposed onto the empirical means (dashed) for each of the scores and stratified into control (grey) and cleft (black) groups.

and low upper lip, with no discernable Cupid's bow.

Since the saturated mean model was fitted, the estimated mean PC score trends fit very closely to the empirical means, as displayed in Figure 5.25. The fitted lines are not exactly the same as their empirical equivalents because of the reasonably high levels of missing data, the extent of which varies across groups and time points. As previously mentioned, the linear mixed effects model takes account of both individual and group trends when estimating the mean (Diggle et al., 2002) and so if a large number of, say, cleft patients with a high PC 1 score dropped out after 12 months, then the estimated group mean at 24 months would likely be higher than the empirical one.

The residuals versus fitted values are displayed in Figure 5.26 and show a substantially better fit of the covariance structure of the model than previously shown, with reasonably constant variance across the fitted values for all of the PC scores. Two of the scores (3 and 8) show a small change in variance, and this is largely due to the cleft group at three months. Since, by definition, PCA finds components that are orthogonal to one another, analysing the scores removes the need to model the large extent of the correlation between coordinates and this results in a better model fit for the covariance structure.

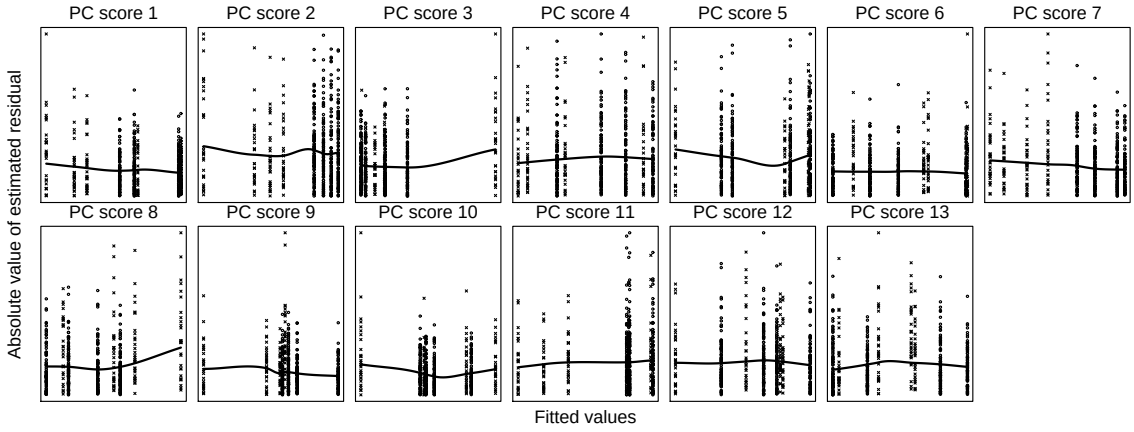

Figure 5.26: Plots of the absolute values of the (unstandardised) residuals versus the fitted values for each PC score, with a loess smooth line fitted

Figure 5.27 gives plots of the estimated random effects SEs and correlations between them for different PC scores. The SEs show similar properties to those from previous models, with the cleft SEs generally being higher than the control ones, and a small selection being in the opposite direction or very close together. Figure 5.27(B) shows an improvement on previous models, with the majority of the correlations lying between -0.5 and 0.5 and none estimated as close or equal to -1 or 1. One might expect all the correlations to be exactly zero since the PCs are orthogonal to one another. However the orthonormality holds on a global level across all of the PCs and here they are considered on a local (pairwise) level, with comparisons considered across various covariates, so it is possible that there may be some correlations that are not particularly close to zero.

5.6 Discussion

This chapter has introduced some alternative ways of applying the pairwise approach, since the application to Procrustes aligned tangent coordinates, as illustrated in Chapter 4, resulted in problems with model convergence and with the estimation of random effects covariance parameters. The use of subset GPA, where two entirely different sets of landmarks were used for alignment and analysis, resulted in fewer such problems and allowed a model to be

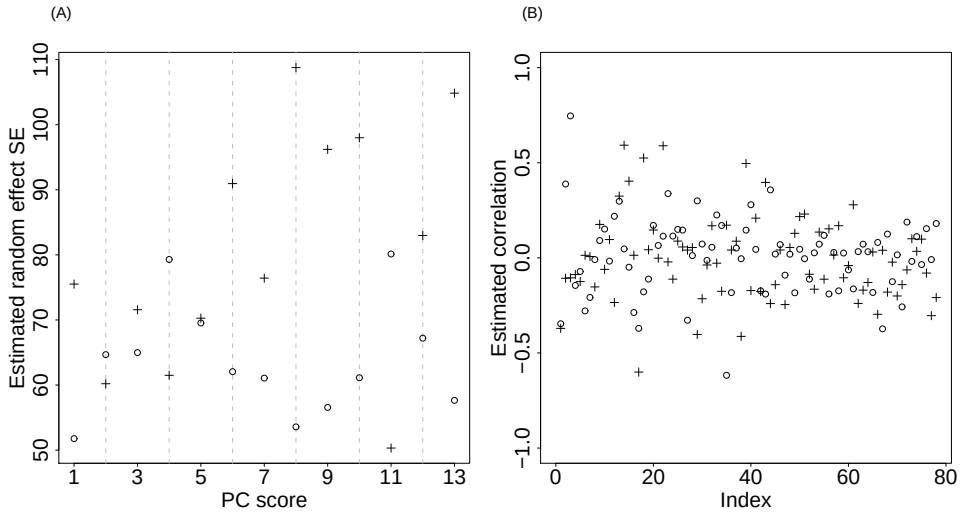

Figure 5.27: (A) Random effects standard error estimates (by coordinate) for the group stratified random intercepts; (B) Estimated correlations between the random effects for pairs of PC scores, stratified by group and in no particular order (both plots: crosses - cleft group; circles - control group)

fitted with correlated random effects which differed for each group. Overlapping sets were considered, but if they were largely coincident then this was likely to result in convergence problems. A small overlap (say, of two landmarks) did not cause such problems.

Since the landmarks included in the alignment were considered to be stable and reasonably unaffected by the cleft defect, the idea was that matching the cleft individuals to each other and to the controls using only these points and excluding others that had very high variability, would result in a more stable and accurate analysis of the resulting tangent coordinates. When compared to the results from the models of Chapter 4, the subset GPA bivariate confidence regions showed more differences between the groups in areas where they would be expected, such as the upper lip, and fewer where they would not, such as the corners of the lips. With data such as these, the analyses are very sensitive to the method of alignment (and the choice of landmarks that are included) and the use of subset GPA allows interpretation of the results as actual differences between the groups rather than facets of the alignment itself. It is possible, however, that although all of the children in this study adopted an open mouth pose, different facial poses may have an effect, particularly since four of the landmarks included in the alignment are located on the lower lip or below. We assume here that differences in facial pose are negligible. These points all highlight the care

that must go into the selection of the landmarks to be included in the alignment.

Fitting the pairwise models to each dimension separately, thereby allowing an unstructured correlation matrix for random effects for different coordinates within each dimension but assuming independence between those in different dimensions (following subset GPA), allowed a larger number of coordinates to be included in the model and resulted in a substantially more detailed analysis. The alignment and analysis landmark sets for the subset GPA overlapped slightly, but this did not appear to be a problem since the two landmarks are fairly stable and unaffected by the cleft. As with the subset GPA analysis, convergence problems were not an issue, but the assumption of independence between dimensions is fairly strong and may not be realistic. The landmarks that were included in both analyses gave comparable results and so it is likely that this assumption makes little real difference in terms of the comparison between the cleft and control groups. If we were to use these results to make predictions about future individuals then there may be more of an important effect.

One issue that arose from allowing correlations between the random effects, in both the subset GPA and dimension separation sections of this chapter, is that some of these correlations were, apparently spuriously, estimated as exactly or very close to one. Simulations showed that when the true value of a random effects SE is very close to zero and the optimisation process is therefore looking for an estimate close to the boundary of the parameter space, correlations containing the corresponding variable may be incorrectly estimated as having an absolute value of one, even if they are actually quite close to zero. This does not appear to have a large effect on the fixed effects estimates, but if prediction or model comparison tests were to be employed using either of these models, then it is possible that this could cause inaccuracies.

This suggests that, although tangent coordinates may be analysed using normal multivariate methods, the fact that in an application such as this there are some which have an extremely large amount of variation (such as the points around the upper lip), at least for a subset of the population, and others that have very little, analysing these coordinates as they are may not be an entirely appropriate way to proceed.

An alternative approach is to apply the pairwise models to the principal component scores of the tangent coordinates, following subset GPA. This enables a large reduction in dimension,

since 13 PCs explained approximately 90% of the variance of 66 coordinates. As with the dimension separation approach, this allows substantially more landmarks to be included in the analysis, but does not make such strong assumptions about the correlation structure. A further benefit of this approach is the substantial reduction in computation time, as there are only 13 different outcomes, compared to 22 for each of three dimensions in the dimension separation approach. Since there are a small number of PCs that explain most of the variance, the results of this analysis do tend to be somewhat driven by these particular directions, but aside from this the results are very similar to those obtained under the dimension separation. The bivariate confidence regions are slightly larger under the dimension separation and this may be due to the assumption of independence between dimensions and there being fewer models fitted as a result. The model covariance structure for the PCA approach shows a good fit to the data and is unique amongst the models considered in this respect.

In general, then, the principal components approach following subset GPA avoids the problems with convergence and spuriously estimated correlations and allows a larger number of landmarks to be included. Although the results from this method are similar to some previous ones, this seems to be more stable and robust for the application of the pairwise approach to three-dimensional shape data.

Chapter 6

Curve analysis

6.1 Introduction

In the previous two chapters, facial shapes were described using anatomical landmarks that were placed on three-dimensional images. These landmarks were placed in standardised positions that were considered to be important with respect to the cleft-lip defect. One of the disadvantages of such a description of facial shape is that substantial amounts of information about the areas between landmarks is lost. Potentially more information can be gleaned by the analysis of curves rather than landmarks alone.

Some work has been done using curves to describe shapes but it is not a widely studied area, particularly in the case of curves measured repeatedly over time. In the area of human motion Faraway (1997) used functional regression to study the arm motion of an individual reaching for certain targets, primarily in order to predict future motion towards a new target given its position. Linear and quadratic models were fitted, with the response variable being the positions of points at various small increments along the curve, with the three-dimensional coordinates of the target (and combinations of them) as the explanatory variables. Under this scenario, the points on the curve are considered to occur at certain time points, t_j, but this has an obvious correspondence with the situation where the curve is parameterised as $X(t)$ where $t \in [0, 1]$ does not necessarily represent time, but rather the distance along the curve (such as will be the case for the curves from the cleft study). A bootstrap-based

approach was used both for model comparison and to provide confidence bands both for the mean curve and for a new observation. This aspect of the work will be discussed more fully in the following chapter.

Faraway (2004) studied the movement of landmarks on the face, specifically during the motion of a smile. The author fitted B-splines to the curve describing the relative change in distance from rest between two landmarks across the entire motion and was thus able to impose an average smile on the face of an individual, in order to compare actual and average motion. A comparison of the average smile on the average face, both of normal individuals, with the average smile of patients with abnormally sized jaws on the average normal face, showed that differences were more likely to result from different facial shapes than from different smile motions. A linear mixed model fitted to the PC scores confirmed no evidence of a difference between these two groups. A further paper (Trotman et al., 2005) applied these methods in a similar way to a study comparing the facial movement of individuals with cleft-lips that have been repaired to normal controls.

In an analysis of facial shape, Bowman & Bock (2006) used a mainly graphical approach in considering the variation in the first few principal components (PCs) of the tangent coordinates of facial curves, obtained from the cleft study considered here. As mentioned in more detail in Chapter 2, the authors describe the differences across the curves at one time point through plots of ± 3 standard deviations on the PC scale either side of the mean, and they adopt a 'tour' approach, which shows smooth deformations of the nose profile curve as it moves throughout PC space. Individual cleft cases were each compared to the closest case in the normal range of controls and a severity index proposed which showed that some of the clefts were within that range. Differences between the groups were detected by a permutation test, but no attempt was made to compare trends over time between the two groups. Bock & Bowman (2006) studied facial curves from the same study in terms of asymmetry and this is also mentioned in greater detail in Chapter 2. Although the authors did fit a repeated measurements model, they considered the different time points separately.

Facial recognition is another area in which facial curves have been used. Samir et al. (2006) represented facial surfaces by unions of curves and formed a metric on the images by finding the geodesic between the curves in shape space and thereby the distance between the surfaces.

They found that with a range of images of an individual and a selection of curves, there was a very high automatic recognition rate. Another example in this area is Gokberk et al. (2006).

Decker et al. (2007) described ankle and knee cyclograms, the movement of the ankle and knee throughout one stride cycle. The individuals in the study had markers placed on their ankle and knee joints and were videotaped in lateral view whilst running. Particular reference frames of the video were chosen and the position of the marker calculated for each frame. In this way, the positions of 14 landmarks in two dimensions were found and these formed a curve outlining the entire cycle of one stride. The curves were aligned using Procrustes matching and the PC scores obtained for the approximate tangent coordinates. The authors studied the main PCs to find the main variant and invariant parts of the cycle and applied a linear discriminant analysis to the PC scores to discriminate between middle-distance runners and sprinters.

A further development in the field of human motion was Faraway et al. (2007) who used Bézier curves to describe the movement of an individual's arm as they reach for an object and move it to a different location. Bézier curves are constructed from a linear combination of degree d polynomials and d control points, which determine the shape of the curve. The first and last of these points are fixed at either end of the curve and the segments connecting them to the second and penultimate control points, respectively, are at tangents to the curve. These line segments, therefore, describe the nature of the curvature at its endpoints. Bézier curves arise as a special case of B-splines, where the first $d + 1$ knots are at zero and the last $d + 1$ are at one (where the curve is parameterised to lie between zero and one), with no internal knots. Two different parameterisations were used to model the effect of various covariates (such as height, age, gender); one taking the response as the difference between the endpoints and their closest control point, and the other the difference between the interior control points and evenly spaced axis points (on the straight line between the two endpoints). These were modelled using standard multivariate modelling techniques and used primarily for prediction of the reaching motion.

A Bayesian approach to the analysis of human motion data was taken by Alshabani et al. (2007). Individuals pointed towards the positions of five different targets, placed at various angles in front and to either side of them. The movement of a marker placed on the index

finger was monitored throughout the movement and a set of points recorded to describe the curve. Procrustes registration was necessary to align the curves in time to an idealised template curve, so that the main features of the movement occurred at the same points. Equally spaced points in time were placed between the main landmarks and connected using piecewise linear interpolation. The starting values of the parameters in the MCMC were obtained from the estimates of the landmarks obtained from the Procrustes registration. This approach removes the need to mark landmarks onto images and the inherent noise within such types of landmark location.

Dryden et al. (2008*b*) analysed the spine curves of various different types of sibling pairs (same sex twins, different sex twins, non-twin siblings), in order to detect whether spine shape is genetic or sex-related. The curves were described by pseudo-landmarks placed along the spine between two fixed landmarks at the top and bottom of the spine. They were Procrustes aligned and the first few PCs taken to test for intra-familial correlation for each of the sibling types. Some individuals had measurements taken repeatedly over time but these were used only to test for repeatability in terms of the correlation of each individual's PC scores.

Although these papers show the variety of work that has been carried out on curve data, there has not been extensive work on this in the field of shape analysis and especially in terms of repeated curves over time. This chapter aims to address such issues and will proceed by employing the pairwise methods previously introduced to the analysis of curve data, providing a strong visual interpretation of the comparison between the groups in the cleft data that was introduced in Chapter 4.

6.2 Facial curve representations

The facial curves are represented by pseudo-landmarks placed at various small equal intervals along particular features of the face between fixed anatomical landmarks, such as those introduced in Chapter 4. The positioning of the pseudo-landmarks was carried out by colleagues in the Department of Computing Science at the University of Glasgow. The anatomical landmarks are fixed in the sense that they correspond to particular points on the

face, whereas the pseudo-landmarks may move along the curve in either direction and do not have a specific anatomical meaning. The pseudo-landmarks could have been placed at unequal intervals, for example, potentially to give more weighting to certain parts of a curve. Since some of the curves cross the midline, at the intersections one of the fixed anatomical landmarks had to be removed from each image so as to avoid duplication. As the initial analysis concentrates on the midline curves, the duplicated points were removed from the other curves.

The curve data arises from the same cleft-lip and palate study introduced in Chapter 4. The full set of curves was not marked onto all of the subjects in the study, however, and so the dataset is smaller here, with 52, 60, 73 and 66 individuals in the control group and 19, 19, 23 and 20 individuals in the cleft group at 3, 6, 12 and 24 months, respectively.

Figure 6.1(a) displays the points used to form the curves that will be included in these analyses, with the fixed landmarks highlighted by solid black circles. The curves presented are the midline, which runs from the centre point between the eyes to the midpoint of the upper lip; the bridge of the nose, which joins the inside corners of the eyes to the midpoint; the rim of the nose, which runs from the corners of the nose through the tip; the base of the nose; and the upper lip, which runs from the corners of the lip through the Cupid's bow to the midpoint. By the nature of the cleft-lip deformity, the dataset does not include points on the upper lip curve for the cleft children at three months of age (prior to their surgery). Hence, as in Figure 6.1(b), we have removed the upper lip data from all the individuals in the study at all time points, since at any chosen time point an individual must have a full set of landmark positions in all dimensions. From this point onwards, the full set of curves refers to those displayed in Figure 6.1(b).

