Linear mixed models for longitudinal shape data with applications to facial modeling

SARAH J. E. BARRY*, ADRIAN W. BOWMAN

Department of Statistics, University of Glasgow, 15 University Gardens, Glasgow G12 8QW, UK
sarah@stats.gla.ac.uk

SUMMARY
We present a novel application of methods for analysis of high-dimensional longitudinal data to a comparison of facial shape over time between babies with cleft lip and palate and similarly aged controls. A pairwise methodology is used that was introduced in Fieuws and Verbeke (2006) in order to apply a linear mixed-effects model to data of high dimensions, such as describe facial shape. The approach involves fitting bivariate linear mixed-effects models to all the pairwise combinations of responses, where the latter result from the individual coordinate positions, and aggregating the results across repeated parameter estimates (such as the random-effects variance for a particular coordinate). We describe one example using landmarks and another using facial curves from the cleft lip study, the latter using B-splines to provide an efficient parameterization. The results are presented in 2 dimensions, both in the profile and in the frontal views, with bivariate confidence intervals for the mean position of each landmark or curve, allowing objective assessment of significant differences in particular areas of the face between the 2 groups. Model comparison is performed using Wald and pseudolikelihood ratio tests.

Keywords: Curves; Mixed models; Multivariate longitudinal profiles; Pairwise modelling; Shape analysis.

1. INTRODUCTION
Statistical shape analysis became a recognized 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, 1984a,b), Kendall (1984), and Goodall (1983) brought together the fields of geometry, biology, and statistics to provide a framework for analyzing shapes, realizing the initial aims of Thompson (1917). Dryden and Mardia (1998) provide an authoritative synthesis of recent work. Shape analysis is currently used for many different applications in fields as diverse as computer science and medicine.

Much work has been done on 2-dimensional shape analysis, including study of variation in reversible images (Theobald and others, 2004), symmetry of children’s faces as they grow (Morris and others, 1999; Bock and Bowman, 2006), a parallel curve model for growth (Morris and others, 2000), and functional models for growth (Kent and others, 2001), among others. Considerably less research has focused on 3-dimensional shapes, particularly those that are measured over time. One exception is Bowman and

*To whom correspondence should be addressed.
Fig. 1. Low-resolution surface mesh of an infant with landmarks (left) and curves in frontal (centre) and profile (right) views marked onto the mesh.


A standard approach to the analysis of longitudinal data is to use random-effects models (Laird and Ware, 1982). This approach may be applied in the case of multivariate outcomes which would assume a random effect for each outcome in the form of subject-specific deviations from the average for that outcome. However, problems arise with this approach as the number of outcomes increases, and hence the dimensionality of the random-effects variance matrix becomes very large. We therefore follow Fieuws and Verbeke (2006), who suggested an alternative method of fitting bivariate models to each of the pairs of outcomes and combining the results.

Our data 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. Images of the children in 3 dimensions were captured at age 3, 6, 12, and 24 months using a sophisticated stereophotogrammetry system (Ayoub and others, 2003). The cleft group underwent surgical repair between the 3- and 6-month points. The background of the study is given in Hood and others (2004).

An example of a low-resolution surface mesh of points describing an individual face is displayed in Figure 1. The picture on the left shows landmarks placed on the mesh at anatomically meaningful points, such as around the edges of the nose and upper lip, whereas the pictures in the center and on the right show curves marked onto the mesh in frontal and profile views, respectively, which similarly describe the shape of the nose and upper lip. There is interest in the differences between the cleft and control children on average and in how the facial shapes change over time as the children grow, particularly after the cleft group undergoes surgery.

The article proceeds as follows. Section 2 describes the approach of Fieuws and Verbeke (2006) and applies it to the setting of landmark data, using the cleft lip and palate study as an example, and finally addresses model comparison. Section 3 extends the methods presented to curve data which describe the shape of a face in more detail than landmarks do alone. Section 4 contains discussion.

2. PAIRWISE LONGITUDINAL MODELING OF SHAPE DATA

2.1 Statistical shape analysis

One definition of shape is, as described by Dryden and 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, which are points of correspondence on each object that match between and within populations. These are often anatomical landmarks with a biological meaning. However, they can also be mathematical landmarks with geometric meaning, such as points of extreme curvature, or pseudolandmarks which generally connect the other types.

