Regression modelling for size-and-shape data based on a Gaussian model for landmarks

Ian L. Dryden
School of Mathematical Sciences, University of Nottingham
Alfred Kume
School of Mathematics, Statistics and Actuarial Sciences, University of Kent
Phillip J. Paine
School of Mathematics and Statistics,University of Sheffield
and
Andrew T. A. Wood
Research School of Finance, Actuarial Studies and Statistics,
Australian National University

January 17, 2020

Regression modelling for size-and-shape data based on a Gaussian model for landmarks

Abstract

In this paper we propose a regression model for size-and-shape response data. So far as we are aware, few such models have been explored in the literature to date. We assume a Gaussian model for labelled landmarks; these landmarks are used to represent the random objects under study. The regression structure, assumed in this paper to be linear in the ambient space, enters through the landmark means. Two approaches to parameter estimation are considered. The first approach is based

directly on the marginal likelihood for the landmark-based shapes. In the second approach we treat the orientations of the landmarks as missing data, and we set up a model-consistent estimation procedure for the parameters using the EM algorithm. Both approaches raise challenging computational issues which we explain how to deal with. The usefulness of this regression modelling framework is demonstrated through real-data examples.

Keywords: EM algorithm, size-and-shape analysis, offset-normal shape distributions, mean shape, shape of the mean.

1 Introduction

Statistical shape analysis has developed rapidly since the 1980s and has had a major impact in many fields of application, such as medicine and the health sciences, biology and forensic science. See, for example, the monographs by Dryden and Mardia (2016) and Kendall et al. (1999). However, despite the many successes of this field, in one important respect statistical methodology for shape analysis is still rather deficient: few if any suitable regression models have been developed for the situation in which the response variable is a shape or size-and-shape; in the latter, size information is also retained; see e.g. Dryden and Mardia (2016, Chapter 5). The aim of this paper is to develop a regression model for size-and-shape analysis of objects described by labelled landmarks in two and three dimensions, where the covariates are not required to have any particular structure. More specifically, the response is assumed to consist of size-and-shape data, i.e. size and shape information is retained, while location and orientation information is discarded. The new regression model is specified in §2.2.

In this paper we focus mainly on labelled landmarks in $m = 2$ or $m = 3$ dimensions. Two distinct approaches to shape analysis of random objects described by labelled landmarks have been developed in the literature. In one approach a statistical model for landmarks is proposed and then one works with the marginal distribution of the shapes, or size-and-shapes. The starting point here is the matrix of landmark means, μ_M, and the focus of interest is the shape, or size-and-shape, of μ_M. In the other approach one works directly with the shape or size-and-shape of the objects of interest, and often the Fréchet mean μ_F is the focus of interest. In general, however, μ_M and μ_F are not the same, in the sense that the shape corresponding to the mean μ_M is not the same as the mean shape μ_F, with similar conclusions holding for size-and-shape. See Kent and Mardia (1997, 2001) , Le (1998) and Le and Kume (2000a, 2000b) for further discussion and results relating to this

issue.

We focus here on the first approach. Our goal is to develop a regression model for μ_M which induces a regression model for size-and-shape response data. The analogous models for shape data are not covered here as they are more computationally challenging, except the $m = 2$ case where the inference involves the closed form expressions of the offset normal shape distributions; see e.g. Dryden and Mardia (2016, Chapter 11) and Kume and Welling (2010).

The structure of the paper is as follows. In §2 we discuss a convenient representation for size-and-shape and present the size-and-shape regression model that we focus on in this paper, while §3 contains the main results needed for fitting the model, either by marginal maximum likelihood or using the EM algorithm. In particular, Theorem 2 specifies the structure of the EM procedure, and we make use of various results which facilitate the calculation of the E-step. In §4 we explain how to deal with the challenging computational issues which arise when fitting the model with $m = 2$ and, especially, $m = 3$. In §5 we focus on the simple IID submodel with a scalar covariance matrix and explain the similarities and differences to the Procrustes approach in this setting. This relatively simple analysis throws light on why the Procrustes approach is consistent when $m = 2$, but inconsistent when $m \geq 3$; these theoretical findings are supported by simulation results. Numerical results for real-data examples are given in §6. In the main body of the paper we focus on the definition of size-and-shape in which size-and-shape is not reflection invariant. In Appendix A we present parallel results for the case where size-and-shape is reflection invariant. All proofs are given in Appendix B. Further results, containing formulae for approximatate standard errors, are given in the Supplementary Material.