Four examples of Procrustes aligned cleft cases superimposed onto control cases, all measured at three months, are displayed in Figure 6.2 with interpolating curves joining the anatomical landmarks. We initially focus on the midline curve. Just as in the landmark examples, there are various ways of aligning the curves in order to remove the effects of location, rotation and scale and leave only shape. Different ways of applying generalised Procrustes analysis (GPA) in the case of curves include the use of the full set of curves in Figure 6.1(b). As in the landmark example, however, this tends to result in midline curves that appear

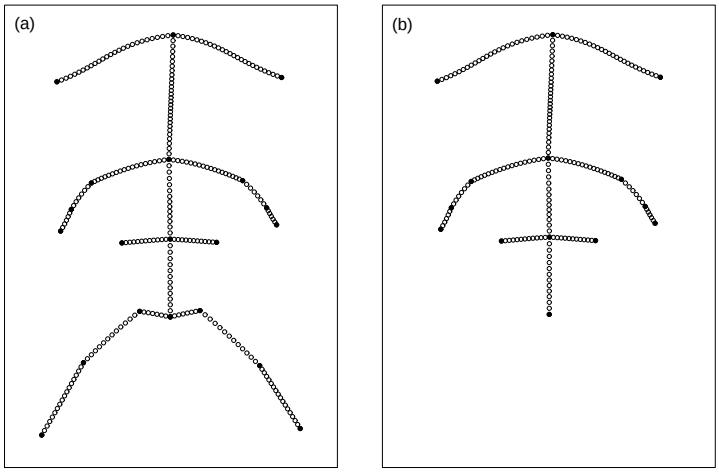

Figure 6.1: (a) Example of the pseudo-landmarks (open circles) forming the curves used to describe a face with (bottom to top) the upper lip curve, base of the nose, rim of the nose and bridge of the nose. The midline of the face runs vertically from the midpoint between the eyes to the midpoint of the upper lip. (b) Example of the curves used to describe a face, with the upper lip curve removed. In both plots the fixed landmarks are represented by solid black circles.

more similar across individuals in terms of asymmetry than they should, which leads to the models estimating zero standard errors for the random effects. Alternatives include using curves other than the midline for the alignment, or using landmarks only, whether they are the full set of landmarks or only those that do not lie on the midline. The landmarks are highlighted by solid black circles in Figure 6.1(b). The use of curves other than the midline gave similar aligned images to the matching using all of the curves because of the effects of the more unstable parts of the image, such as around the rim and base of the nose. The inclusion of the landmarks on the midline also tends to artificially reduce asymmetry and so the alignment was carried out using only the landmarks that do not lie on the midline.

The cases in Figure 6.2 illustrate the extent of the cleft defect when compared to controls, particularly in terms of asymmetry and the collapse of the nose rim. The figure also shows the large amount of variation within the cleft group, due to the varying severity of the cleft defect. The two cleft cases on the right side of the figure have substantial levels of asymmetry, with the upper individual having more of a collapsed nasal area. The lower right case has considerable asymmetry in the level of the eyes, while the lower left cleft case has eyes that are very much lower than the bridge of the nose. The cleft case on the upper left seems to

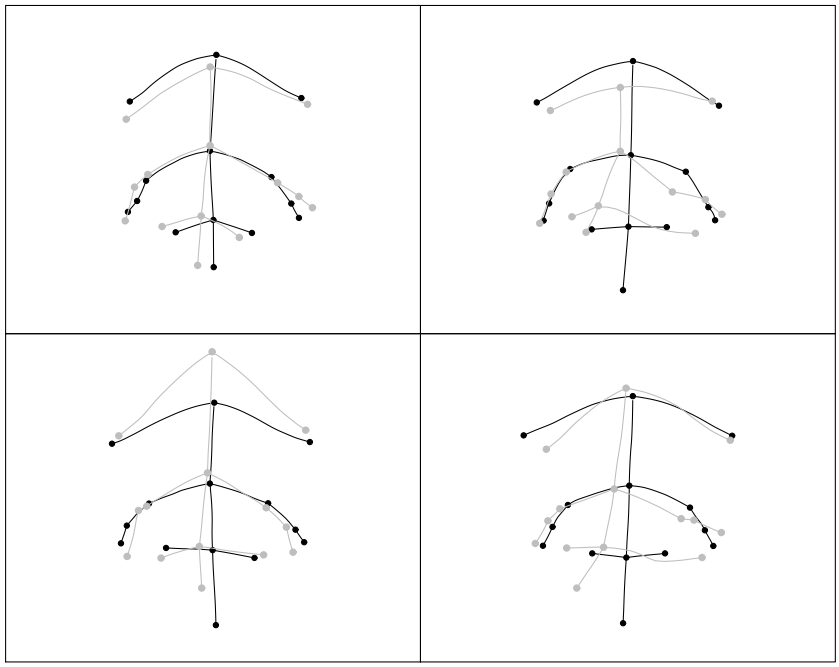

Figure 6.2: Four examples of randomly chosen cleft (grey) and control (black) cases at three months, showing both the interpolating curves and superimposed landmarks

have a relatively mild deformity, with curves that are reasonably comparable with control positions.

It is clear from Figure 6.1 that the curves convey considerably more information about facial shape than the landmarks do alone. However the number of points used to construct each curve is markedly higher (midline 57, bridge 52, rim 64, base 24, all in three dimensions) and it is necessary to find a way of representing this information in a more parsimonious manner, before we can fit the pairwise models. This chapter proceeds by discussing two ways of approaching this dimension reduction issue, those of B-splines and of principal components analysis (PCA). In both cases, the pairwise models will be fitted and the results subsequently compared.

6.3 B-splines of profile curves

The dimension of curves such as the midline may be reduced by fitting B-splines, cubic splines that smoothly join piecewise polynomials at a selected number of knots (Ramsay & Silverman, 1997). The mean midline curves for the cleft and control individuals are displayed for each time point in Figure 6.3, with the landmarks superimposed. It is clear from the frontal view that there is a great deal of asymmetry in the cleft group at three months which is substantially reduced by six months, but that a small amount persists to two years of age. The cleft group on average has a considerably flatter nose at three months than the controls and a much shorter upper lip, but although there is a small amount of residual deformity remaining at 24 months in both, the groups become much more similar in these aspects. The higher and asymmetric bridge of the nose in the cleft group remains and becomes only a little closer in terms of symmetry to the corresponding position for the controls.

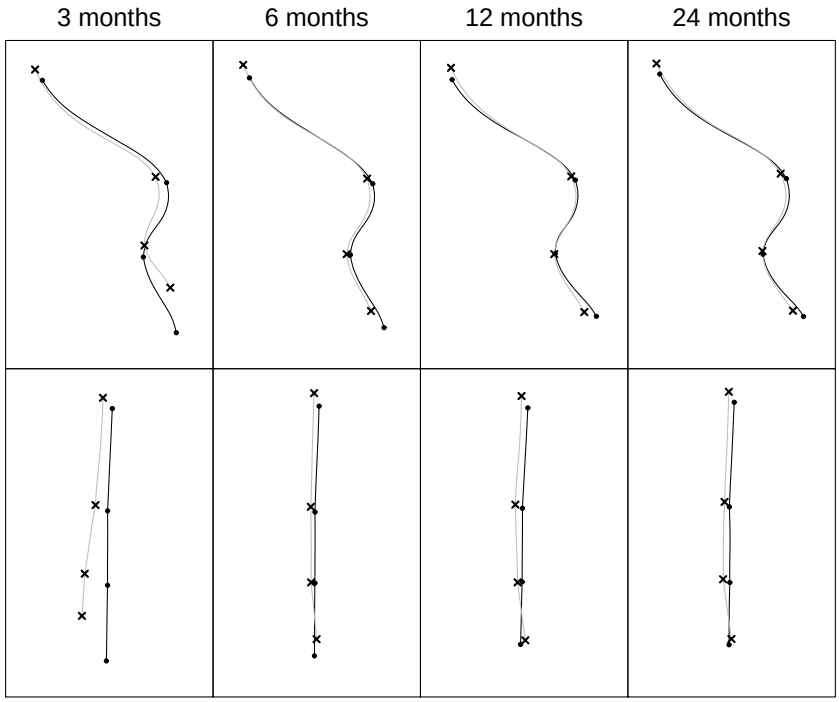

Figure 6.3: Interpolated mean curves for each group at 3, 6, 12 and 24 months for the cleft (grey line, crosses) and control (black line, solid circles) groups, with landmarks superimposed. Upper - profile view; lower - frontal view.

In order to fit the B-splines, the curves were parameterised as $(x(d), y(d), z(d))$, where $d \in [0, 1]$ represents the distance travelled along the curve as a proportion of its entire length. For any individual i (at some time point), the continuous midline curve in the j^{th} dimension $(j = 1, \ldots, m)$, $u_{ij}(d)$, has discrete realisations v_{ijl}, the value of the function at the proportion $d = d_l$ along the curve. The function is described by the following vector of points:

$$u_{ij}(d) = c_{ij0}(d) + \sum_{\kappa=1}^{K} c_{ij\kappa}(d)\phi_{i\kappa}(d),$$

where the $c_{ij\kappa}$ are spline coefficients, K is the chosen number of knots and $\phi_{i\kappa}(d)$ is the basis vector describing the part of the curve corresponding to knot κ (Ramsay & Silverman, 1997). This holds for each dimension, so there are Km spline coefficients describing each midline curve. The vector of spline coefficients for the curve for individual i in dimension j, c_{ij}, is determined by minimising the least squares criterion $(v_{ij} - \Phi_i c_{ij})'(v_{ij} - \Phi_i c_{ij})$, where Φ_i is the $q \times K$ matrix whose columns are the basis vectors $\phi_{i\kappa}$, q is the number of points used to describe the curve and v_{ij} is the vector containing all v_{ijl}. The solution gives the vector of spline coefficients as $c_{ij} = (\Phi_i'\Phi_i)^{-1}\Phi_i' v_{ij}$ (Ramsay & Silverman, 1997).

Each individual curve has a B-spline fitted and the resulting set of spline coefficients are used to form the outcome variable, which is the response in the linear mixed effects model below. Figure 6.4 shows the fit of the B-splines to the cleft and control group mean curves for different numbers of knots and suggests that seven knots is the smallest number that fits a spline which adequately describes the mean curves for both groups. However, for the purposes of this analysis, we proceed with the eight knot spline as this proves to be a substantially better overall fit of the data, since using seven knots misses some of the curvature around the nose. In the frontal view all fit extremely well since there is substantially less curvature than in profile.

The 24 mean spline coefficient trends (eight in each of three dimensions), with 95% confidence intervals, are displayed for each group in Figure 6.5. For the most part, these show a large surgery effect between 3 and 6 months, as for the landmark tangent coordinates, and then either a linear or slightly curved trend between 6 and 24 months, at least for the cleft group. The control group shows a roughly linear trend across the entire time period for most

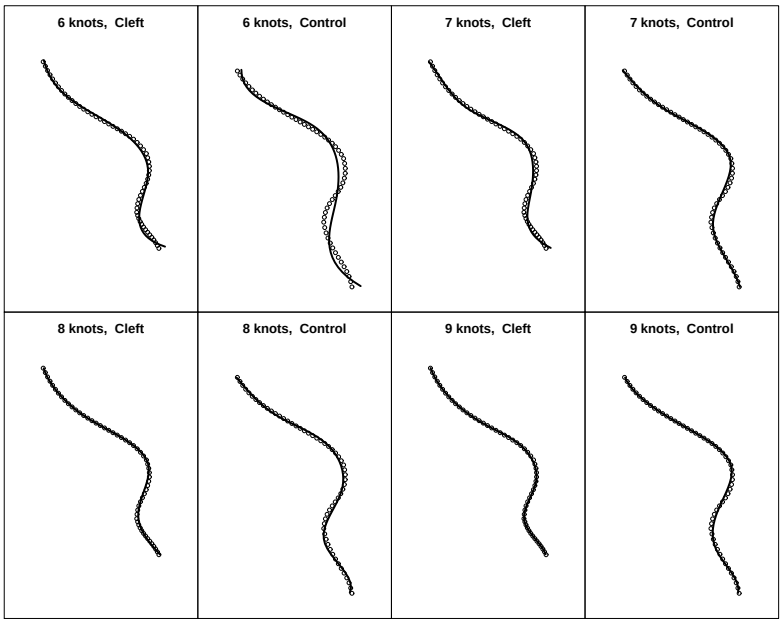

Figure 6.4: B-splines fitted to the mean cleft and control group curves in profile with different numbers of knots

coefficients, although some have curvature. There are clearly differences in the variability between groups and coefficients and these must be accounted for in the model.

The fitted model, for a single outcome, where the value of spline coefficient $r = 1, \ldots, Km$ is denoted as $c_{ir}(t)$ for individual i at time t months after the beginning of the study, is

$$c_{ir}(t) = \beta_{0r} + b_{irg_i} + \beta_{1r}g_i + \beta_{2r}t + \beta_{3r}t^2 + \beta_{4r}g_i \cdot t + \beta_{5r}g_i \cdot S(t) + \beta_{6r}g_i \cdot t^2 + \epsilon_{ir}(t), \quad (6.1)$$

where the surgery effect $S(t) = 1$ if $t > 3$ and is zero otherwise. The random error term, $\epsilon_{ir}(t)$, is distributed as $N(0, \sigma_r^2)$ and therefore the error variance is specific to the coefficient. The covariate g_i is an indicator taking value 1 if individual i is in the cleft group and 0 otherwise. A quadratic time trend has been included, from 6 to 24 months amongst the clefts and for the entire period for the controls, to allow for any curvature in the coefficient trend. For any individual i who is in the control group, their random intercept vector for

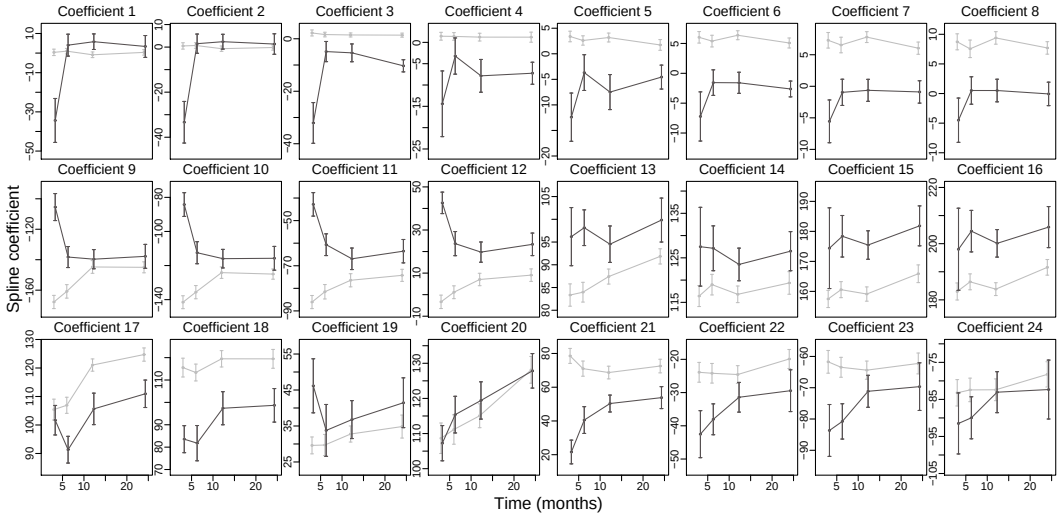

Figure 6.5: Mean trends over time, with 95% confidence intervals, of the spline coefficients for the cleft (black lines) and control (grey lines) groups (note the different scales)

pair $p = (r, s)$, where $r = 1, \ldots, m-1$ and $s = r+1, \ldots, m$, is distributed as

$$\mathbf{b}_{ip0} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{r0}^2 & \tau_{rs0} \\ \tau_{rs0} & \tau_{s0}^2 \end{pmatrix}\right).$$

For a cleft group subject,

$$\mathbf{b}_{ip1} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{r1}^2 & \tau_{rs1} \\ \tau_{rs1} & \tau_{s1}^2 \end{pmatrix}\right).$$

In order to present the estimated mean curves in shape space, as displayed in Figure 6.6, the results from the model must be backtransformed from the spline coefficient scale. This backtransformation may be calculated for each dimension using $\bar{\Phi} X \beta^*$, where $\bar{\Phi}$ is the basis matrix for the Procrustes mean shape, β^* is the subset of parameters of θ^* containing only fixed effects and X is a design matrix corresponding to the particular group and time point of interest. The variance matrix in shape space is determined by $\bar{\Phi} X \mathrm{var}\beta^* X' \bar{\Phi}'$.

Figure 6.6 displays the estimated mean curves for the cleft and control groups at each time point, in the profile and frontal views, following transformation back into shape space. Bivariate 95% confidence regions were calculated at 100 closely located points along the mean

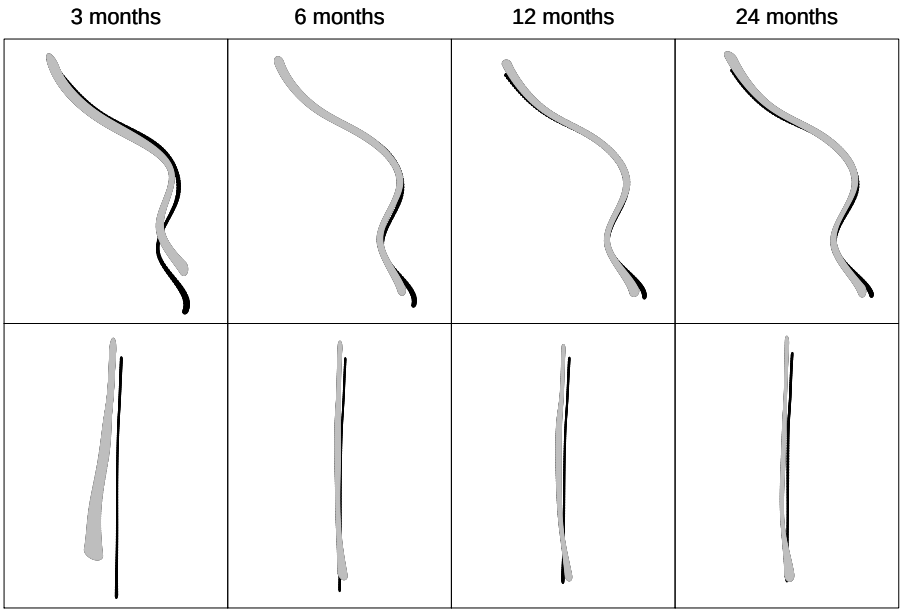

Figure 6.6: Model estimates for the midline curves, with 95% bivariate confidence bands for cleft (grey) and control (black) groups for the mean midline curves at each time point in the profile (upper) and frontal (lower) views

curves which create the impression of overall two-dimensional confidence bands. There is a substantial level of asymmetry amongst the cleft individuals prior to their surgery, along with a much flatter nose and shorter upper lip than the average control subject. The figure shows that on average the flatter nose of the cleft group subjects becomes more similar to that of the controls as the children grow subsequent to their surgery, whereas there is still some asymmetry in the frontal view of the curve leading down through the nose to the upper lip. By 24 months, the short upper lip of the average cleft individual has lengthened to become almost the same length as that of the controls. In general the profile of the clefts becomes reasonably similar to the controls, albeit with a slightly flatter lip, but in the frontal view there remains some considerable asymmetry.