When a sample of shapes is measured, there are likely to be differences between individuals according to their location, rotation, and scale. These factors therefore must be removed if we are to concentrate solely on differences in shape. Procrustes analysis is an established method of removing the effects of location, rotation, and scale by matching the configurations to one another (Dryden and Mardia, 1998). The facial configurations analyzed here were aligned by generalized Procrustes analysis (GPA), matching the configurations using a subset of landmarks which were chosen to be stable points (the midline below the lower lip and points around the eyes) that are relatively unaffected by the cleft deformity. Other landmarks, not included in the matching set, were then used to assess shape changes in the areas of interest. This was carried out so that any group differences observed in the final analysis could reasonably be attributed to true differences and not to the matching process.

Following alignment of the sample of shapes, the resulting Procrustes coordinates may be approximated by tangent coordinates since the shapes are relatively close together in shape space (Dryden and Mardia, 1998). The tangent coordinate space is defined as a linearized version of shape space at a tangent to a nominated point, usually the mean. We approximate the tangent coordinates by the Procrustes residuals, calculated by subtracting the Procrustes mean shape, \( \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i^P \), from each individual shape, where \( X_i^P \) are the set of Procrustes coordinates for individual \( i \). This allows general multivariate methods to be applied to the data (Dryden and Mardia, 1998).

### 2.2 Pairwise approach: motivation and methodology

Each facial shape in this application is described by a set of tangent coordinates representing landmark positions in the \( x, y, \) and \( z \) dimensions. By the very nature of shapes, within any individual there will be correlation between coordinates and, since the data are longitudinal, also over time. For a small number of landmarks (say, up to 2 in 3 dimensions), we can fit a fully joint model to the tangent coordinates such that \( \mathbf{y}_{ri} \), the \( n_{ri} \) vector of repeated measurements across time for subject \( i = 1, \ldots, N \) and coordinate \( r = 1, \ldots, m \), is assumed to be distributed as multivariate normal, conditional upon a vector of random effects, \( \mathbf{b}_{ri} \). If the vector \( \mathbf{b}_i \) contains all random effects for individual \( i \) and \( \mathbf{D} \) represents its covariance matrix, then it is clear that as the number of coordinates increases the dimension of \( \mathbf{D} \) becomes very large and its parameters difficult or impossible to estimate (Molenberghs and Verbeke, 2006).

Fieuws and Verbeke (2006) suggest instead a pairwise modeling approach, which preserves the advantages of joint modeling while avoiding the computational difficulties inherent in estimating the parameters of \( \mathbf{D} \) for a large number of coordinates. This involves fitting \( m(m - 1)/2 \) bivariate models for all possible pairs

\[
(\mathbf{Y}_1, \mathbf{Y}_2), (\mathbf{Y}_1, \mathbf{Y}_3), \ldots, (\mathbf{Y}_1, \mathbf{Y}_m), (\mathbf{Y}_2, \mathbf{Y}_3), \ldots, (\mathbf{Y}_2, \mathbf{Y}_m), \ldots, (\mathbf{Y}_{m-1}, \mathbf{Y}_m),
\]

where \( \mathbf{Y}_r \) is a vector of measurements of the \( r \)th outcome across individuals and times. This vector is of length \( \sum_{i=1}^{N} n_{ri} \) and contains the stacked \( \mathbf{y}_{ri} \) for \( i = 1, \ldots, N \).

Fitting the pairwise models may be treated as the maximization of a pseudolikelihood (Besag, 1975), and, in this framework, the vector of all parameter estimates from all models, \( \hat{\theta} \) (which may contain repetitions due to different models containing the same outcome), is distributed asymptotically as

\[
\sqrt{N} (\hat{\theta} - \theta) \sim \mathcal{M} \mathcal{V} \mathcal{N}(0, J^{-1} K J^{-1}),
\]

where \( N \) is the number of individuals and \( J \) and \( K \) are matrices of second and first derivatives of the log-likelihood, respectively. The required vector containing a single estimate of each parameter can then
be obtained by \( \hat{\theta}^* = A \hat{\theta} \), where \( A \) is the requisite matrix of coefficients to calculate the averages of the repeated parameter estimates. Hence, the variance matrix for \( \hat{\theta}^* \) is \( \text{Var}(\hat{\theta}^*) = AJ^{-1}KJ^{-1}A' / N \).

### 2.3 Application to cleft lip and palate data

Fieuws and Verbeke (2006) apply the pairwise approach to a set of data on hearing thresholds at different frequencies. The application to shape data is a novel yet natural one, given its inherent high dimensionality and, in this case, longitudinal nature. The approach also allows us to carry out a considerably more detailed analysis than would otherwise have been possible.

Since a cleft defect may occur on either side of the face and we are interested in the extent of asymmetry when considering the frontal view of the face, we reflected the images for those individuals who had a cleft on the right side. This effectively left us with a cleft group made up entirely of left-side cleft patients, thus allowing comparison of the asymmetry in this group with that of the controls, without being concerned about the side of the cleft.