2 Size-and-Shape Modelling

2.1 Representation of size-and-shape

We are interested in a sample of objects in $\mathbb{R}^m$, where each of these objects is represented by the Cartesian coordinates of $k+1$ *labelled* landmarks. The configuration matrix for a typical object may be written as

$$\check{X} = \begin{pmatrix} \check{x}_{1,1} & \check{x}_{1,2} & \cdots & \check{x}_{1,m} \\ \check{x}_{2,1} & \check{x}_{2,2} & \cdots & \check{x}_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \check{x}_{k+1,1} & \check{x}_{k+2,2} & \cdots & \check{x}_{k+1,m} \end{pmatrix} = \begin{pmatrix} \check{x}_{[1]}^\top \\ \check{x}_{[2]}^\top \\ \vdots \\ \check{x}_{[k+1]^\top} \end{pmatrix},$$

where $\check{x}_{[j]}$ is an m-vector containing the coordinates of landmark j. Since we are interested in the size and shape of the objects, but not their location or orientation, we need to remove information relating to location and orientation. A convenient way to remove location information is to *Helmertize* the landmarks; see e.g. Dryden and Mardia (2016, p.49). This entails working with the $k \times m$ Helmertized configuration matrix $X = H\check{X}$ rather than $\check{X}$, where H is the $k \times (k+1)$ Helmert submatrix of the $(k+1) \times (k+1)$ whose j^{th} row is given by

$$(-d_j, -d_j, ..., -d_j, jd_j, 0, ..., 0), \tag{1}$$

where $d_j = \{j(j+1)\}^{-\frac{1}{2}}$. The Helmertized configuration matrix X is called a pre-form.

In order to remove the effects of orientation, we assume $k \geq m$ and consider the singular value decomposition of the Helmertized configuration matrix X:

$$X = U\Delta R^\top, \quad R \in O(m), \quad U \in \mathcal{V}_{k,m} \tag{2}$$

where $O(m)$ is the space of $m \times m$ orthogonal matrices, $\mathcal{V}(k,m) = \{V \in \mathbb{R}^{k \times m} : V^\top V = I_m\}$ is the (k,m)-Stiefel manifold, and $\Delta = \text{diag}(\delta_1, \ldots, \delta_m)$ is a diagonal matrix with non-negative elements satisfying $\delta_1 \geq \cdots \geq \delta_m \geq 0$. It is easy to see that if we apply a general

orthogonal transformation to the landmarks, i.e. we post-multiply X by an arbitrary $S \in O(m)$, then the singular value decomposition of the resulting matrix has a different R in (2), but U and Δ are unchanged. We conclude from this that the orientation information in X is contained in R in (2), and the size-and-shape of the configuration X is characterized by U and Δ. To check that we have retained no more and no less than the size-and-shape information in $\check{X}$, suppose that we add an arbitrary vector $t \in \mathbb{R}^m$ to each landmark and pre-multiply each landmark vector by an arbitrary $S \in O(m)$. Then $\check{X}$ transforms to

$$\check{X}^* = \left(\check{X} + 1_{k+1}t^\top\right) S^T,$$

where 1_{k+1} is the $(k+1)$-vector of ones. From the definition of the Helmert submatrix H via (1), $H1_{k+1} = 0_k$, where 0_k is the k-vector of zeros. Therefore

$$\begin{aligned} H\check{X}^* &= H\left(\check{X} + 1_{k+1} \otimes t\right) S^\top \\ &= XS^\top \\ &= U\Delta(SR)^\top, \end{aligned}$$

so the orientation information R in (2) is replaced by SR but the size-and-shape information $U\Delta$ is unchanged.

In the above, reflection information in the configuration has not been retained because we are assuming that R in (2) lies in $O(m)$, the space of $m \times m$ orthogonal matrices, in which case $|R| = \pm 1$, rather than assuming that R lies in $SO(m)$, the space of $m \times m$ rotation matrices, in which case $|R| = +1$. If we wish to retain reflection information as part of the size-and-shape information, it is necessary to apply the following version of the singular value decomposition in which we restrict R to lie in $SO(m)$:

$$X = U\Delta R^\top, \quad R \in SO(m), \quad U \in \mathcal{V}_{k,m}, \tag{3}$$

and Δ is a positive-definite diagonal matrix. Note that we can always arrange that $R \in SO(m)$, because if $|R| = -1$, we can change the sign of one of the columns of R, change

the sign of the corresponding column of U and leave Δ unchanged, in which case X is unchanged, the new R has $|R| = 1$ and the new U still lies in $\mathcal{V}_{k,m}$.