Figure 6.7 shows the fitted mean trends over time, superimposed on the empirical means of each of the spline coefficients and all estimated trends fit the data well. Figure 6.8 shows a reasonably good fit of the model to the data, with the fitted means (with confidence bands) on the whole closely matching their empirical equivalents in shape space. The exceptions are the control curves in profile at the earlier time points, where the fitted curves do not quite

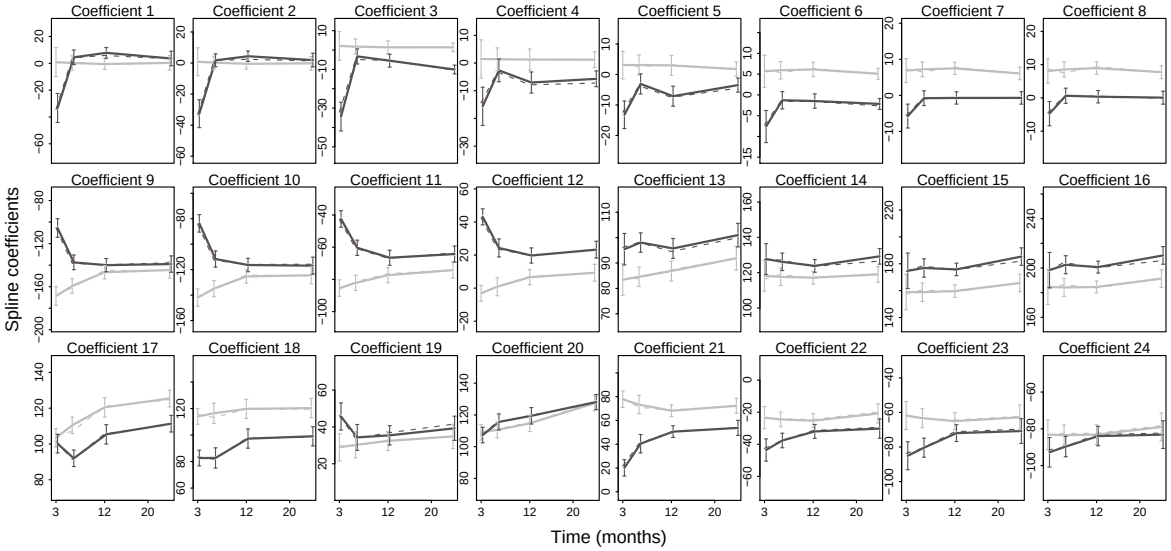

Figure 6.7: Fitted mean trends over time (solid) and 95% confidence intervals for each of the spline coefficients, superimposed onto the empirical means (dashed) and stratified into control (grey) and cleft (black) groups.

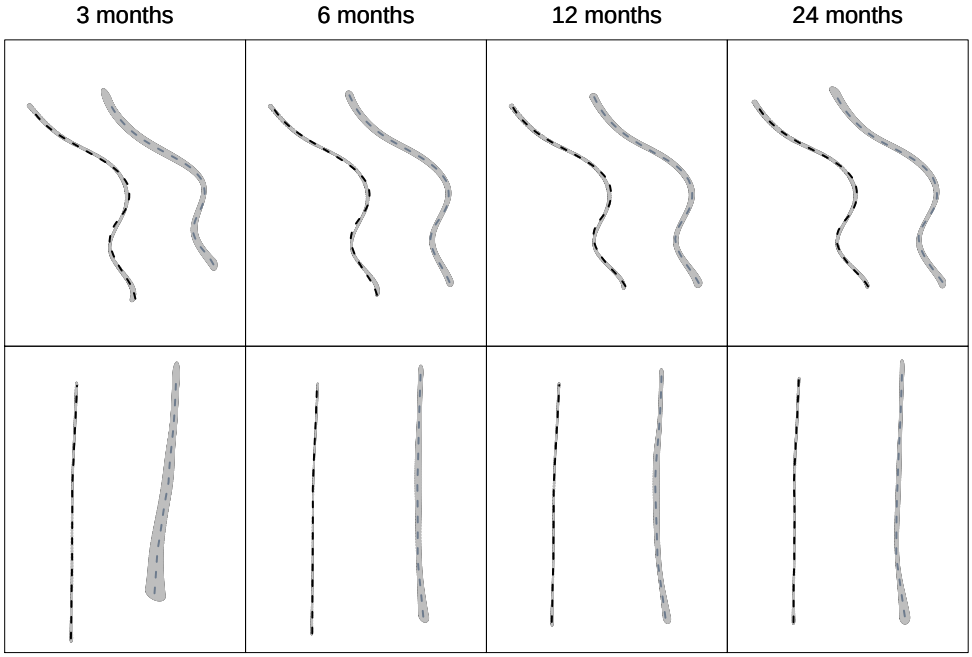

Figure 6.8: Model estimates for the midline curves and 95% bivariate confidence bands superimposed onto the empirical mean midline curves for the cleft (grey confidence interval, dark grey dashed line) and control (grey confidence interval, black dashed line) groups, at each time point in the profile (upper) and frontal (lower) views. The group curves have been shifted apart for ease of viewing.

express the points of most extreme curvature. If the model is fitted with seven knots, but with a saturated mean part, the estimated spline coefficients show a very close fit to the data but in shape space there are a number of discrepancies, with the splines being inadequate at describing some of the areas of highest curvature for either group (plots shown in Figures 6.9 and 6.10). Fitting a saturated model with eight knots leads to some convergence problems and so this is not possible, and interestingly, although Figure 6.7 shows a substantial improvement in model fit over a model with only linear effects of time (plus a surgery effect), there is very little difference between them in shape space (plots shown in Figures 6.11 and 6.12).

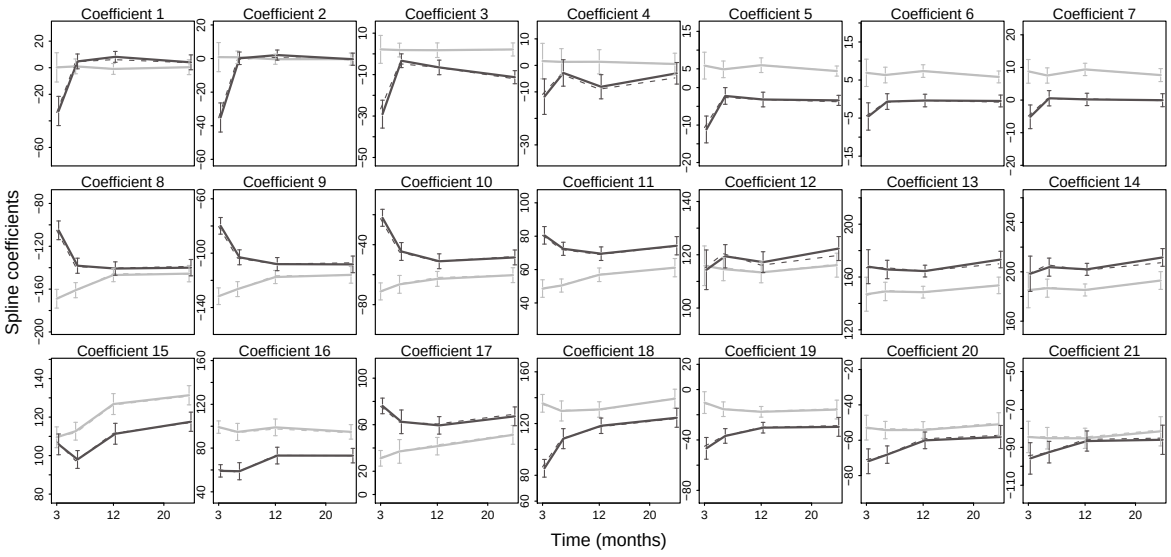

Figure 6.9: Fitted mean trends over time (solid) and 95% confidence intervals for each of the spline coefficients, for a spline with seven knots and a model with a saturated mean part, superimposed onto the empirical means (dashed) and stratified into control (grey) and cleft (black) groups.

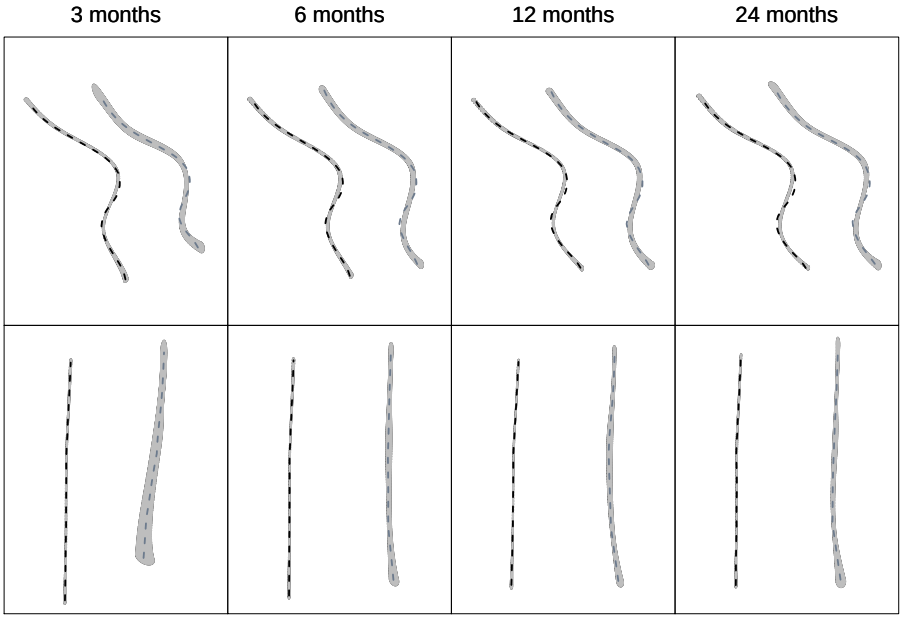

Figure 6.10: Model estimates for the midline curves and 95% bivariate confidence regions, for a spline with seven knots and a model with a saturated mean part, superimposed onto the empirical mean midline curves for the cleft (grey confidence interval, dark grey dashed line) and control (grey confidence interval, black dashed line) groups, at each time point in the profile (upper) and frontal (lower) views

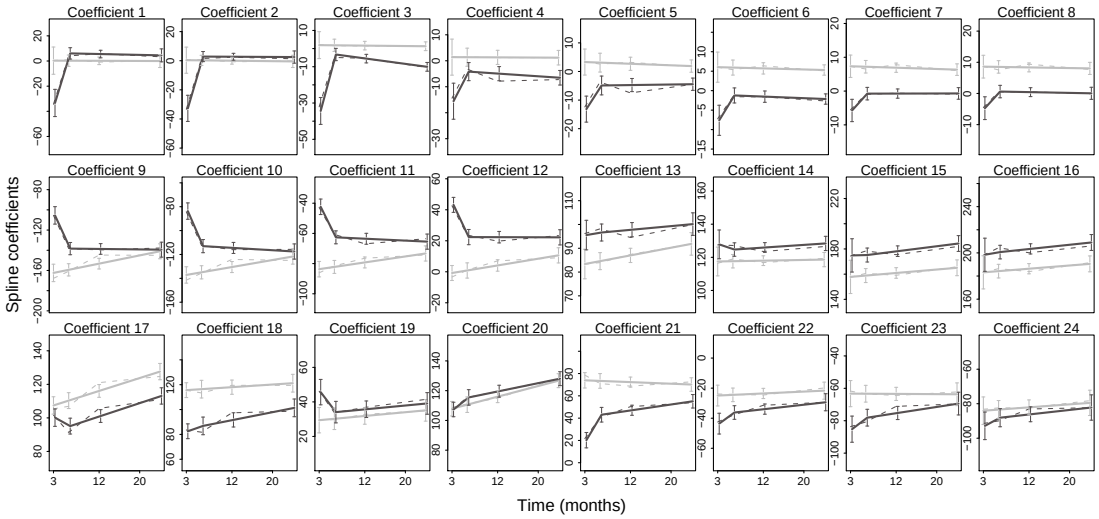

Figure 6.11: Fitted mean trends over time (solid) and 95% confidence intervals for each of the spline coefficients, for a spline with eight knots and a model with a linear effect of time from six months, superimposed onto the empirical means (dashed) and stratified into control (grey) and cleft (black) groups.

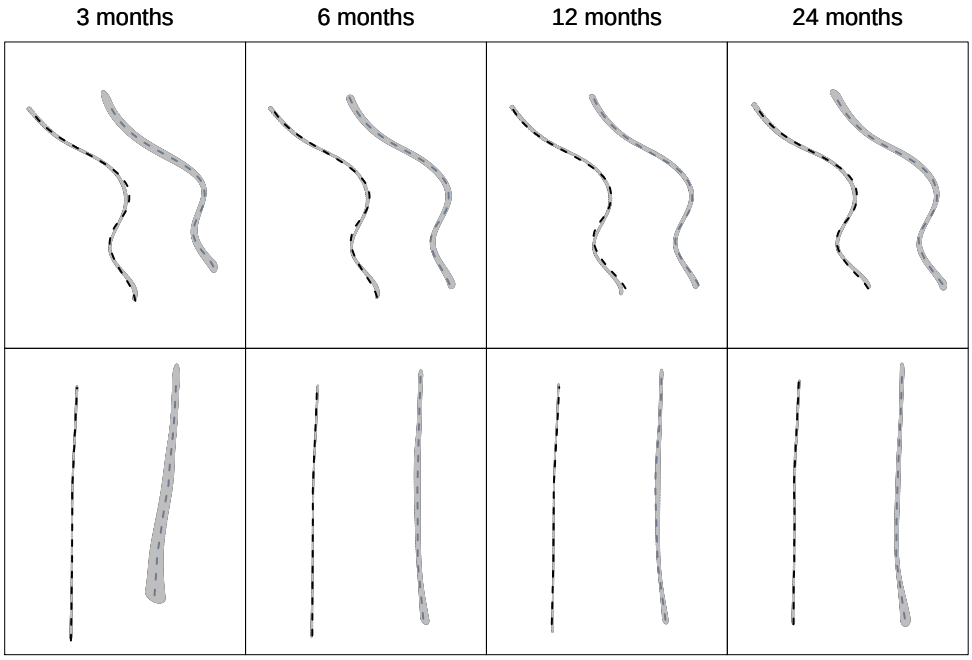

Figure 6.12: Model estimates for the midline curves and 95% bivariate confidence regions, for a spline with eight knots and a model with a linear effect of time from six months, superimposed onto the empirical mean midline curves for the cleft (grey confidence interval, dark grey dashed line) and control (grey confidence interval, black dashed line) groups, at each time point in the profile (upper) and frontal (lower) views

Figure 6.13 gives the residuals versus fitted values and shows that, for the most part, the variance is constant. The exceptions are the first three coefficients along with nine and ten and the high residuals are for the cleft group at three months. The trends for these coefficients are not particularly extreme, which is an improvement over previous models that take the tangent coordinates themselves as the response variable. Furthermore, the size of the cleft group has been reduced because not all of the individuals who had landmarks marked onto their images also had curves. The high variance residuals, which arise from the three month cleft group images, are clearly related to the small sample size. Despite these issues, however, clearly the fit of the model could still be improved with respect to the covariance structure.