Four sets of measurements from 3 individuals in the study were removed from the data set due to data errors. We applied the pairwise model approach to the remaining data using a program written in R (R Development Core Team, 2004). For any specified mean function, this program fits the bivariate linear models for all pairwise combinations of outcomes using either the lme function in the nlme package (Pinheiro and others, 2004) or the lmer function in the lme4 package (Bates and Sarkar, 2007), depending on the covariance structure required. The estimates of all fixed- and random-effects parameters are extracted from the output, and the first and second derivatives that make up the \( K \) and \( J \) matrices introduced in (2.1) are calculated and evaluated using the deriv function in the stats package. The resulting code allows any model of interest to be fitted easily in a single computational procedure with considerable saving in analytical manipulation.

Figure 2 gives the mean trends over time of the tangent coordinates for each group. The 9 landmarks represented are the corners of the lips, the midpoint of the upper lip and two highest points on the “Cupid’s

\[ \text{Fig. 2. Mean trends over time, with 95\% confidence intervals, of the tangent coordinates for the cleft (black lines) and control (gray lines) groups (note that the plots are not all on the same scale). The columns represent the } x, y, \text{ and } z \text{ coordinates of each landmark.} \]
Linear mixed models for longitudinal shape data

559

bow,” the midpoint and corners of the base of the nose, and the tip of the nose. The plots show different
levels of variation between coordinates and groups, with the cleft group generally being more variable at
time points. This is unsurprising given the different natures of the cleft and control populations and
the effect of the cleft on different parts of the face. The cleft group also shows a large jump from 3 to 6
months which is not present among the controls and which, it can be assumed, is attributable to a surgery
effect. Plots of the individual trends (not shown) suggest that the assumption that each subject follows a
similar trajectory is reasonable, and a random slope is therefore unnecessary but clearly random intercepts
should be included.

The fitted model, where \( y_{ir}(t) \) is the value of a single outcome \( r \) for individual \( i \) at time \( t \) months after
the beginning of the study, allows for different random-effects variances for each coordinate which also
differ between the 2 groups. The model is given by

\[
y_{ir}(t) = \beta_0 r + b_{ir} g_i + \beta_1 r g_i + \beta_2 r t + \beta_3 r g_i \cdot t + \beta_4 r g_i \cdot s(t) + \epsilon_{ir}(t),
\]

(2.2)

where \( t \) is a continuous variable representing time and \( s(t) \) is an indicator variable taking value 0 at \( t = 3 \)
months and 1 at all later times to account for the effects of surgery. The effect of time is assumed to be
linear for the control group and from 6 months onward for the cleft group, and Figure 2 indicates that
this is generally appropriate. Since the group differences in variance are accounted for in the random-
effects term, the random error, \( \epsilon_{ir}(t) \), is distributed as \( N(0, \sigma^2) \), and therefore, the error variance differs
by coordinate only. The covariate \( g_i \) is an indicator taking value 1 if individual \( i \) is in the cleft group
and 0 otherwise. For any individual \( i \) who is in the control group, their random intercept vector for pair
\( p = (r, s) \), where \( r, s = 1, \ldots, m \), is distributed as

\[
b_{ip0} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ \tau_{r0} \\ \tau_{s0} \end{pmatrix}, \begin{pmatrix} \tau^2_{r0} & \tau_{rs0} \\ \tau_{rs0} & \tau^2_{s0} \end{pmatrix}\right).
\]

For a cleft group subject,

\[
b_{ip1} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ \tau^2_{r1} \\ \tau_{rs1} \end{pmatrix}, \begin{pmatrix} \tau^2_{r1} & \tau_{rs1} \\ \tau_{rs1} & \tau^2_{s1} \end{pmatrix}\right).
\]

This example involves landmarks that describe a region of the face that is of particular interest in cleft
children as it is highly affected by the cleft defect. The main parameters of interest are those that de-
scribe the differences between the groups over time but, since these are estimated in tangent space and
separately for each coordinate, they are not easy to interpret directly. A clear visual interpretation of the
results may be given by transforming back into shape space. The estimated mean for each group and time
combination is calculated by adding the Procrustes mean shape onto the relevant set of estimated tangent
coordinates.

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 sub-
matrices 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 \( (u - \mu)'\Sigma^{-1} (u - \mu) \sim \chi^2 \)
where the random 2-vector \( u \) is distributed as \( N(\mu, \Sigma_{rs}) \) (Mardia and others, 1979). In this context, \( \Sigma_{rs} \)
is the \( (2 \times 2) \) submatrix of \( \Sigma \) that corresponds to the coordinates of interest and \( \mu \) is the corresponding
mean. The boundaries of the confidence regions are therefore easily constructed by drawing the appropri-
ate contour of a bivariate normal density function with mean given by the estimate of interest and with
covariance matrix \( \Sigma_{rs} \).