For the remainder of this paper we shall focus on the version (3) in which $R \in SO(m)$, i.e. we retain the reflection information as part of the size-and-shape. However, only minor adjustments in the calculations specified below are required to implement the case in which we use (2) and $R \in O(m)$; full details are given in Appendix A.

2.2 A regression model for size-and-shape

Suppose that our observed dataset is of the form

$$(X_1, z_1), \ldots, (X_n, z_n), \tag{4}$$

where, for each $i = 1, \ldots, n$, each X_i is a $k \times m$ Helmertized configuration, or pre-form, and $z_i = (z_{i1}, \ldots, z_{ip})^\top$ is a vector of covariates associated with configuration X_i. We consider the underlying linear model

$$X_i | z_i \overset{\text{indep}}{\sim} \mathcal{N}_{k\times m}(\mu_i, I_m \otimes \Sigma), \tag{5}$$

where $\otimes$ denotes the Kronecker product (see e.g. Muirhead, 1982, p. 73), Σ is a general $k \times k$ covariance matrix, assumed non-singular, and

$$\mu_i = \sum_{j=1}^{p} z_{ij} B_j, \quad i = 1, \ldots, n, \tag{6}$$

where B_j, $j = 1, \ldots, p$, are $k \times m$ parameter matrices.

Remark 2.1. An important point is that, in the EM procedure we propose below in Theorem 2 for estimating the unknown parameters, we deliberately discard the information concerning the observed orientations and treat this as missing data.

Remark 2.2. The covariance structure $I_m \otimes \Sigma$ assumed in (5) deserves some comment. If, for example, we were to replace the identity I_m by a general $m \times m$ covariance matrix Ω, say,

then the covariance structure will depend in an explicit way on the choice of orientation. However, given that we are basing inference on the marginal distribution of $U_i\Delta_i$, $i = 1, \ldots, n$, there will be relatively little information in the data with which to estimate Ω, and consequently we would typically expect there to be identifiability problems. If, on the other hand, we believe that the observed landmark orientations are potentially important predictors of size-and-shape, then the orientation information should be explicitly modelled rather than discarded, a possibility we do not consider here.

Remark 2.3. Although we do not do so here, one could also consider models in which μ_i is a nonlinear function of the covariate vector z_i and the parameters.

2.3 Standardization of the parameter matrices

One important practical point arises from the fact that we are only interested in the size-and-shape of the mean configuration μ. If μ has singular value decomposition $V\Psi W^\top$, then the size-and-shape of μ is determined by $V \in \mathcal{V}_{k,m}$ and $\Psi = \mathrm{diag}\{\psi_1, \ldots, \psi_m\}$, while $W \in SO(m)$ just determines the orientation of the mean configuration and does not affect its size or shape. To avoid redundancy in the specification of the size-and-shape of the mean configuration μ, we need to remove the dependence of μ on its orientation through multiplication on the right by a suitable matrix $R \in SO(m)$. One way to remove this dependency is now explained.

Suppose, as will nearly always be the case in the applications we consider, that the parameter matrix B_1 corresponds to the covariate $z_{i1} \equiv 1$ for all $i = 1, \ldots, n$. Then we find a matrix $\Gamma \in SO(m)$ to standardize as follows:

$$(B_1\Gamma)_{j\ell} = 0, \quad \ell > j; \qquad (B_1\Gamma)_{\ell\ell} \geq 0, \quad \ell = 1, \ldots, m-1. \tag{7}$$

We briefly show how to determine $\Gamma = [\gamma_1, \ldots, \gamma_m] \in SO(m)$ in non-degenerate cases when

$m = 2$ and $m = 3$. Suppose $B_1 = [a_1, \ldots, a_m]^\top$. When $m = 2$, define

$$\gamma_1 = a_1/||a_1||; \quad \gamma_2 = \pm \left\{a_2 - (a_2^\top \gamma_1)\gamma_1\right\} / ||a_2 - (a_2^\top \gamma_1)\gamma_1||, \tag{8}$$

where the sign of γ_2 is chosen to make the determinant of Γ equal to $+1$. When $m = 3$, define γ_1 and γ_2 as in (8), taking the plus sign in the latter case, and define

$$\gamma_3 = \pm \left\{a_3 - (a_3^\top \gamma_1)\gamma_1 - (a_3^\top \gamma_2)\gamma_2\right\} / ||a_3 - (a_3^\top \gamma_1)\gamma_1 - (a_3^\top \gamma_2)\gamma_2||,$$

where now the sign of γ_3 is chosen so that $|\Gamma| = +1$. A similar type of Gram-Schmidt construction for Γ may be used when $m > 3$.