The separate random effect SE estimates for the two groups are displayed in Figure 6.14(a), along with the random error variances for each coefficient in Figure 6.14(b). There are clear differences between the coefficients in both, and also substantial differences between the groups in terms of the random effect SE estimates. There are several very small estimates of

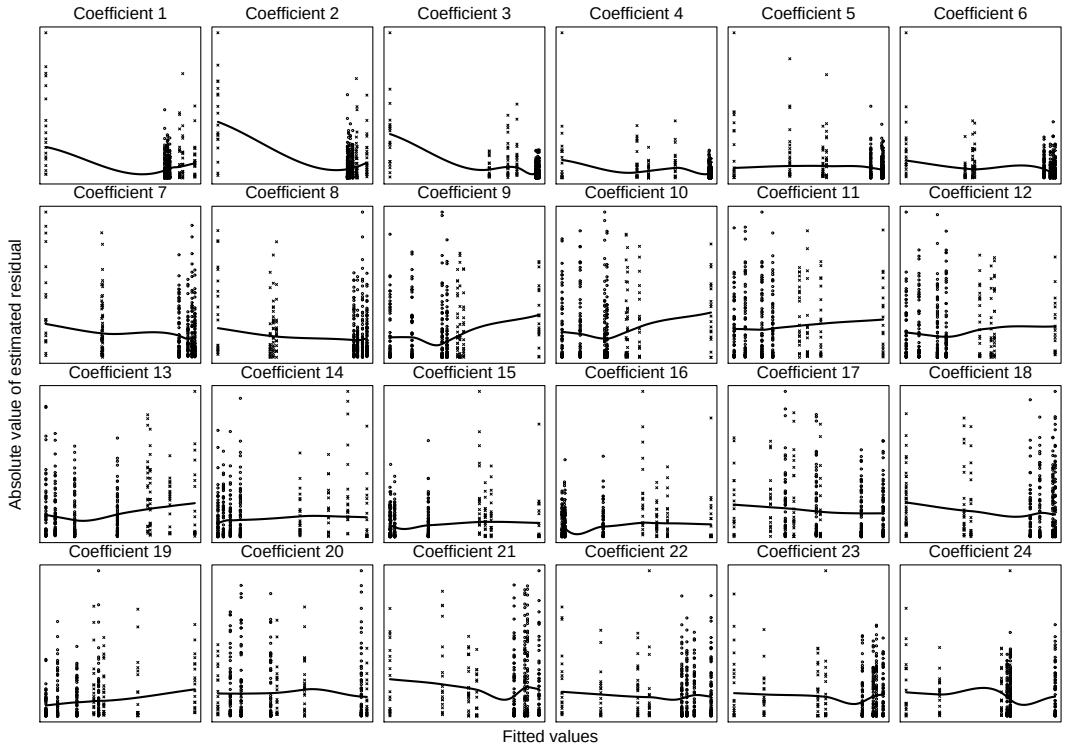

Figure 6.13: Plots of the absolute values of the (unstandardised) residuals versus the fitted values for each spline coefficient, with a loess smooth line fitted

the random effect SEs for the control group and this, in turn, leads to the correlation patterns that may be observed in Figure 6.14(c), which shows the model-estimated correlations of the random effects for pairs of coefficients. There are numerous estimated correlations that are very close to one in absolute value, particularly for the control estimates. This results from the small random effect SE estimates for this group, which do not affect the within dimension correlations to such an extent because these have very small covariance estimates.

This method of analysing the profile curves gives a useful way of reducing the dimension so as to fit the pairwise models, but the model checks suggest that there are substantial numbers of spuriously estimated correlations and very small random effect SEs. Therefore, this does not seem a particularly robust method to employ alongside the pairwise approach. Furthermore, the number of spline coefficients necessary to adequately describe the profile curves is close to the outcome limit that the pairwise models can fit. Therefore an analysis of a set of curves (such as displayed in Figure 6.1(b)) would require an assumption of independence between

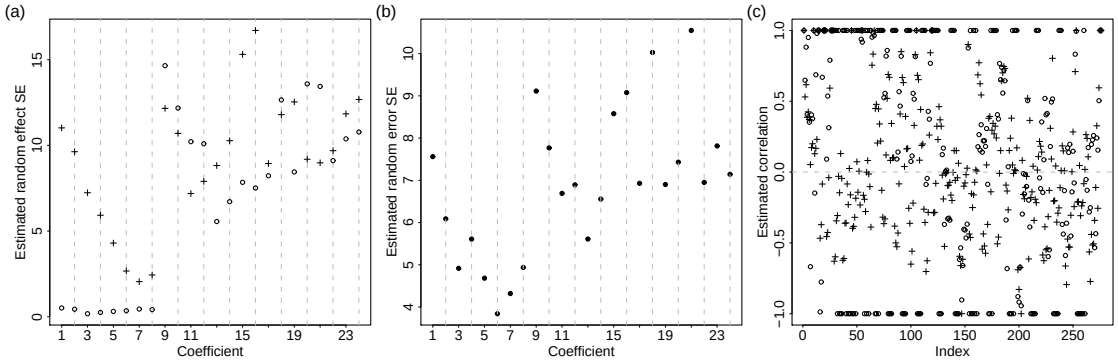

Figure 6.14: (a) Random intercept SE estimates, by spline coefficient, for the cleft (plusses) and control (circles) groups; (b) Random error SE estimates for each coefficient; (c) Estimated correlations between the random effects for pairs of coefficients, for the cleft (plusses) and control (circles) groups and in no particular order

the different curves and the fitting of the pairwise models to each separately. This is clearly not an appropriate assumption as independence between different parts of the face is highly unlikely. This chapter proceeds, therefore, by instead employing the method of principal components analysis in order to reduce the dimension of the curve data.

6.4 Principal components of facial curves

The methodology behind principal components analysis and the application to shape data is outlined in Section 5.5. The full set of curves is displayed in Figure 6.1(b), with examples of randomly chosen cleft and control individuals in Figure 6.2. Due to the high correlation in the curve data, it is possible to substantially reduce the dimension using PCA and thereby analyse the full set of curves using the pairwise approach.

The curves were aligned using GPA in the same way as for the B-spline example of the previous section, except in this case all of the anatomical landmarks highlighted by solid black circles in Figure 6.1(b) were included in the alignment. The resulting Procrustes mean configurations for each group and time combination are displayed in Figure 6.15 in the profile and frontal views.

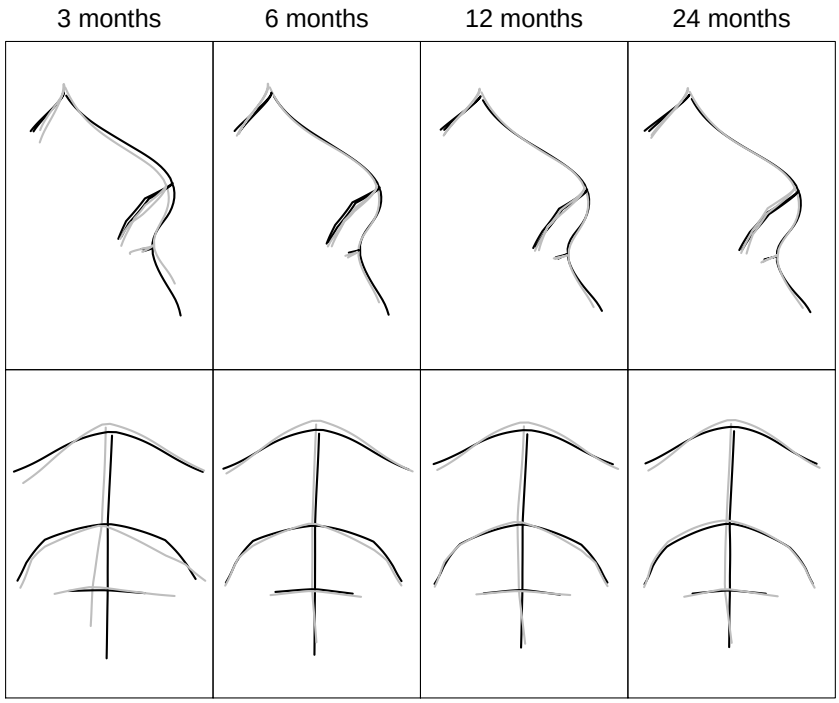

Figure 6.15: Interpolated mean curves for each group at 3, 6, 12 and 24 months for the cleft (grey) and control (black) groups. Upper - profile view; lower - frontal view.

As in Section 5.5, the tangent coordinates of the Procrustes aligned curves were approximated by the Procrustes residuals and the sample covariance of these obtained. The standardised PC scores were calculated from the eigenvalues and eigenvectors of this matrix, as in (5.2). Figure 6.16 gives the mean group trends over time of the PC scores and suggests that a saturated model is necessary for the group means, since assumptions of linearity of the time trends made in previous models do not hold here. The model is therefore defined, for $c_{ir}(t)$ the r^{th} PC score for individual i at time t, as

$$\begin{aligned} c_{ir}(t) = \beta_{0r} + b_{irg_i} &+ \beta_{1r} g_i + \beta_{2r} t_6 + \beta_{3r} t_{12} + \beta_{4r} t_{24} \\ &+ \beta_{5r} g_i \cdot t_6 + \beta_{6r} g_i \cdot t_{12} + \beta_{7r} g_i \cdot t_{24} + \epsilon_{ir}(t), \end{aligned} \tag{6.2}$$

where t_6, t_{12} and t_{24} are indicators which take value one if $t = 6$, 12 and 24, respectively, and zero otherwise allowing for discrete, rather than continuous, time effects. The variance terms are distributed as in (6.1) earlier in this chapter, with score-specific random intercepts

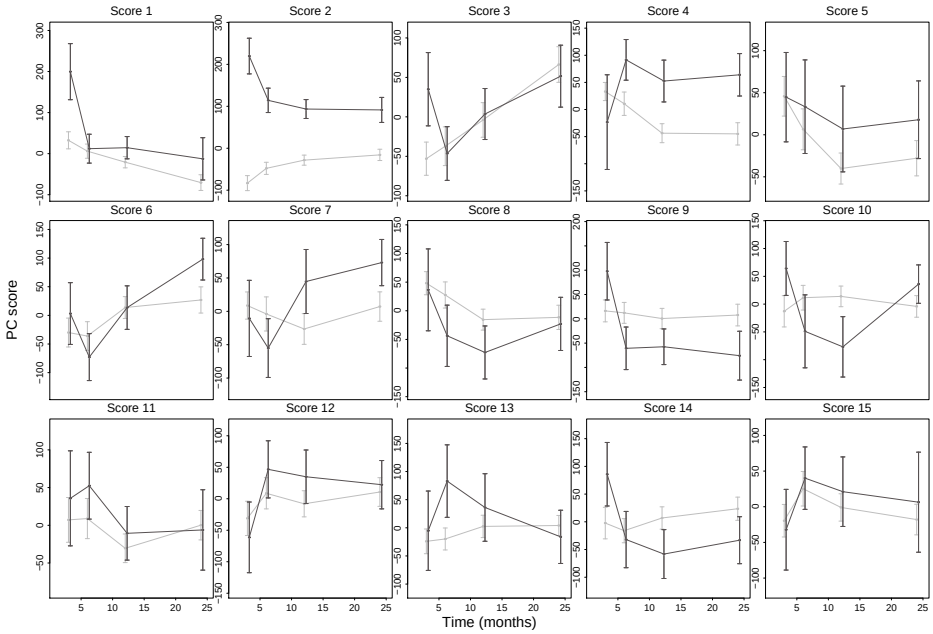

Figure 6.16: Mean principal component score trends over time, with 95% confidence intervals, for the cleft (black lines) and control (grey lines) groups (note the different scales)

that are correlated across PC scores and allowed to differ across the two groups.

The group means at each time point with bivariate 95% confidence bands are displayed in Figure 6.17. These plots give the most detailed analysis of the differences between the groups and their trends over time. Some effects were observed in previous analyses, such as the asymmetry in the midline, the shorter upper lip and flatter nose and higher midpoint between the eyes of the average cleft patient. This analysis also identifies the primary differences between the groups out to 24 months as being in the entire bridge of the nose, with the position of the eyes being depressed amongst the clefts and the midpoint being higher; and shows that the rim of the nose for the cleft patients, whilst being substantially different to that of the controls prior to surgery, becomes very similar by 24 months, as does the length of the upper lip although there are still some areas of minor difference, particularly in terms of asymmetry between the base of the nose and the upper lip. One area that this analysis misses is around the nostrils (although this would be entirely possible had these been marked onto the images as curves). However, this area of the face could be added into the curve analysis as landmarks.

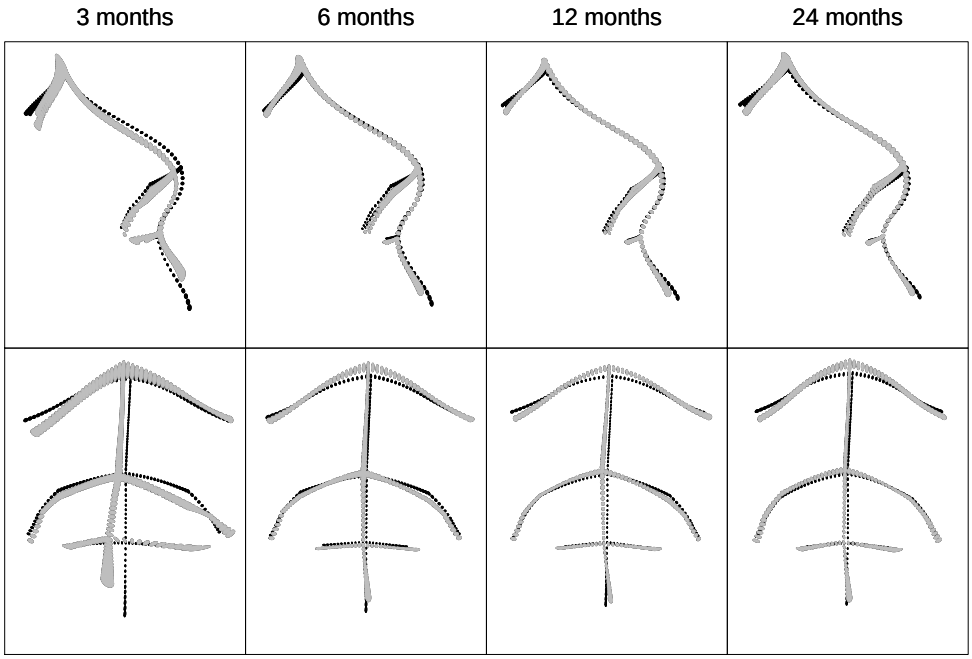

Figure 6.17: Model estimates for the full set of curves, with 95% bivariate confidence bands for cleft (grey) and control (black) groups for the mean midline curves at each time point in the profile (upper) and frontal (lower) views

Figure 6.18 gives the model estimates with 95% bivariate confidence bands superimposed onto the empirical mean curves. The plots illustrating the model fit of the PC scores have been omitted here since they show an extremely good fit and Figure 6.18 is rather more informative. The figure shows that for all group and time combinations, and in both the profile and frontal views, the model fits extremely well to the empirical mean curves.

The residuals versus fitted values are displayed in Figure 6.19 and show a good fit of the model covariance structure for all of the scores except one and nine. As with the spline coefficient example, these high variance residuals arise for the cleft group at three months and are largely due to the small sample size of this group. Clearly the model fit could be improved for these scores, but that would involve extra complication in the model for the sake of a minority of outcomes and for one subset of the data. It seems unlikely that this would have a substantial effect on the overall results, especially given the small sample size.

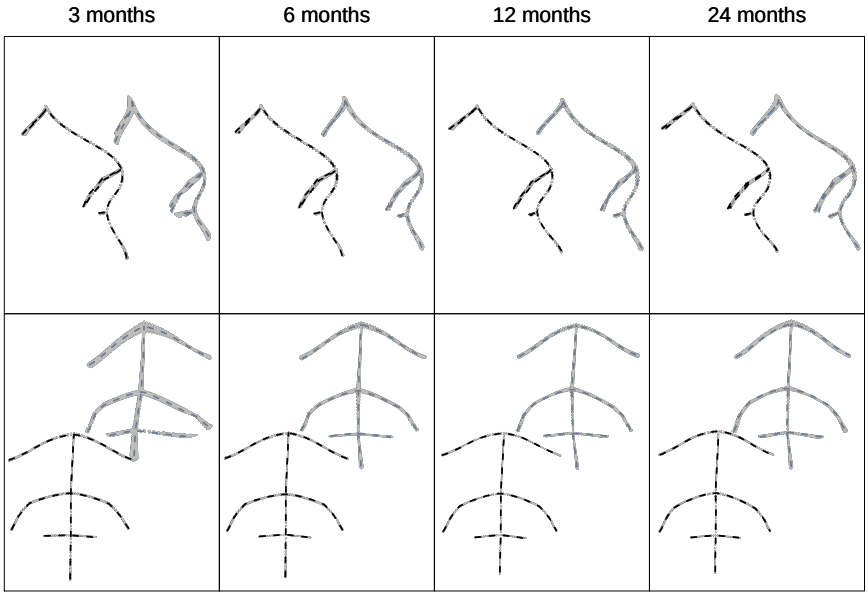

Figure 6.18: Model estimates for the full set of curves and 95% bivariate confidence bands superimposed onto the empirical mean curves for the cleft (grey confidence bands, dark grey dashed lines) and control (grey confidence bands, black dashed lines) groups, at each time point in the profile (upper) and frontal (lower) views. The group curves have been shifted apart for ease of viewing.

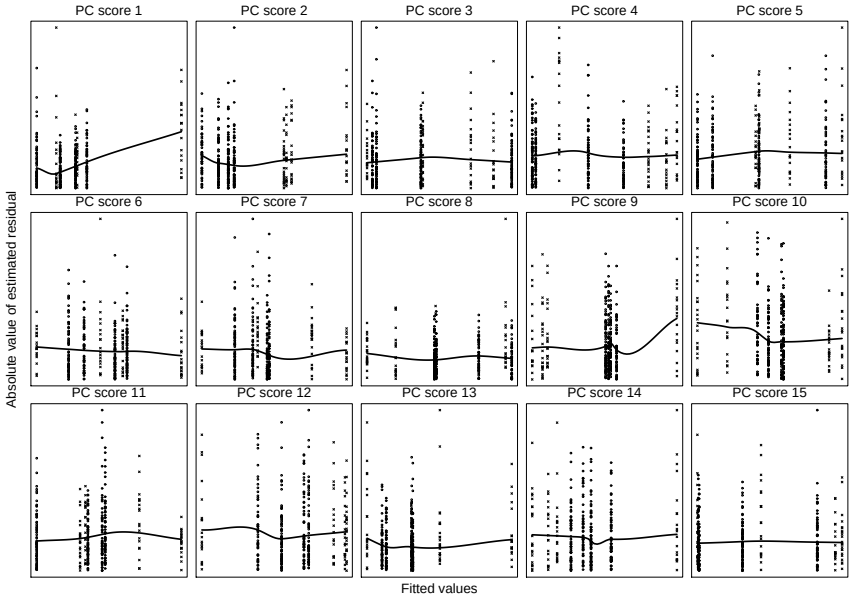

Figure 6.19: Plots of the absolute values of the (unstandardised) residuals versus the fitted values for each PC score, with a loess smooth line fitted

Figure 6.20 shows the estimated random effects variances and correlations from the model. The figure shows that there are now no correlations being estimated as close to one and no SE estimates close to zero, suggesting that this model provides a more suitable fit of the covariance structure of the data.