The resulting bivariate 95% confidence intervals for each landmark are presented in Figure 3, in both
the profile and the frontal views. These regions allow observation of the extent of shared variation in the
Fig. 3. Bivariate 95% confidence regions for the cleft (grey) and control (black) group mean positions of the landmarks describing the upper lip and nose at each time point (upper: profile view, 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.

2 dimensions for any particular landmark and comparison of the group means in any direction within either view. The figure shows clear differences between the groups and highlights the trends over time. In particular, the landmarks around the upper lip and base of the nose are highly affected, with some 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 nose becomes slightly more pointed.

2.4 Model checking and comparison

A variety of checks were made on this model to assess its fit, including the comparison of mean tangent coordinate trends over time for each coordinate with the corresponding model fit, superimposition of mean observed landmarks onto the fitted equivalents in Figure 3, and comparison of estimated random effects across coordinates and groups with ordinary least squares residuals from a simple linear model. All checks showed that the model is an adequate fit to the data (plots not shown).

We used a Wald test for the existence of a group by time interaction in the above example by testing the null hypothesis that all the coordinate-specific group by time interactions were equal to zero. We obtained a $\chi^2$-statistic of 197 on 27 degrees of freedom, which suggests strong evidence of a significant difference between the 2 groups in how the shape of the face, as described by the lower face landmarks, changes over time from 6 to 24 months of age.

Since the Wald test is known to be sensitive to the particular parameterization of a model, we also applied the pseudolikelihood ratio (PLR) test (Geys and others, 1999). For the null hypothesis $H_0 : \gamma = 0$, where $\gamma$ is a $k$ subvector of the vector of regression parameters $\theta$, the test statistic

$$G^* = 2[pl(\hat{\theta}) - pl(\hat{\theta}_0)]/\lambda,$$

where $pl(\cdot)$ is the log of the pseudolikelihood (calculated by summing the pairwise log-likelihoods), approximately follows a $\chi^2_k$ distribution. The vector $\hat{\theta}_0$ contains the estimates of the remaining parameters.
of $\theta$ under the reduced model. Recall that $\theta$ 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 $\theta$ rather than to $\theta^*$. The denominator of the test statistic, $\bar{\lambda}$, is the mean of the eigenvalues of $(J_0)^{-1} \Sigma_0$, where $J_0$ and $\Sigma_0$ are $k \times k$ submatrices of $J^{-1}$ and $J^{-1}KJ^{-1}$, respectively (as defined in Section 2.2). These are evaluated here under the null hypothesis, as Geys and others (1999) found that if evaluated under the alternative, the power may be misleadingly high. In this example, the test of whether all the group by time interaction parameters are equal to zero gives a $\chi^2$-statistic of 5751 on 702 degrees of freedom, which again gives strong evidence to reject the null hypothesis.

3. Curve analysis

3.1 Methods and motivation

Describing a face by curves instead of landmarks gives more information about the shape. For a random selection of cleft and control cases, Figure 4 displays the curves which interpolate pseudolandmarks that are equally spaced between the superimposed anatomical landmarks. For the purposes of this article, we focus on the midline curve, which runs from the midpoint of the bridge of the nose to the top of the upper lip. Each midline curve contains 57 points in 3 dimensions. The curves were aligned using GPA, with landmarks on curves other than the midline used as the matching subset. Since the lmepair function can fit the pairwise models to a maximum of about 30 outcomes (because of the large $J^{-1}KJ^{-1}$ matrix), it is more convenient to fit a B-spline to each curve and use the spline coefficients as the response variable for each individual at each time point instead of the actual curve points. Since we could adequately describe a curve using a B-spline with 7 knots, this reduced the number of points for each curve to 21 (7 coefficients in 3 dimensions).

3.2 Fitting the model

The mean spline coefficient trends, with 95% confidence intervals, are displayed in Figure 5. For the most part, these show a large surgery effect between 3 and 6 months, as for the landmark tangent

Fig. 4. Four examples of randomly chosen cleft (black) and control (grey) cases at 3 months, showing both the interpolating curves and the superimposed landmarks.
coordinates, and then a rather curved trend between 6 and 24 months, at least for the cleft group. The control group shows more of a linear trend across time for most coefficients. There are clearly, once again, differences in the variability between groups and coordinates and this must be accounted for in the model.