To use this standardization for a given $\mu_i = \sum_{j=1}^p z_{ij} B_j$, all we need to do is post-multiply μ_i by Γ, which means calculating

$$B_j \mapsto B_j \Gamma, \quad j = 1, \ldots, p. \tag{9}$$

We also recommend adopting some form of centering of the covariates z_{ij}, $2 \leq j \leq p$. If z_{ij} is a continuous covariate, we suggest centering by replacing z_{ij} by $z_{ij} - n^{-1} \sum_{k=1}^n z_{kj}$. In the case of factors, we suggest centering slightly differently, in that we take advantage of structural zeros, as illustrated in the following example.

Example: one-way ANOVA. It is instructive to see how the standardization for factors may be implemented in a one-way ANOVA-type model. Suppose that there are p groups, where group j has n_j observations and $k \times m$ mean configuration A_j, unstandardized at this point. Define $z_{i1} \equiv 1$ and, prior to centering, define

$$z_{ij} = \begin{cases} +1 & \text{if observation } \ i \ \text{ is in group } \ j-1 \\ -1 & \text{if observation } \ i \ \text{ is in group } \ p \\ 0 & \text{otherwise,} \end{cases}$$

for $j = 2, \ldots, p$. The centered z_{ij} for $j = 2, \ldots, p$ are given by

$$z_{ij} = 1 - \frac{n_{j-1} - n_p}{n_{j-1} + n_p} = \frac{2n_p}{n_{j-1} + n_p} \quad \text{or} \quad z_{ij} = -1 - \frac{n_{j-1} - n_p}{n_{j-1} + n_p} = -\frac{2n_{j-1}}{n_{j-1} + n_p} \tag{10}$$

or $z_{ij} = 0$, depending on whether observation i is in group $j-1$, group p, or one of the other groups, respectively. For each $i = 1, \ldots, n$, μ_i, the mean configuration for observation i, is equal to one of $A_1, \ldots, A_p$. Moreover, if μ_i is given by (6) where $z_{i1} = 1$ and the z_{ij} for $2 \leq j \leq p$ are defined in (10), each mean configurations will be equal to one of $A_1, \ldots, A_p$ if we define

$$B_1 = (n_1 + \cdots + n_p)^{-1} \sum_{j=1}^{p} n_j A_j,$$

and for $j = 1, \ldots, p-1$,

$$B_{j+1} = \frac{(n_j + n_p)}{2n_p}(A_j - B_1).$$

The standardization is then applied to $B_1, \ldots, B_p$ using (7) and (9). Our rationale is that it seems preferable to derive the standardization from B_1, which makes the same contribution to all the observations since $z_{i1} \equiv 1$, than to arbitrarily select one of the A_j to standardize.

Finally, we determine the number of free parameters in the regression model (5) and (6) when the standardization (7) and (9) is used. The number of free parameters in B_1 when standardization (7) is used is $km - m(m-1)/2$; the number of free parameters in each of $B_2, \ldots, B_p$ is km; and the number of free parameters in Σ is $k(k+1)/2$. Summing, it is seen that the number of free parameters in the model defined by (5), (6), (7) and (9) is

$$kmp + \frac{1}{2}k(k+1) - \frac{1}{2}m(m-1). \tag{11}$$

3 Marginal likelihood and EM approaches

In §3.1 relevant results from multivariate distribution theory are presented in Lemma 1 and Theorem 1. In §3.2 the relevant marginal likelihood is specified, while a full specification of the EM algorithm is presented in Theorem 2 in §3.3, and further points are briefly discussed in §3.4. Details of how to perform the required computations are given in §4. Parallel results for the case in which size-and-shape is defined to be invariant with respect

to reflections are given in Appendix A.