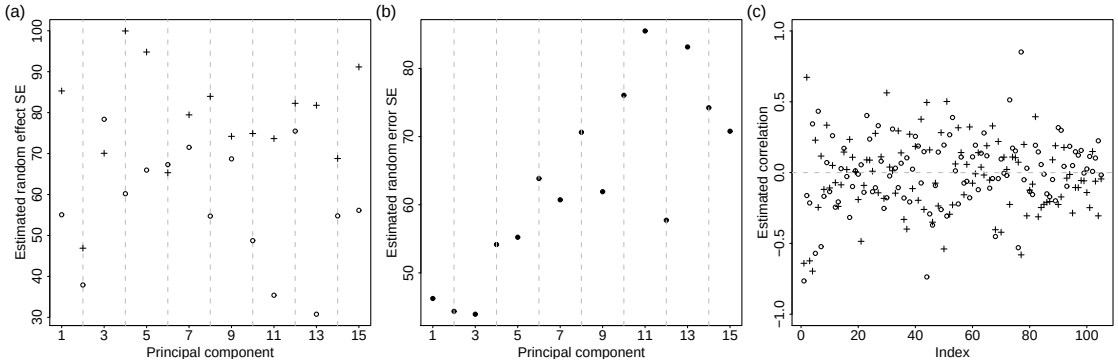

Figure 6.20: (a) Random intercept SE estimates, by PC score, for the cleft (plusses) and control (circles) groups; (b) Random error SE estimates for each PC score; (c) Estimated correlations between the random effects of pairwise combinations of PC scores for the cleft (plusses) and control (circles) groups

6.5 Discussion

This chapter has involved the extension of the pairwise approach to curve data. B-splines and principal components analysis (PCA) were employed to reduce the dimension of the data and, as in the case of the landmarks, the most effective method of doing so is PCA. The B-spline approach gave useful results in terms of midline curves, but these did not fit particularly closely to all curves, and in particular missed some of the more extreme curvature amongst the controls at earlier time points. Increasing the number of knots would probably have reduced the discrepancies, but that is liable to result in convergence and memory problems since 30 outcomes is around the limit that the method can cope with, and this tends to reduce when fitting more complicated models including effects such as a saturated mean, different random effects variances for each group or different random error variances for each outcome. Furthermore, the B-spline approach could not be readily applied to the full set of curves without some restrictive assumptions about the covariance structure.

The PCA approach allowed the full set of curves to be analysed since the high correlation within faces results in a small number of PCs describing a large extent of the variance in the data. This approach was substantially quicker computationally than using B-splines, even when the latter was applied to the midline curve only. The results showed a closer fit to the data in shape space than the B-splines and no problems with the estimated random effects variances or correlations. The model covariance structure was shown to be a reasonable fit of the data, with a small number of high residuals for a particular subset of individuals which were largely a result of small sample size. Given the low number of PCs required to describe a set of curves with 197 points per face, this approach could easily be extended to combinations of landmarks and curves, or to surfaces.

Chapter 7

Model Comparison

7.1 Introduction

The graphical presentation of results from the pairwise models in terms of bivariate confidence regions, as in the previous chapters, is a useful guide as to the differences between groups and their trends over time, but does not allow any definitive inferences. Comparison between the models fitted and particular submodels could answer questions such as whether there is a difference between the group trends (i.e. slopes over time) or whether there is a significant surgery effect. The answer to a question such as the former could be found by testing the null hypothesis that the relevant group by time interactions are equal to zero and this involves simultaneously testing this hypothesis for these parameters for each of the coordinates. Clearly this can lead to the existence of a large number of effects being tested at once. An added complication is that the model is constructed by fitting a set of pairwise models and averaged so we obtain a pseudo-likelihood rather than a proper likelihood.

Various model comparison tests are considered and simulations carried out to verify their effectiveness when testing for the existence of a large number of effects, with and without distributional assumptions. A Wald test may be applied using the vector of averaged parameter estimates, θ^*, where in Section 3.3 it was shown that asymptotically

$$\hat{\theta}^* \sim MVN(\theta^*, V),$$

where $V = AJ^{-1}KJ^{-1}A'/N$. Hence to test the null hypothesis $H_0 : \theta_S^* = 0$, where θ_S^* is a p^*-dimensional subvector of θ^* and the $p^* \times p^*$ matrix V_S contains the corresponding covariance parameters of V we assume that $\hat{\theta}_S^* \sim MVN(\theta_S^*, V_S)$. It follows that (Verbeke & Molenberghs, 2000)

$$\hat{\theta}_S^{*\prime} V_S^{-1} \hat{\theta}_S^* \sim \chi^2_{p^*} \tag{7.1}$$

where p^* is the degrees of freedom. The Wald test is not always appropriate in the case of random effects models because of the way in which a model is specified conditionally upon the random effects b, as $Y|b \sim N(X\beta + Zb, \Sigma)$, where Z and X are matrices of coefficients, b is assumed normally distributed with mean zero and covariance matrix D and Σ is the covariance matrix of the the random errors. Inference, however, is based on the marginal model, $Y \sim N(X\beta, Z'DZ + \Sigma)$ (Verbeke & Molenberghs, 2000). This can result in problems applying the Wald test to random effects models, since it can be sensitive to the parameterisation of the model (Geys et al., 1999) and the consequence of the conditional specification is that the marginal means may have a complex structure that does not correspond to that of a simple regression (Diggle et al., 2002). The other aspect is that the Wald test does not account for the variation in estimating the random effects parameters, and therefore may underestimate the variation in the fixed effects (Verbeke & Molenberghs, 2000).

In the case of a model without multivariate outcomes a likelihood ratio (LR) test may be easily applied, such that $2[l(\theta) - l(\theta_0)] \sim \chi^2_{p^*}$, where $l(\cdot)$ is a log-likelihood (maximised using maximum likelihood), θ and θ_0 are the vectors of parameters from the full and sub-models, respectively, and p^* is the difference in the number of parameters (or the number that are zero under the null hypothesis) (Verbeke & Molenberghs, 2000). For tests of the existence of large numbers of parameters, however, the LR test is likely to be anti-conservative and the extent of this increases with the number of parameters being tested (Pinheiro & Bates, 2000). Faraway (1997) explains that the LR statistic depends on the log of the ratio of eigenvalues of the covariance matrices from the full and sub-models, and that this ratio does not become small as one progresses through a high number of eigenvalues. Therefore the LR statistic can be highly influenced by terms that represent directions of variation which are

not important. Clearly this is a disadvantage in a situation where there is a large number of outcomes.

One could argue that since we are not dealing with a proper log-likelihood, but rather with a pseudo-log-likelihood formed by summing the log-likelihoods for all of the pairwise models, then neither the Wald nor the LR tests are suitable. An alternative approach is to calculate the adjusted pseudolikelihood ratio (PLR) statistic

$$G_a^{*2} = 2[pl(\hat{\theta}) - pl(\hat{\theta_0})]/\bar{\lambda},$$

where $pl(\cdot)$ is the log of the pseudo-likelihood (Geys et al., 1999). The pseudo-log-likelihoods may be calculated by summing the pairwise log-likelihoods. Under the null hypothesis H_0 : $\gamma = 0$, where γ is a p-subvector of the vector of regression parameters, θ, and θ_0 is the remaining subvector containing the parameters of θ that are included in the submodel, the test statistic, G_a^{*2}, approximately follows a χ_p^2 distribution (Geys et al., 1999). The denominator of the test statistic, $\bar{\lambda}$, is the mean of the eigenvalues of $(J_0)^{-1}\Sigma_0$, where J_0 and Σ_0 are $p \times p$ submatrices of J^{-1} and $J^{-1}KJ^{-1}$, respectively (where, as defined in Section 3.3, J and K are matrices containing the second and first derivatives of the log-likelihood). These are evaluated under the null hypothesis, as Geys et al. (1999) found that if evaluated under the alternative, the power may be misleadingly high.

Recall that θ is the long vector of parameters, containing all repetitions and used here because the derivatives in J and K are necessarily calculated from all the pairwise models and therefore correspond to θ, rather than to θ^*. This results in a test of a larger number of parameters, p, than for the Wald or LR tests although, technically, these are not all different parameters but multiple copies. The number of parameters of interest under the null hypothesis is denoted here by p^* and arises from the short parameter vector θ^*, whereas the value p represents the number of copies of all the relevant parameters which are included in θ. It is unclear as to how the test could be modified to account for this issue of multiple copies and strictly test for each of the relevant parameters being equal to zero, rather than for multiple versions of the same parameters being zero, since the off-diagonal entries in the matrices are based upon pairs of parameters and not individual ones. The PLR test may

also suffer the same problems as the ordinary LR test in terms of being anti-conservative when there is a large number of outcomes.

An alternative way of carrying out the LR test following the pairwise approach could be to calculate the LR statistic by evaluating the pairwise log-likelihoods at the averaged parameter estimates in θ^* rather than obtaining the values for the estimates from each individual model. This removes the issue of evaluating the statistic using multiple estimates of the same parameter, although it still may be unduly affected by the large number of outcomes and the log-likelihood must still be formed by summing all of the pairwise log-likelihoods.

A method which makes no distributional assumptions, in contrast to those mentioned above, was suggested by Faraway (1997). The author presents a bootstrap-based method that uses the sum, as opposed to the ratio, of the eigenvalues of the difference between the covariance matrices for the full and submodels, so that the greatest differences will dominate rather than those in unimportant directions. The test statistic is defined as

$$T = \frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{M}(\hat{y}_{ij}^{\omega} - \hat{y}_{ij}^{\Omega})^2 \rightarrow \sum_{i=1}^{n}\int(\hat{y}_{i}^{\omega}(d) - \hat{y}_{i}^{\Omega}(d))^2 \text{ as } M \rightarrow \infty, \tag{7.2}$$

where y_{ij} is the value of the j^{th} point ($j = 1, \ldots, M$) on the curve for individual $i = 1, \ldots, n$. The responses for each point may be estimated under the full model as $\hat{y}_{ij}^{\Omega}$ and under the null model as $\hat{y}_{ij}^{\omega}$. In the limit, the curve is parameterised as $y(d)$, for $d \in [0, 1]$. Faraway (1997) sets out a bootstrapping procedure wherein the test statistic, T, is calculated along with the estimated residuals under the null hypothesis, $\hat{\epsilon}_{i}^{\omega}(d) = y_{i}^{\omega}(d) - \hat{y}_{i}^{\omega}(d)$. Resampling with replacement is carried out on these residuals and a new set of responses formed as $y^* = \hat{y}^{\omega} + \hat{\epsilon}^{\omega}$, where the covariates match the estimated responses. The models are refitted and the new test statistic, T^*, obtained. This is carried out B times and the p-value calculated as $\#\{T^* > T\}/B$.

This chapter proceeds by applying all of the methods discussed and comparing their performances using simulations. Finally, a model comparison test will be considered to test for differences in the time trends for the two groups in terms of their full set of facial curves, using the model fitted to the principal component scores in Section 6.4.

7.2 A simulation study

7.2.1 Five outcomes

In order to calibrate the performance of the various model comparison tests mentioned above on high-dimensional longitudinal data, such data were simulated from the conditional model $Y|b \sim N(X\beta + Zb, \Sigma)$, following the simulation of the random effects, b, from $N(0, D)$. The mean part of the model and the random effects variances were specified as in (6.2), while the random error variance was assumed to be pair-specific for computational reasons (which will be discussed in more detail later).

The test is that both groups have the same time trend, so under the null hypothesis all of the group by time interaction terms are equal to zero, and for each coordinate there are three such parameters. Two original sets of data were simulated: (A) from the full model, including a group by time interaction term for each time point; and (B) from the null model, which has the group by time interaction terms removed. Therefore, under scenario (A) the true model corresponds to the alternative hypothesis and under (B) the true model corresponds to the null. In each case, the data were simulated for 200 individuals at four time points. To compare against the performance of the tests on the fully joint model, initially the individuals had responses simulated from a multivariate normal distribution for five outcomes only so that both the fully joint and the pairwise models could be fitted.

For each of the scenarios, the full (under the alternative) and the null (under the null) models were fitted, using both the fully joint and pairwise approaches. The bootstrap approach proposed by Faraway (1997) and outlined above was used to obtain null distributions for each of the test statistics by resampling residuals from the fitted null model in each scenario. The type I error can therefore be calculated for the tests which have distributional assumptions by finding the proportion of the bootstrapped test statistics that exceed the 95% quantile for the relevant distribution in each case. The parametrically obtained p-value (for the tests with distributional assumptions) can be compared with the nonparametric proportion of test statistics greater than the original for each scenario.

The vector of estimated responses for each individual, $\hat{y}_i$, which must be calculated for the bootstrapping, could be computed at the individual or population (group) levels, the former

by adding on the estimated random effects and the latter using only the fixed effects. Since the fixed effects are of primary interest here and the random effects estimated only so that they are taken account of, the estimated responses were taken to be the population level estimates, dependent on each individual's particular covariates. This results in the random effects variance being incorporated into the estimated random error term in the resampling, rather than in the estimated response. When the Faraway (1997) test statistic in (7.2) is computed, the estimated random effects are likely to be reasonably similar under the full and null models and so should make little difference to the value of the statistic.

Two hundred bootstrap datasets were produced from the original simulated datasets of each of the two scenarios, and the test statistics calculated under the fully joint and pairwise approaches for the Wald and LR tests, and just for the pairwise for the PLR and Faraway tests. The test statistic was also produced for the LR test using the average estimates of the parameters in θ^*, resulting from the pairwise approach, and this is defined as the LR* test. The number of parameters equal to zero under the null hypothesis is $p^* = 5 \times 3 = 15$ for tests using the averaged parameters and $p = 60$ for tests using all of the individual parameter estimates.

The test statistics for each of the model comparison tests are given in Table 7.1 for the original datasets simulated under both the null and alternative hypotheses. Figures 7.1 and 7.2 give the null distributions of the test statistics arising from the bootstrap simulations from the null model under each of the two scenarios, with the actual test statistics from the original simulated datasets marked onto the figures where appropriate. The distributions in the two figures are very similar, which is to be expected as they are both representing the null distributions for each test statistic (and the only marked difference should be in the original test statistic). For the dataset simulated under the alternative hypothesis (scenario A), comparing the test statistics in the first row of Table 7.1 against Figure 7.1, all of the tests correctly reject the null hypothesis whether using distributional assumptions or calculating the p-values as the proportion of bootstrapped test statistics that exceed the original, and regardless of whether the fully joint or pairwise models were fitted. For the data simulated under the null hypothesis (scenario B and test statistics in second row of Table 7.1), all of the tests using distributional assumptions incorrectly reject the null hypothesis, although

the LR* test correctly accepts the null hypothesis using the bootstrapped test statistics as a reference distribution. This suggests that the assumption of this statistic following the χ^2 distribution with p (or p^*) degrees of freedom is false. The Faraway test correctly accepts the null hypothesis based upon the bootstrapped samples.

	Wald		LR				
	Joint	PW	Joint	PW	PLR	LR*	Faraway
Original	**Test statistic**						
(A) Alternative	463	473	430	1730	3372	1374	35982
(B) Null	32	38	32	129	303	105	2525
	Type I error						
(A) Alternative	0.070	0.135	0.090	0.200	0.210	0.815	-
(B) Null	0.040	0.070	0.045	0.220	0.210	0.820	-
	Proportion of test statistics > original (empirical p-value)						
(A) Alternative	0	0	0	0	0	0	0
(B) Null	0	0	0.005	0.005	0	0.625	0.600

Table 7.1: For the datasets simulated under the alternative and null hypotheses (with and without the group by time interactions, respectively): Test statistics from the fully joint and pairwise (PW) models for the Wald and likelihood ratio (LR) tests, and from the pairwise model only for the pseudo-likelihood ratio (PLR), LR* (likelihood ratio using the final, averaged parameter estimates) and Faraway tests; Type I error rates from the bootstrap simulations for the tests with distributional assumptions; and empirical p-values using the proportion of bootstrap test statistics that are greater than the original.

There are some important discrepancies in test performance. The Wald test shows similar performance for both the fully joint and pairwise models, but has slightly higher type I error for the latter. The bootstrap simulations arising from the original data simulated under the alternative hypothesis showed inflated type I error for both fully joint and pairwise, with the latter being particularly high at almost 14%. Despite the type I error being almost correct for the fully joint model when the data are simulated under the null hypothesis, this hypothesis is still incorrectly rejected for the original dataset, based upon both the distributional assumptions and the bootstrap sample. It could be argued that this is an outlier, but the Faraway test correctly accepts the null hypothesis in this case, which suggests that the Wald test is in error.