The model fitted to the spline coefficients, $\eta_{ir}(t)$, is therefore similar to (2.2) but with a mean function of

$$
\eta_{ir}(t) = \beta_{0r} + b_{ir} + \beta_{1r}t_6 + \beta_{2r}t_6 + \beta_{3r}t_{12} + \beta_{4r}t_{24} + \beta_{5r}g_i \cdot t_6 + \beta_{6r}g_i \cdot t_{12} + \beta_{7r}g_i \cdot t_{24} + \epsilon_{ir}(t),
$$

where $t_6$, $t_{12}$, and $t_{24}$ are indicators which take value 1 if the time point is 6, 12, or 24 months, respectively, and 0 otherwise. The random effects and random error have the same distributions as in (2.2).

Figure 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 closely located points along the mean curves to give the impression of overall 2-dimensional confidence regions. This figure shows that on average the flatter nose of the cleft group subjects becomes more similar to that of the controls at 24 months, whereas there is still substantial asymmetry in the frontal view of the curve leading down through the nose to the upper lip.

### 3.3 Model comparison

As in the landmark example, we obtain the Wald and PLR statistics to test for the presence of a group by time interaction. These result in a $\chi^2$-statistic of 963 on 63 degrees of freedom for the Wald and a $\chi^2$-statistic of 8824 on 1260 degrees of freedom for the PLR test, giving clear evidence to reject the null hypothesis of no interaction. These both corroborate the inferences made from the model comparison test on the landmark data.
4. DISCUSSION

The lmepair function will fit the pairwise models to various kinds of multivariate longitudinal data, with a general mean function and a selection of variance structures, and is included as supplementary material to this article, available at *Biostatistics* online at http://www.biostatistics.oxfordjournals.org.

The models fitted in this article had random intercepts only, and the random error variances were assumed constant over time. These assumptions may be somewhat restrictive, but since there were only 4 measurements over time on each individual (and sometimes fewer, due to missing data), it was unlikely that there was enough information to fit a correlated error structure. Furthermore, the nature of facial growth and the corresponding change in tangent coordinates or spline coefficients give no particular reason to assume that the correlation over time should not be approximately the same regardless of the actual time points. The most important factors in the trends appear to be the different group and coordinate/coefficient variances, and the fitted models account for these. It is possible, also, that the trend in the controls for coordinate 27 in the landmark data warrants a nonlinear effect of time. In the interests of parsimony, however, and since the coordinate is unique in this aspect, the trend was approximated by a linear effect. This highlights a drawback with the pairwise model approach, in that the same model is assumed for all outcomes.

Some other drawbacks with the methods as we have applied them consist of an upper limit (of about 30) in the number of coordinates that may be included due to the calculation of the large $\text{Var}(\hat{\theta}^*)$ matrix, and some problems with convergence and estimation on the boundary of the parameter space because of very small variances in individual bivariate models. One side effect of the latter is that some of the correlations between random intercepts for different coordinates are estimated as very close to 1 or negative 1, which may affect model comparison tests. An alternative approach is to take the first few principal components of the tangent coordinates and use the corresponding scores as the responses in the pairwise models. This allows a larger number of landmarks to be included in the model and avoids the aforementioned

---

**Fig. 6.** Bivariate 95% confidence regions for cleft (grey) and control (black) groups for the mean midline curves at each time point in the profile (upper) and frontal (lower) views.
problems with variance estimation and model convergence. Another method is to assume independence between dimensions and allow a general covariance structure for the random intercepts with dimension, which also allows inclusion of more landmarks but does not avoid the estimation issues. All approaches give very similar fixed-effects means and confidence intervals.

In future work, we intend to modify the pairwise approach to allow for a specification of the covariance structure that, for example, could take into account the spatial nature of the face. This could potentially lead to fewer covariance parameters being estimated, resulting in gains in efficiency, fewer estimation problems, and a larger number of coordinates being included in the model. We also intend to produce models for the inclusion of all curves in Figure 4 rather than just the midlines. Other future activities involve further work on model comparison, in particular a PLR test that includes only one estimate for each parameter.

ACKNOWLEDGMENTS

The facial modeling application that is described in the paper was funded by the Chief Scientist’s Office of the Scottish Executive, whose support is gratefully acknowledged. The research project on facial modeling was carried out in collaboration with colleagues from Glasgow Dental Hospital and the Department of Computing Science at the University of Glasgow. Conflict of Interest: None declared.

FUNDING

Engineering and Physical Sciences Research Council (EP/P500249/1) to S.B.

REFERENCES


[Received 31 October, 2006; first revision 19 June, 2007; second revision 19 November, 2007; accepted for publication 17 December, 2007]