There is close correlation between the test statistics for the LR test under the fully joint and pairwise models. However this correlation does not follow the line of equality as the

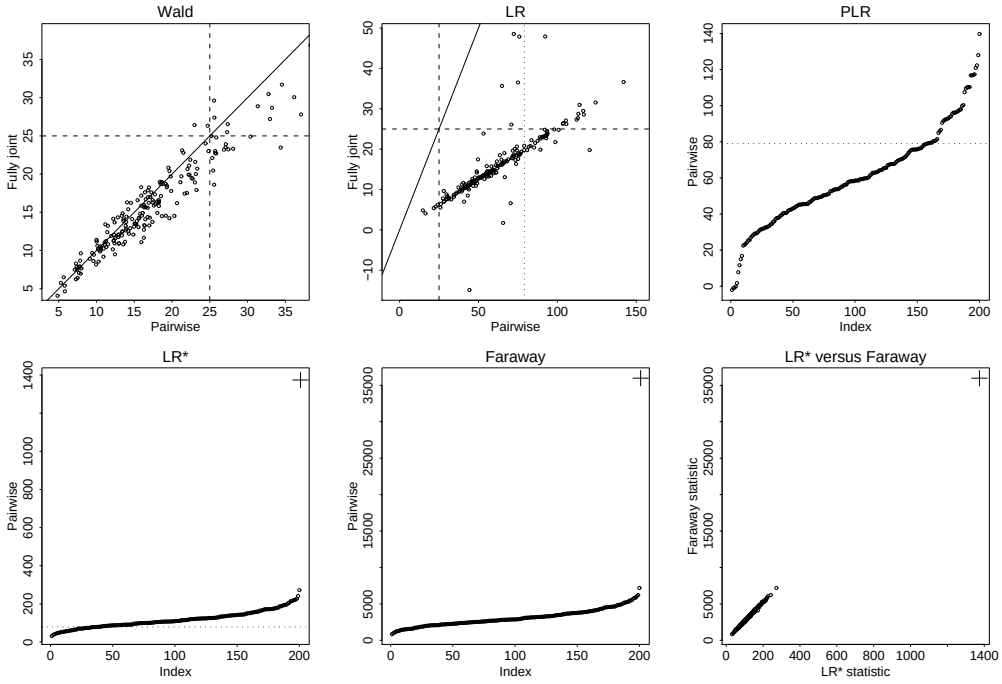

Figure 7.1: Bootstrap distributions of test statistics from the Wald, LR, PLR, LR* and Faraway tests arising from an original dataset simulated under the alternative hypothesis (scenario A). Lower right - plot of LR* versus Faraway test statistics. Original test statistics are displayed by a large cross when they are not too far from the distributions. The Wald and LR tests have been calculated under the fully joint and pairwise models and the statistics from the other tests have been ordered in terms of size. The tests which make distributional assumptions (Wald, LR, PLR, LR*) have lines marking the 95% quantile for the assumed null distribution, where the dotted and dashed lines correspond to $p^* = 60$ and $p = 15$ degrees of freedom, respectively. Diagonal lines, where present, mark the line of equality.

pairwise model produces test statistics that are substantially higher than for the fully joint. If the type I error is calculated using the χ^2 cutoff with p^* degrees of freedom, it is close to one; but even if the degrees of freedom is taken as p, the type I error is around 20%. This result is in line with the comments made in the introduction of this chapter regarding the performance of the LR test with multiple outcomes and suggests that the use of the pairwise models exacerbates the issue.

The PLR test also shows anticonservative behaviour, with a type I error of 21% for the bootstrapped data resulting from original datasets simulated under both the null and alternative hypotheses. The test also rejects the null hypothesis in both cases. On further examination, the set of eigenvalues of $(J_0)^{-1}\Sigma_0$ which, when averaged, becomes the denominator of the test

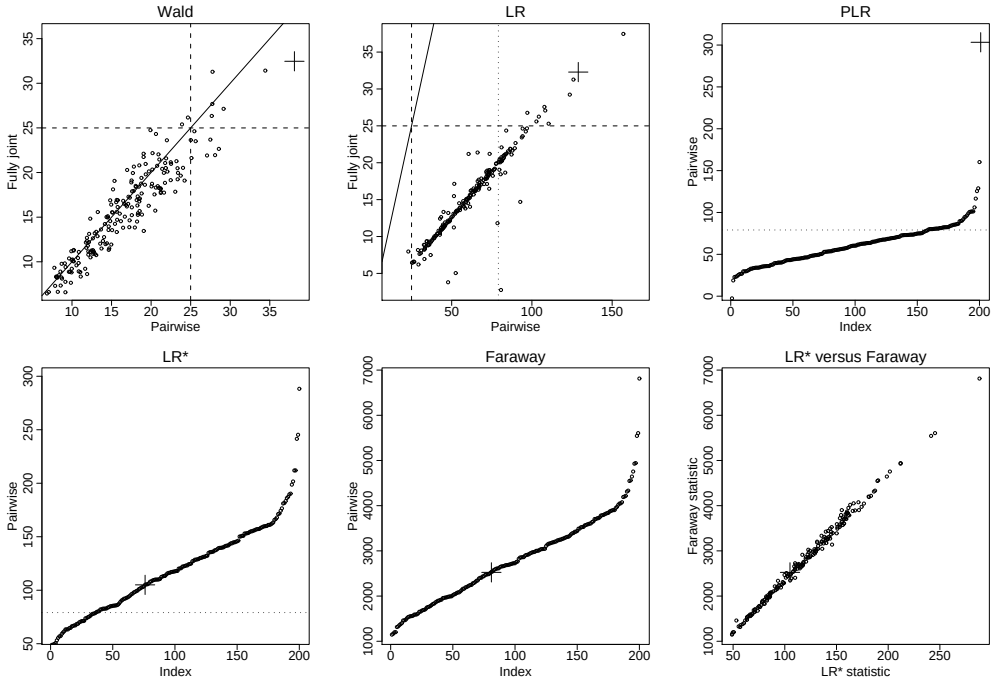

Figure 7.2: Bootstrap distributions of test statistics from the Wald, LR, PLR, LR* and Faraway tests with the corresponding statistic from an original dataset simulated under the null hypothesis (scenario B) marked by a large cross. Lower right - plot of LR* versus Faraway test statistics. The Wald and LR test statistics have been calculated under the fully joint and pairwise models. The tests which make distributional assumptions (Wald, LR, PLR, LR*) have lines marking the 95% quantile for the assumed null distribution, where the dashed and dotted lines correspond to $p^* = 60$ and $p = 15$ degrees of freedom, respectively. Diagonal lines, where present, mark the line of equality.

statistic for the PLR test, tends to contain a large number of very small eigenvalues which artificially skew the test statistic so that it becomes large, despite these eigenvalues being of little importance. This seems to correspond with the comments of Faraway (1997), who noted that with multiple outcomes the LR test may be unduly influenced by unimportant directions. There may also be some effect of the multiple estimates of each parameter, as previously mentioned. Taking the average of subsets of the eigenvalues substantially alters the test statistic, but clearly there is no rigourous way of selecting the size of such a subset.

There is close correlation between the test statistics from the LR* and Faraway tests, as demonstrated by the lower right plots of Figures 7.1 and 7.2. Clearly the assumption of the LR* statistic following a χ^2_p distribution does not hold, with a type I error of around 82%,

but comparison of the original test statistic with the bootstrap reference distribution gives the correct result in both cases, in line with the Faraway test.

These simulations therefore suggest that the only tests that are appropriate to apply to the results of the pairwise models for multiple outcomes are the Faraway and LR* tests. The Wald, LR and PLR tests all show anticonservative behaviour even for such a small number of outcomes whereas, at least for five outcomes, the other two tests correctly reject both the null and alternative hypotheses based on bootstrap simulations.

7.2.2 Fifteen outcomes

We carried out a further simulation study, allowing each individual to have responses on fifteen outcomes to investigate whether the test results were affected by increasing the number of outcomes. Clearly there can be no comparison between the fully joint and pairwise models in this case. The data were again simulated under the null and alternative hypotheses, with both being similar models to those of the previous section; and the only differences being the additional outcomes. Here 100 bootstrap simulations were carried out since the additional outcomes meant that the computations were much more time-consuming.

The simulated test statistics are displayed in Figures 7.3 and 7.4, with the originals from the datasets simulated under the alternative and null models (respectively) marked with crosses (the original PLR test statistics are not marked on the plots because at 786758 for the dataset from the full model and 63888 for the null model they are very high relative to the others). In this case, all five tests correctly reject the null hypothesis and all but the PLR test correctly accept the null hypothesis. The PLR test shows an even more extreme test statistic for the original dataset simulated under the null hypothesis, relative to the null distribution, than with five outcomes which suggests that this tests performs more poorly as the number of outcomes increases. This test also shows some negative values. It is unclear why these have occurred, but it is possible that the eigenvalues contain negative values for some simulated datasets and if there are many of these then the mean may be negative. Clearly this is undesirable and again points to the influence of unimportant directions of variation.

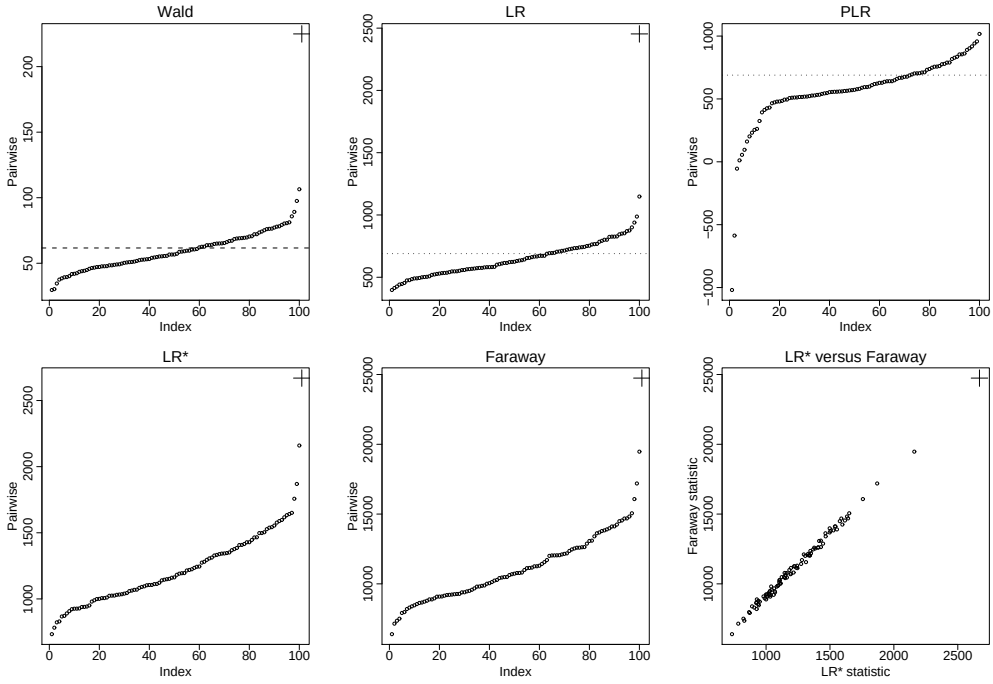

Figure 7.3: Bootstrap distributions of test statistics from the Wald, LR, PLR, LR* and Faraway tests with the corresponding statistic from an original dataset simulated under the alternative hypothesis (marked by a large cross). The tests that make distributional assumptions (Wald, LR, PLR) have lines marking the 95% quantile for the assumed null distribution, where the dashed and dotted lines correspond to $p^* = 45$ and $p = 630$ degrees of freedom, respectively. Lower right - plot of LR* versus Faraway test statistics

The Wald and LR tests do give the correct result for the original datasets simulated under each scenario, but both tests are very anticonservative, with type I error rates of approximately 0.42 for the Wald and 0.36 for the LR under both models. The test statistics obtained from the LR* and Faraway tests indicate that the original dataset from the null hypothesis is at the opposite end of the spectrum to the one of the previous section in terms of the differences in time trends between groups, and that seems likely to be the reason for the Wald and LR tests correctly accepting the null hypothesis in this case.

7.3 Application to the cleft-lip and palate data

Since the LR* and Faraway tests gave consistently correct results for the simulated data, these will be employed to compare models fitted to the cleft-lip and palate data. The tests

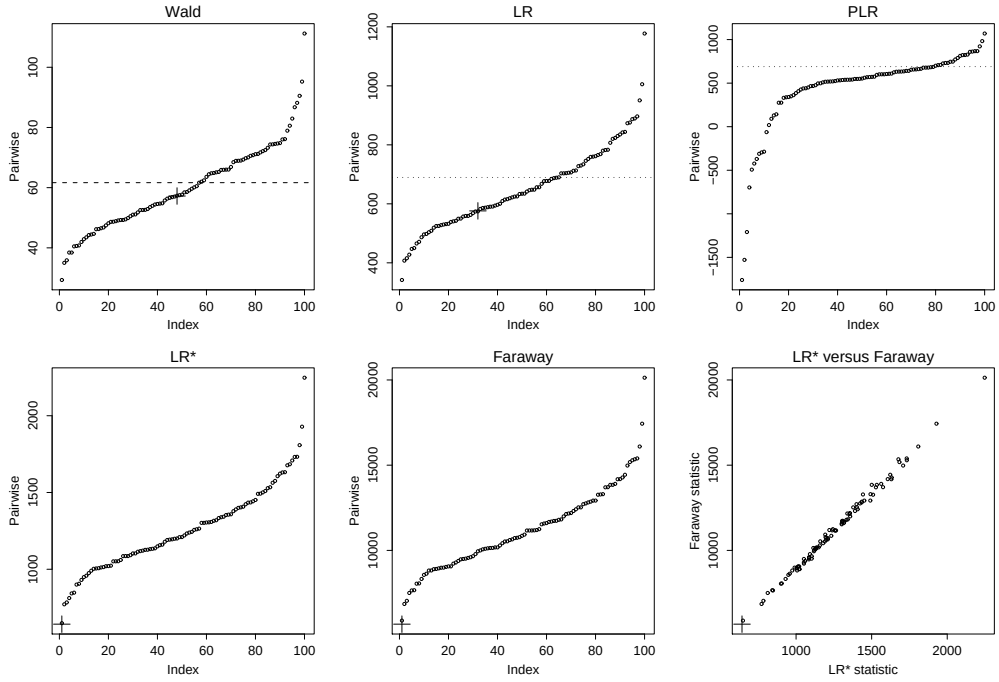

Figure 7.4: Bootstrap distributions of test statistics from the Wald, LR, PLR, LR* and Faraway tests with the corresponding statistic from an original dataset simulated under the null hypothesis (marked by a large cross). The tests that make distributional assumptions (Wald, LR, PLR) have lines marking the 95% quantile for the assumed null distribution, where the dashed and dotted lines correspond to $p^* = 45$ and $p = 630$ degrees of freedom, respectively. Lower right - plot of LR* versus Faraway test statistics

are carried out here for the example in Section 6.4, where principal components analysis was used to reduce the dimension of the curve data and the resulting 15 PC scores taken as the responses for the model. The tests could equally be carried out for the models fitted to the landmark data (and similar conclusions would be expected).

Since neither test makes distributional assumptions, for each it is necessary to obtain a bootstrap reference distribution with which to compare the test statistic. This is performed in the same way as for the simulated data, so that the fitted values for each individual are computed under the null model, and from them the estimated residuals. The residuals are resampled with replacement and a new set of responses formed from the original fitted values and the resampled residuals. The primary interest here is whether the time trends for the two groups are different from six months onwards. There is clearly a qualitative difference between the groups at three months and an obvious change from this point to six months in

the case of the clefts, but it is less clear as to whether there is a significant difference in the way the facial shapes of the two groups change on average as they grow following the cleft group surgery. Since in (6.2) in Section 6.4 the mean part of the model is saturated, there are two parameters of interest for each coordinate. These correspond to the $g_i \cdot t_{12}$ and $g_i \cdot t_{24}$ effects, where t_j is a dummy variable representing time point j and g_i represents the group of individual i. The $g_i \cdot t_6$ parameter is not included because this represents the difference between the two groups in the change from three to six months, which is not of primary interest. Note that the test of the existence of a non-zero interaction between group and time does not answer the question of whether any differences remain between the groups at 24 months, but is a test of any differences in the way the groups change as they grow from 6 to 24 months. This is therefore a test of differences in slopes, but does not assume the same intercepts or the same group positions at 6 months.

The full model fitted to the PC scores here is not identical to (6.2). Where each coordinate in that model was allowed to have a different random error variance, here the variance is pair-specific. The reason for this is that to allow coordinate-specific random error variance, at present one must employ the `lme` function in `R`, whereas `lmer` may be used for pair-specific variance. The fitting process of `lme` seems to be considerably less stable than that of `lmer` (at least for these models) and as a result the pairwise model fitting often fails if using `lme`, with obscure error messages. Since repeated bootstrap-simulated datasets must be modelled using the pairwise approach in order to carry out the model comparison tests, this makes `lme` virtually impossible to use for this procedure. Assuming a pair- rather than coordinate-specific random error variance should not make major differences to the fit of the model since the random effects variances (which are coordinate and group-specific) should absorb some of the extra variation.

The full and null models were fitted to the PC scores and the Faraway and LR* test statistics calculated as 1805491 and 6476, respectively. The fitted values from the model fitted under the null hypothesis were used to produce 150 bootstrapped datasets, with the resulting null distribution of Faraway statistics ranging from 490424 to 1296363 and the LR* statistics from 1582 to 3971. In both cases, therefore, the proportion of bootstrapped test statistics greater than the original is equal to zero and there is strong evidence to reject the null hypothesis.

Thus the evidence suggests that the average time trends from six to twenty-four months are not the same for the cleft and control groups. This is to be expected since the plots of bivariate confidence regions in Chapters 5 and 6 suggest that on average the cleft individuals change more than the controls, even after the clefts undergo surgery.

7.4 Discussion

Comparison of nested models is not straightforward with high-dimensional data and the fitting of the models using the pairwise approach further complicates matters. The comparison of models fitted to high-dimensional data by definition involves a null hypothesis concerning multiple parameters for which there are certain suitable methods, such as the Wald test, and appropriate concessions that can be made, such as Bonferroni corrections of the confidence level. However, methods such as the Wald test are not necessarily suitable for use with mixed effects models because of the way they are specified and the extra variance involved in estimating the random effects. Other methods, such as the likelihood ratio test, suffer similar problems in comparing models for multivariate data when the number of outcomes is high, which is generally the case with shape data.

It is also not immediately clear how the usual application of tests such as the Wald and LR may be modified to use alongside the pairwise approach. The obvious problem is that many parameters are estimated multiple times, whereas some are estimated only once, but it is not as clear how this issue manifests itself in the application of model comparison tests. Clearly, from the models fitted to the simulated data, there are substantial differences in the application of the LR test for the fully joint and pairwise models, while the Wald test shows rather less extreme discrepancies. As pointed out by an anonymous reviewer of Barry & Bowman (2008), the test statistics may not in fact be entirely comparable for the two different models, because in using a sandwich estimator the pairwise approach adjusts for possible misspecification of the covariance structure, whereas the fully joint model does not. This may be related to the differences in type I error for the Wald test when applied to the pairwise and fully joint models. However this seems unlikely to be the reason behind the discrepancies in the LR test statistics when applied to the pairwise and fully joint models

since, while the correlation between the test statistics was high, the bootstrap distribution for the pairwise models showed very high type I error. This seems more likely to be related to the fact that the likelihood for the pairwise approach is calculated by summing all of the bivariate likelihoods, which involve multiple estimates of the same parameter.

The pseudo-likelihood ratio test was applied with the view that it might be more appropriate than the LR test in the setting of the pairwise approach, since the sum of the log-likelihoods from the pairwise models is a pseudo-likelihood rather than a true likelihood. However, this test performed very poorly on the simulated data, and it is not clear how to correctly apply the test with multiple estimates of some of the parameters. Possibly some combination of the estimates could be taken, but this still does not explain how to treat the $(J_0)^{-1}\Sigma_0$ matrix and its corresponding eigenvalues in the denominator of the test statistic, since each off-diagonal element of the matrix of second derivatives (J) refers to two parameters. Alternatively the averaged parameter estimates could be used, but then it is unclear how to construct the aforementioned matrices at all, since they would then depend on the full likelihood and this would be substantially more complex. If the matrices are to be constructed based upon the pairs then some parameters must necessarily be included multiple times.

The Faraway and LR* tests performed well on the simulated data, correctly rejecting both the null and alternative hypotheses, and were subsequently successfully used with the cleft-lip and palate data. There is high correlation between the test statistics from these two tests and it is likely that there is an algebraic relationship between them, although that is not of particular interest here. Faraway (1997) mentions that it would be possible to derive the null distribution of his test statistic by making some assumptions, but it is sufficient to use the bootstrapping procedure to empirically obtain the null distribution. One downside of this approach is that the computations are very time consuming, since for each bootstrap simulation the full and null models must be fitted using the pairwise approach and clearly as the number of outcomes increases the computational time becomes large. Another computational issue is that the model allowing different random error variances for each coordinate cannot, at present, be used in the bootstrap simulations due to instability of the `lme` function. Since the `lmer` function is being modified and updated all the time, this should not remain a problem for long.

Chapter 8

Discussion

This thesis has addressed the problem of analysing longitudinal three-dimensional shape data, with an application in the comparison of the facial shapes of cleft-lip and palate children with controls, all of whom had the positions of various points on their faces recorded at 3, 6, 12 and 24 months of age. Much work has been done on the analysis of two-dimensional shape data, including some modelling growth, but three dimensional shapes have been less visited, particularly in terms of change over time. Historically, shapes have generally been made up of anatomical landmarks that are placed at meaningful points and we use this method as well as curves, which consist of landmarks that are placed at small increments along particular arcs of interest, usually joining up anatomical landmarks. This results in a more detailed analysis of the shape, but the problems that arise with landmarks are compounded with curves.

The major issues involved include the alignment of individual shapes so as to remove the effects of location, rotation and scale; the inherent high dimensionality of the data; the correlation between different points in the shape; different levels of variability in certain areas of the shape or in subsets of the measured sample of individuals; summarising curves that are made up of closely located points that are highly correlated; correlation between an individual's measurements over time; how to make definitive inferences about the questions of interest. Each of these will be addressed in turn in this discussion.

When interest lies solely in shape, the effects of location, rotation and scale must be removed from the data. In some cases scale may be left in, but here it is not of interest. The most widely used method in shape analysis is generalised Procrustes analysis (GPA), which involves minimising the sum of squared differences between all of the shapes. In preliminary analyses on the cleft-lip and palate data, involving only four landmarks on the midline of the face, it was found that if the GPA was carried out only on these landmarks, then the asymmetry present in the cleft individuals was artificially removed since the alignment naturally tries to bring the corresponding landmarks on different shapes as close together as possible. This resulted in almost perfect vertical lines down the midline for all individuals, which is clearly not realistic, for the cleft individuals at least. Therefore, in subsequent analyses the entire set of landmarks was used for the alignment, and the particular landmarks of interest were extracted from the resulting aligned shapes for analysis. This seemed to be an appropriate way to proceed, since the landmarks off the midline all correspond with others on the opposite side of the face and should therefore anchor the midline in the correct position, retaining any existing asymmetry.

Since shapes reside in a high-dimensional non-Euclidean space, the standard method to transform the data such that the distances between the shapes become Euclidean is to find the tangent coordinates. This involves projecting the shapes onto a linearised space which is tangent to the shape space at a particular point on its surface, which is usually taken to be the mean shape. The resulting Euclidean distances between shapes are good approximations to the non-Euclidean distances in shape space provided the shapes are reasonably close together in the latter. Given the high dimensionality of the set of landmarks defining facial shapes and their inherent similarities (even if some have a deformity), the assumption of closeness is reasonable in this case. The tangent coordinates may be approximated by the Procrustes residuals, which are computed by subtracting the mean shape. Although approximate tangent coordinates may still contain some variability in directions that their exact equivalents would remove, the former are more straightforward to backtransform and so they are used throughout.

The analyses used these tangent coordinates for each individual to describe their facial shapes. However, it became apparent that there were large discrepancies in the levels of

variation, across individuals, for different coordinates. Two aspects were of particular importance: that the controls showed substantially less variation in the positions of landmarks than the cleft individuals; and that certain areas of the face, such as the x-coordinates of the midline, showed relatively low variability particularly amongst the controls and even amongst the cleft group at later time points, in comparison to points such as those along the upper lip which are highly affected by the cleft. Attempts were made to model these differences in variation, but problems remained with model convergence resulting from some of the smallest variances. So, the GPA was reconsidered. Since a selection of landmarks in some individuals are highly unstable, such as those around the nose and mouth, they were deemed not to be particularly suitable to use in the alignment process. A process in which some landmarks from some individuals do not correspond because a deformity alters them, while others do, does not seem sensible. Instead, while ensuring that as many symmetric landmarks were included as possible, an alignment process was carried out that matched individuals only on stable landmarks that were not highly affected by the cleft. The remaining landmarks were adjusted using the same transformation and included in further analysis. The result of this is that the differences between individuals are likely to be more realistic and attributable to true discrepancies in shape rather than aspects of the alignment process itself. This also led to fewer landmarks showing very small levels of variation, although this problem was not eradicated completely by the alteration in the alignment process, which had to be used in conjunction with modelling techniques and principal component analysis, as will be discussed presently. This work shows the care that is necessary in selecting and carrying out the method of alignment for shape data. Although Procrustes alignment is a well-established method, the subsequent analysis will be sensitive to the way in which it is carried out and it is important that resulting differences between shapes are exactly that, and not due to the particular method by which they are matched.

A further issue with longitudinal shape data is its inherent high dimensionality. The set of landmarks in the cleft-lip and palate study consists of 36 points in three dimensions, although not all of the landmarks were included in the analysis. The full set of curves includes 197 points in three dimensions and this will be discussed later. One of the aims of this thesis is to analyse a shape as a whole and quantify how it changes over time, rather than

deal with individual landmarks separately. Analysis of an entire shape over time involves modelling the correlation between points within a shape and within individuals over time. The methodology employed here to model the change in shapes over time is the linear mixed effects (LME) model. This allows each individual to have their own intercept (and slope, if desired) and assumes some estimated correlation structure over time, depending on user choice. Since there are multiple outcomes per individual in the form of tangent coordinates (even at a single time point), which may have different levels of variation, it is necessary to allow a random intercept term for each. However, this leads to a large number of variance parameters which must be estimated, if correlation between random intercepts for different coordinates is assumed - and this is clearly necessary. Therefore, a pairwise LME model approach was adopted, whereby for all the pairwise combinations of outcomes, a LME model was fitted to the tangent coordinate responses, allowing a different random intercept variance for each of the two coordinates. The parameter estimates were obtained from each pairwise model and combined into a long vector. Since some parameters may appear in more than one model, the repeated estimates were averaged and their standard errors calculated using a sandwich estimator resulting from pseudo-likelihood theory (as the sum of the pairwise log-likelihoods is a pseudo-likelihood).

This approach allows a substantially higher number of outcomes to be included in the model than would be the case otherwise. Parameter estimates may be obtained and bivariate confidence regions plotted for each landmark in two dimensions to study the differences between the groups in their shapes at different time points. Estimates may also be obtained regarding the level of variation in the random intercepts of the coordinates along with the correlations between them. However, a downside is that the same model has to be fitted for each pair of coordinates, and there is no room for variation in the model specification between coordinates that, for example, exhibit substantially different variances. This can lead to over-parameterisation for coordinates that show behaviour and dependencies that are not as complex as others.

There are also some computational issues. A function was written in `R` in order to fit the pairwise models for a general mean function and in which `deriv` (`stats` package) is used to analytically find the first and second derivatives of the relevant log-likelihood for each

pair (for the sandwich variance estimator) before `lme` (`nlme`) or `lmer` (`lme4`) fits the models. The function then combines and returns all of the resulting parameter estimates. The `lmer` function is still a work in progress and does not yet have all the facilities of `lme`, but the latter is substantially slower and less stable than the current version of `lmer`. This led to simplification of some models so that `lmer` could be used. The pairwise fitting process is quite time consuming computationally, which is a particular problem when carrying out simulations in which it is involved. It would be useful to develop a function that fits LME models in this context to substitute for `lme`/`lmer` so that some of these problems may be avoided, and this could form part of future work. Having said that, in most cases the computational process runs smoothly and the pairwise models are fitted successfully with one line of code. With further developments of `lmer` this should become even simpler.

The next issue is how to parameterise the pairwise models in order to allow for correlation between different points within an individual shape. Some structure could have been assumed, whereby points close together had higher correlation than those further away. Dealing with the coordinates in pairs made this difficult, however, as this gives no overall sense of how the correlation varies. An alternative idea was to estimate the empirical correlation between each pair of coordinates and if it was higher than some cutoff (say, 0.4), then the correlation would be estimated in the model and otherwise assumed zero. However, this involved reusing the data and was not felt to be an appropriate method. Models assuming independent random effects were compared with others which assumed independence between random effects for coordinates in different dimensions and estimated the correlations for those in the same dimension, and with further models which estimated all of the correlations. Generally the latter was the best performing, with high estimated correlations between some coordinates and low for others, and for some dimensions, a relationship with the Euclidean distance between them. High negative correlations were estimated for corresponding x-coordinates on either side of the face. Different random effects variances for each of the groups were included to account for the high variability amongst the cleft individuals relative to the controls. Different random error variances were estimated for the individual coordinates, since there may also be variation in measurement error across coordinates as some are more difficult to measure than others (this being one effect that `lme` can include but `lmer` cannot).

Related to these modelling aspects is the issue of some coordinates exhibiting very low levels of variance relative to others. As previously mentioned, the cleft patients in general showed more variation in landmark position than the controls, especially at three months and in particular areas of the face, such as around the nose and upper lip. This caused problems in the model fitting, since when one coordinate with high and another with low variability were included in a pairwise model, the latter would often have an estimated random intercept variance of zero, or very close to it. This occurred even (or, perhaps, especially) when the random effects variances were allowed to differ for the two groups, with the control group variance for the relevant coordinate being estimated as very close to zero. The result was that the correlation between the two coordinates would be estimated as very close to one or negative one (or infinity, if the variance was exactly zero). Simulations suggested that the models struggled to estimate variances whose true values were reasonably close to zero (but not as close as the model estimates), the result being that the estimated variances were artificially small and the estimated correlations may be nowhere near their true values.

Apparently valid results were obtained for the fixed effects of models where some of the variances were estimated as close to zero and corresponding correlations close to one (although those with variances estimated as exactly zero were discarded). However, this is not a desirable situation as it is unclear how much effect the spurious correlations have and there was a suggestion, from simulations, that the corresponding fixed effects standard errors may be slightly anticonservative. As previously mentioned, one approach to avoid this issue was to align the shapes on a subset of landmarks but this did not remove the problem. Another was to fit the model which assumed independence between coordinates in different dimensions while estimating the correlations within those in the same dimension but, aside from enabling the inclusion of a larger number of landmarks, this made little difference. The most successful method was to align the shapes on a subset of the landmarks and carry out principal components analysis (PCA) on the resulting tangent coordinates. Due to the high correlation between the landmarks within individual shapes, the dimension of 66 (22 landmarks in three dimensions) was reduced to 13 PCs explaining around 90% of the variance in the data. Currently 22 landmarks cannot be analysed using the pairwise approach applied to the tangent coordinates, and in most analyses up until this point a maximum of

10 landmarks has been included. Going beyond this number can cause problems with memory, especially if the model becomes more complex. The responses for the model were taken to be the PC scores and the computation time was substantially reduced by only having 13 outcomes. Furthermore, the issue with spuriously estimated correlations was eliminated. PCA dictates that the different components are orthogonal to one another, but since this was carried out on a global level and the pairwise analysis was performed on a local level with individual covariates, the random effects were not assumed independent. The model estimated correlations varied from about -0.5 to 0.5, with a small selection outside this range, and none were close to 1 or -1. The parameter estimates could be easily transformed back onto the original scale of the landmarks in order to plot the bivariate confidence regions and thereby provide a useful graphical presentation of the differences between groups and the time trends. The improved fit of the model for the PC scores may, of course, be in part due to the increase in the number of fixed effects arising from a saturated model for group and time, as opposed to assuming time is (partly) continuous.

One of the main conclusions of this study, then, is that the most stable, effective and time-efficient way of applying the pairwise approach to longitudinal shape data is to take the principal components of the tangent coordinates as the model responses. This approach allows a large number of points on the face to be included in the analysis, but at low computational cost compared to other methods.

Moving on from landmarks, the next issue is how to analyse curves. These consist of points placed at close increments along a curve of interest, such as the rim of the nose. Generally these points join up anatomical landmarks which may be placed at the ends of the curve and at other positions along it. As previously mentioned, the full set of curves consists of 197 landmarks in three dimensions. Given that the pairwise approach can only handle up to about 10 landmarks in three dimensions, clearly the tangent coordinates themselves may not be taken as the model responses. The initial approach was to reduce the dimension of the curves by fitting B-splines and using the estimated spline coefficients as the responses. B-splines with eight knots demonstrated reasonably good fits to the midline curves and so the pairwise models were fitted to the resulting 24 spline coefficients. The parameter estimates were transformed back onto their original scale and bivariate confidence regions

plotted at very small increments along the curve (which did not necessarily match the original landmarks). This provided a graphical presentation of the average curves in the two groups across time that resembled two-dimensional confidence bands around these means. Clearly these bands are not rigourous as confidence intervals and cannot be used to make definitive inference, but as a guide to the differences between groups and over time they are useful. However, the model fitted to the spline coefficients shared similar problems with the earlier landmark analyses in that many of the correlations were close to or equal to one or negative one. Therefore, the B-spline method did not seem to be a suitable method of dimension reduction in this context.

Instead, a similar approach to that finally taken with the landmarks was followed. PCA was carried out on the tangent coordinates of the full set of curves and 15 PCs found that explained around 90% of the variance. Since the set of curves involves $197 \times 3 = 591$ dimensions, this is an extraordinary reduction. However, given that the landmarks are so close together within the curve set, the correlations between them are likely to be extremely high and so this result is perhaps not surprising. It does show a considerable advantage of the PCA method over the B-splines, since the midline curve alone requires 24 spline coefficients and if other curves in the set were to be included their analysis would have to be done separately. This would involve an assumption of independence between curves, which is clearly unreasonable.

As in the landmark example, the parameter estimates obtained from modelling the PC scores may be transformed back onto the original scale. As for the B-spline representation of the curves, the backtransformed means and standard errors may be used to form two-dimensional confidence bands around the mean curves for the two groups at each time point by calculating the bivariate regions at many small increments along the curves. This gives the most detailed analysis of the cleft-lip and palate data yet and paves the way for a combination of landmarks and curves, and subsequently surfaces, to be analysed in future work. Another aspect of this future work could involve the presentation of results as trivariate confidence regions, using the `R` package `rgl`. This package allows three-dimensional viewing of landmarks, curves and surfaces and, while already useful as an exploratory tool, also has potential in graphical presentation of the results.

An issue that affects both the landmark and curve analyses is that of correlation over time. In all of the fitted models, only a random intercept was included and not a random slope. The reason for this was drawn partly from a desire to keep already complex models as parsimonious as possible, but mainly because computationally it is considerably more complicated to involve a random slope in the pairwise models. Taking this route does, however, involve some assumptions about the correlation over time that may be unreasonable. Including only a random intercept assumes that the correlation between measurements on a particular coordinate from the same individual at different time points follows a compound symmetry structure; in other words, the correlation between two measurements is the same regardless of the actual time points or the distance between them. In actual fact, due to difficulties with analytically inverting the covariance matrix for a pair of outcomes at four different time points, the compound symmetry structure in the estimates of the standard errors of the parameter estimates from the pseudo-likelihood was approximated further, so that different measurements over time were assumed independent. The sandwich estimator for the covariance matrix of the parameter estimates, however, should go some way towards correcting for this. This issue did not affect the parameter estimates themselves, since they were estimated by `lme`, and this may result in a discrepancy between the parameter estimates and their standard errors.

These assumptions may or may not be realistic in the case of shape data because it is possibly unlikely that the correlation between time points will follow any kind of linear course such that (for example) points further away from each other in time have lower correlation than those close together. At the same time it is probably not reasonable to assume that the correlations are the same between all time points. In particular, for the cleft-lip and palate study, the three month measurements for the cleft individuals are very different to any others and, even if the compound symmetry assumption is appropriate for the control group, it is unlikely to hold for the cleft group when comparing the correlation between the three month and other time points, with the correlation between the later ones and each other. This is related to the differences in variance which occur between the three month cleft group measurements and those from other group by time combinations, and was highlighted in plots of the unstandardised residuals, which show that a subset of the data, the cleft group

at three months, have residuals with high variance relative to the rest. Since the cleft group may consist of a range of individuals, from those with very severe cleft-lip and palate to others with a mild cleft, the variance in this group at three months is very high.

An alternative to a random slope, for this study, may be to allow a random intercept for the surgery effect, which would allow a higher variance for the cleft group at three months. For a quicker fix, the six to twenty-four month data could be analysed alone, ignoring the three month measurements, since the primary interest is in the differences in trend over this period. These areas are promising avenues for future research and if a random slope could be incorporated into the pairwise model fitting process then it would broaden the flexibility of these models and extend their usefulness.

The final issue is that of model comparison. The primary interest of the analysis was to quantify the differences in trend for the two groups between the six and twenty-four month time points. The bivariate confidence regions provide a useful graphical presentation of the results and allow identification of particular differences between the groups and over time. However, they do not allow definitive inference, while an appropriate model comparison hypothesis test would provide rigourous quantification of the significance of a particular effect, in this case the group by time interaction between certain time points.

Any hypothesis test on the results of the pairwise approach applied to the cleft-lip and palate data must be carried out on multiple parameters simultaneously, even if only one effect is of interest, because each parameter is coordinate-specific. As the number of coordinates increases, so too does the number of parameters being tested under the null hypothesis, as it is necessary that the shape is treated as a whole. Various model comparison tests were considered here, including the Wald, likelihood ratio (LR), pseudolikelihood ratio (PLR), the LR test using the averaged parameter estimates rather than the individual estimates from the pairwise models (LR^*), and a test suggested by Faraway (1997) which relies on bootstrapping. The Wald, LR and PLR tests all have disadvantages when applied to either high-dimensional data or random effects models. The Wald is sensitive to the parameterisation of the model and tends to underestimate the variance of the fixed effects because it does not explicitly take account of the random effects. The LR and PLR tests have statistics which can be highly influenced by directions of variation that are not important and this tends to get

worse as the number of outcomes increases. There is also an issue with the PLR test in that it tests whether all of the multiple copies of parameter estimates are equal to zero, rather than the individual parameters. These disadvantages are illustrated in simulations here, where all three tests perform poorly. The test suggested by Faraway (1997) concentrates on the differences (rather than the ratio, as for the LR test) between the eigenvalues of the covariance matrices from the full and submodels and is less affected by unimportant directions of variation. This test, and the LR* test, both perform well when compared to their bootstrapped null distributions for cases where either the null or alternative hypotheses hold. These two tests show high correlation in their statistics and it seems likely that either would provide an appropriate result in a scenario such as this. The downside of using model comparison tests that require bootstrapping, however, is that they are extremely time consuming, with both the pairwise full and submodels necessarily being fitted for each simulation. Computational efficiency could be increased if future work were to focus on making some distributional assumptions and thereby finding the null distribution of the test statistic.

Bibliography

Alshabani, A., Dryden, I., Litton, C. & Richardson, J. (2007), 'Bayesian analysis of human movement curves', *Applied Statistics* **56**, 415–428.

Ayoub, A., Garrahy, A., Hood, C., White, J., Bock, M., Siebert, J., Spencer, R. & Ray, A. (2003), 'Validation of a vision-based, three-dimensional facial imaging system.', *The Cleft Palate-Craniofacial Journal* **40**, 523–529.

Bacon, A. (2000), 'Principal components analysis of distal humeral shape in pliocene to recent african hominids: The contribution of geometric morphometrics', *American Journal of Physical Anthropology* **111**(4), 479–487.

Barry, S. & Bowman, A. (2006), Longitudinal models for shape data, *in* J. Hinde, J. Einbeck & J. Newell, eds, 'Proceedings of the 21st International Workshop on Statistical Modelling', Corrib and Data Printers Ltd., Galway, Ireland, pp. 82–89.

Barry, S. & Bowman, A. (2007), Modelling longitudinal spatial curve data, *in* J. del Castillo, A. Espinal & P. Puig, eds, 'Proceedings of the 22nd International Workshop on Statistical Modelling', l'Institut d'Estadística de Catalunya, IDESCAT, Barcelona, Spain, pp. 87–92.

Barry, S. & Bowman, A. (2008), 'Linear mixed models for longitudinal shape data with applications to facial modelling', *Biostatistics*. To Appear; published in Advance Access Feb 5 2008.

Bates, D. (2007), *lme4: Linear mixed-effects models using S4 classes*. R package version 0.99875-6.

Besag, J. (1975), 'Statistical analysis of non-lattice data', *Statistician* **24**, 179–195.

Bock, M. & Bowman, A. (2006), 'On the measurement and analysis of object asymmetry with applications to facial modelling', *Applied Statistics* **55**, 77–91.

Bookstein, F. (1978), *The measurement of biological shape and shape change. Lecture Notes in Biomathematics*, Vol. 24, Springer, Berlin.

Bookstein, F. (1984*a*), 'A statistical method for biological shape comparisons', *Journal of Theoretical Biology* **107**(3), 475–520.

Bookstein, F. (1984*b*), 'Tensor biometrics for changes in cranial shape', *Annals of Human Biology* **11**(5), 413–437.

Bookstein, F. (1991), *Morphometric tools for landmark data: geometry and biology*, Cambridge University Press, New York.

Bowman, A. & Bock, M. (2006), 'Exploring variation in three-dimensional shape data', *Journal of Computational and Graphical Statistics* **15**(3), 524–541.

Dean, D., Hans, M., Bookstein, F. & Subramanyan, K. (2000), 'Three-dimensional bolton-brush growth study landmark data: ontogeny and sexual dimorphism of the bolton standards cohort', *Cleft Palate-Craniofacial Journal* **37**(2), 145–156.

Decker, L., Berge, C., Renous, S. & Penin, X. (2007), 'An alternative approach to normalization and evaluation for gait patterns: Procrustes analysis applied to the cyclograms of sprinters and middle-distance runners', *Journal of Biomechanics* **40**(9), 2078–2087.

Dempster, A., Laird, N. & Rubin, D. (1977), 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society, Series B* **39**, 1–22.

Diggle, P., Heagerty, P., Liang, K.-Y. & Zeger, S. (2002), *Analysis of Longitudinal Data*, second edn, Oxford University Press Inc, Clarendon.

Dryden, I. & Mardia, K. (1998), *Statistical shape analysis*, Wiley, Chichester.

Dryden, I., Kume, A., Le, H. & Wood, A. (2008*a*), 'A multidimensional scaling approach to shape analysis', *Submitted for publication*.

Dryden, I., Oxborrow, N. & Dickson, R. (2008*b*), 'Familial relationships of normal spine shape', *Statistics in Medicine* **27**(11), 1993–2003.

Faraway, J. (1997), 'Regression analysis for a functional response', *Technometrics* **39**, 254–261.

Faraway, J. (2004), 'Modeling continuous shape change for facial animation', *Statistics and Computing* **14**, 357–363.

Faraway, J., Reed, M. & Wang, J. (2007), 'Modelling three-dimensional trajectories by using bézier curves with application to hand motion', *Applied Statistics* **56**(5), 571–585.

Fieuws, S. & Verbeke, G. (2006), 'Pairwise fitting of mixed models for the joint modelling of multivariate longitudinal profiles', *Biometrics* **62**(2), 424–431.

Fink, B., Grammer, K., Mitteroecker, P., Gunz, P., Schaefer, K., Bookstein, F. & Manning, J. (2005), 'Second to fourth digit ratio and face shape', *Proceedings of the Royal Society B* **272**, 1995–2001.

Franklin, D., Freedman, L., Milne, N. & Oxnard, C. (2007*a*), 'Geometric morphometric study of population variation in indigenous southern african crania', *American Journal of Human Biology* **19**(1), 20–33.

Franklin, D., Oxnard, C., O'Higgins, P. & Dadour, I. (2007*b*), 'Sexual dimorphism in the subadult mandible: Quantification using geometric morphometrics', *Journal of Forensic Sciences* **52**(1), 6–10.

Free, S., O'Higgins, P., Maudgil, D., Dryden, I., Lemieux, L., Fish, D. & Shorvon, S. (2001), 'Landmark-based morphometrics of the normal adult brain using mri', *NeuroImage* **13**, 801–813.

Geys, H., Molenberghs, G. & Ryan, L. (1999), 'Pseudolikelihood modeling of multivariate outcomes in developmental toxicology', *Journal of the American Statistical Association* **94**(447), 734–745.

Gokberk, B., Irfanoglu, M. & Akarun, L. (2006), '3d shape-based face representation and feature extraction for face recognition', *Image and Vision Computing* **8**(8), 857–869.

Golland, P., Grimson, W., Shenton, M. & Kikinis, R. (2001), Deformation analysis for shape based classification, *in* 'International Proceedings in Medical Imaging'.

Golland, P., Grimson, W., Shenton, M. & Kikinis, R. (2005), 'Detection and analysis of statistical differences in anatomical shape', *Medical Image Analysis* **9**(1), 69–86.

Goodall, C. (1983), The statistical analysis of growth in two dimensions, PhD thesis, Harvard University, Massachusetts, U.S.A.

Harmon, E. (2007), 'The shape of the hominoid proximal femur: a geometric morphometric analysis', *Journal of Anatomy* **210**, 170–185.

Harvati, K. (2003*a*), 'The neanderthal taxonomic position: models of intra- and inter-specific craniofacial variation', *Journal of Human Evolution* **44**(6), 107–132.

Harvati, K. (2003*b*), 'Quantitative analysis of neanderthal temporal bone morphology using three-dimensional geometric morphometrics', *American Journal of Physical Anthropology* **120**(4), 323–338.

Hennessy, R., Baldwin, P., Browne, D., Kinsella, A. & Waddington, J. (2007), 'Three-dimensional laser surface imaging and geometric morphometrics resolve frontonasal dysmorphology in schizophrenia', *Biological Psychiatry* **61**, 1187–1194.

Hennessy, R., McLearie, S., Kinsella, A. & Waddington, J. (2005), 'Facial surface analysis by 3d laser scanning and geometric morphometrics in relation to sexual dimorphism in cerebral-craniofacial morphogenesis and cognitive function', *Journal of Anatomy* **207**, 283–295.

Hood, C., Bock, M., Hosey, M., Bowman, A. & Ayoub, A. (2003), 'Facial asymmetry - 3d assessment of infants with cleft lip and palate', *International Journal of Paediatric Dentistry* **13**(1), 404–10.

Hood, C., Hosey, M., Bock, M., White, J., Ray, A. & Ayoub, A. (2004), 'Facial characterization of infants with cleft lip and palate using a three-dimensional capture technique', *Cleft Palate-Craniofacial Journal* **41**(1), 27–35.

Hutton, T., Buxton, B., Hammond, P. & Potts, H. (2003), 'Estimating average growth trajectories in shape-space using kernel smoothing', *IEEE Transactions on Medical Imaging* **22**(6), 747–753.

Kaliontzopolou, A., Carretero, M. & Llorente, G. (2007), 'Multivariate and geometric morphometrics in the analysis of sexual dimorphism variation in *Podarcis* lizards', *Journal of Morphology* **268**, 152–165.

Kendall, D. (1984), 'Shape-manifolds, procrustean metrics and complex projective spaces', *Bulletin of the London Mathematical Society* **16**, 81–121.

Kent, J. & Mardia, K. (2001), 'Shape, procrustes tangent projections and bilateral symmetry', *Biometrika* **88**(2), 469–485.

Kent, J., Mardia, K., Morris, R. & Aykroyd, R. (2000), Procrustes growth models for shape, *in* 'Proceedings of the First Joint Statistical Meeting New Delhi', India, pp. 236–238.

Kent, J., Mardia, K., Morris, R. & Aykroyd, R. (2001), Functional models of growth for landmark data, *in* K. V. Mardia & R. G. Aykroyd, eds, 'Proceedings in Functional and Spatial Data Analysis', Leeds University Press, Leeds, pp. 109–115.

Kieser, J., Bernal, V., Waddell, J. & Shilpa, R. (2007), 'The uniqueness of the human anterior dentition: A geometric morphometric analysis', *Journal of Forensic Sciences* **52**(3), 671–677.

Kume, A. (2007), 'Shape space smoothing splines for planar landmark data', *Biometrika* **94**, 513–528.

Laird, N. & Ware, J. (1982), 'Random-effects models for longitudinal data', *Biometrics* **38**, 963–974.

Larsen, R., Hilger, K. & Wrobel, M. (2002), Statistical 2d and 3d shape analysis using non-euclidean metrics, *in* 'Medical Image Computing and Computer-Assisted Intervention - MICCAI 2002, 5th Int. Conference', Springer, Tokyo, Japan.

Le, H. & Kume, A. (2000), 'Detection of shape changes in biological features', *Journal of Microscopy* **200**(2), 140–147.

Lockwood, C., Lynch, J. & Kimbel, W. (2002), 'Quantifying temporal bone morphology of great apes and humans: an approach using geometric morphometrics', *Journal of Anatomy* **201**(6), 447–464.

Mardia, K., Bookstein, F. & Moreton, I. (2000), 'Statistical assessment of bilateral symmetry of shapes', *Biometrika* **87**(2), 285–300.

Mardia, K., Kent, J. & Bibby, J. (1979), *Multivariate Analysis*, Academic Press, Inc, London.

Molenberghs, G. & Verbeke, G. (2005), *Models for discrete longitudinal data*, Springer, New York.

Morris, R., Kent, J., Mardia, K. & Aykroyd, R. (2000), A parallel growth model for shape, *in* S. Arridge & A. Todd-Pokropek, eds, 'Proceedings in Medical Imaging Understanding and Analysis', BMVA, London, pp. 171–174.

Morris, R., Kent, J., Mardia, K., Aykroyd, R., Fidrich, M. & Linney, A. (1999*a*), Exploratory analysis of facial growth, *in* K. Mardia, R. Aykroyd & I. Dryden, eds, 'Proceedings in Spatial Temporal Modelling and its Applications', Leeds University Press, pp. 39–42.

Morris, R., Kent, J., Mardia, K., Fidrich, M., Aykroyd, R. & Linney, A. (1999*b*), Analysing growth in faces, *in* H. Arabnia, ed., 'Proceedings Conference on Imaging Science, Systems and Technology', CSREA Press, Las Vegas, pp. 404–410.

Mutsvangwa, T. & Douglas, T. (2007), 'Morphometric analysis of facial landmark data to characterize the facial phenotype associated with fetal alcohol syndrome', *Journal of Anatomy* **210**, 209–220.

O'Higgins, P. & Jones, N. (1998), 'Facial growth in *Cercocebus torquatus*: an application of three-dimensional geometric morphometric techniques to the study of morphological variation', *Journal of Anatomy* **193**, 251–272.

Pinheiro, J. & Bates, D. (2000), *Mixed-Effects Models in S and S-PLUS*, Springer-Verlag, New York.

R Development Core Team (2007), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0; http://www.R-project.org.

Ramsay, J. & Silverman, B. (1997), *Functional Data Analysis*, Springer-Verlag, New York.

Samir, C., Srivastava, A. & Daoudi, M. (2006), 'Three-dimensional face recognition using shapes of facial curves', *IEEE Transaction on Pattern Analysis and Machine Intelligence* **28**(11), 1858–1863.

Schaefer, K., Mitteroecker, P., Gunz, P., Bernhard, M. & Bookstein, F. (2004), 'Craniofacial sexual dimorphism patterns and allometry among extant hominids', *Annals of Anatomy* **186**, 471–478.

Šefčáková, A., Katina, S., Velemínská, J., Brůžek, J. & Velemínský, P. (2003), 'Geometric analysis of sexual dimorphism in upper palaeolithic skulls from Předmostí (Czech Republic)', *Slovak Anthropology (Slovenská antropológia)* **6 (n.s.1)**, 141–146.

Small, C. (1996), *The statistical theory of shape*, Springer-Verlag, New York.

Srivastava, A., Joshi, S., Mio, W. & Liu, X. (2004), Statistical shape analysis: clustering, learning and testing, *in* 'ECCV', Prague, Czech Republic.

Styner, M., Lieberman, J., McClure, R., Weinberger, D. & Gerig, G. (2005), 'Morphometric analysis of lateral ventricles in schizophrenia and healthy controls regarding genetic and disease-specific factors', *Proceedings of the National Academy of Sciences of the United States of America* **102**(13), 4872–4877.

Theobald, C., Glasbey, C., Horgan, G. & Robinson, C. (2004), 'Principal component analysis of landmarks from reversible images', *Applied Statistics* **53**(1), 163–175.

Thompson, D. (1917), *On growth and form*, Cambridge University Press, Cambridge.

Trotman, C.-A., Faraway, J. & Phillips, C. (2005), 'Visual and statistical modeling of facial movement in patients with cleft-lip and palate', *Cleft Palate-Craniofacial Journal* **42**(3), 245–254.

Velemínská, J., Katina, S., Šmahel, Z. & Sedláčková, M. (2006), 'Analysis of facial skeleton shape in patients with complete unilateral cleft lip and palate: geometric morphometry', *Acta Chirurgiae Plasticae* **48**(1), 26–32.

Verbeke, G. & Molenberghs, G. (2000), *Linear mixed models for longitudinal data*, Springer, New York.

White, J., Ayoub, A., Hosey, M., Bock, M., Bowman, A., Bowman, J., Siebert, J. & Ray, A. (2004), 'Three-dimensional facial characteristics of caucasian infants without cleft and correlation with body measurements', *Cleft Palate-Craniofacial Journal* **41**(6), 593–602.