3.1 Distribution theory

Consider a random $k \times m$ matrix X whose elements are jointly multivariate Gaussian. Then we write $X \sim \mathcal{N}_{k\times m}(\mu, \Sigma_0)$ where $\mu = E[X]$ is the $k \times m$ mean matrix of X and $\Sigma_0 = \mathrm{Cov}\{\mathrm{Vec}(X)\}$ is the $(km) \times (km)$ covariance matrix of $\mathrm{Vec}(X)$, where $\mathrm{Vec}(X)$ is the vector obtained by stacking the columns of X; see e.g. Muirhead (1982, p.74). In what follows, we use $|A|$ to denote the determinant and $\mathrm{tr}(A)$ to denote the trace of a square matrix A.

When $\Sigma_0 = I_m \otimes \Sigma$, where Σ is a $k \times k$ covariance matrix assumed to be of full rank, it is seen using Lemma 2.2.3 of Muirhead (1982) that

$$\phi_{k\times m}(X; \mu, \Sigma_0) = \frac{1}{(2\pi)^{km/2}|\Sigma|^{m/2}} \exp\left[-\frac{1}{2}\mathrm{tr}\left\{(X-\mu)\Sigma^{-1}(X-\mu)^\top\right\}\right], \qquad (12)$$

where $\phi_{k\times m}(X; \mu, \Sigma_0)$ is the probability density function of $\mathcal{N}_{k\times m}(\mu, \Sigma_0)$ with respect to Lebesgue measure (dX) on $\mathbb{R}^{k\times m}$.

Recall that we define the size-and-shape of the pre-form X to be $U\Delta$ where $X = U\Delta R^\top$; see (2) and (3). Inference will be based on the marginal distribution of $U\Delta$ where $X \sim \mathcal{N}_{k\times m}(\mu, I_m \otimes \Sigma)$.

As a first step towards obtaining this marginal distribution we present a result which is essentially Theorem 3.1 of Diaz-Garcia et al (1997); brief additional details of the proof are given in Appendix B.

Lemma 1. *Suppose* $X = (x_{ij})_{i=1,\ldots,k;j=1,\ldots,m} = U\Delta R^\top$ *where* $U \in \mathcal{V}_{k,m}$, $\Delta = diag(\delta_1, \ldots, \delta_m)$ *and* $R \in SO(m)$. *Then Lebesgue measure* (dX) *on* $\mathbb{R}^{k\times m}$ *decomposes according to*

$$(dX) = \mathcal{D}(\Delta)(d\Delta)(dU)(dR), \qquad (13)$$

where $(dX) = \prod_{i=1}^{k}\prod_{j=1}^{m} dx_{ij}$*;* $(d\Delta) = \prod_{j=1}^{m} d\delta_j$*;* (dU) *and* (dR) *are, respectively, the (unnormalized) invariant measures on* $\mathcal{V}_{k,m}$ *and* $SO(m)$*; and*

$$\mathcal{D}(\Delta) = 2^{-m+1}|\Delta|^{k-m}\prod_{i<j}^{m}(\delta_i^2 - \delta_j^2). \tag{14}$$

Using the above lemma we obtain the following result concerning the marginal distribution of $U\Delta$ and the conditional distribution of R given $U\Delta$; details of the proof are given in Appendix B.

Theorem 1. *Suppose that* $X \sim \mathcal{N}_{k\times m}(\mu, I_m \otimes \Sigma)$*, where* $k \geq m$ *and* Σ *has full rank* k*. Consider the singular value decomposition* $X = U\Delta R^\top$ *given by (3). Then the density* $f_1(U, \Delta; \mu, \Sigma)$ *with respect to the measure* $(d\Delta)(dU)$ *defined via Lemma 1, is given by*

$$f_1(U, \Delta; \mu, \Sigma) = \frac{\mathcal{D}(\Delta)\mathcal{C}(A)}{(2\pi)^{km/2}|\Sigma|^{m/2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\Delta U^\top \Sigma^{-1} U\Delta + \mu^\top\Sigma^{-1}\mu\right)\right\}, \tag{15}$$

where $\mathcal{D}(\Delta)$ *is defined in (14),* $A = \mu^\top\Sigma^{-1}U\Delta$ *is an* $m \times m$ *matrix and* $\mathcal{C}(A)$ *is defined by*

$$\mathcal{C}(A) = \int_{R\in SO(m)} \exp\left\{\mathrm{tr}\left(RA^\top\right)\right\}(dR). \tag{16}$$

Moreover, the conditional distribution of R *given* U *and* Δ *has density* $f_2(R|A)$ *with respect to the unnormalized geometric, or Haar, measure* (dR) *on* $SO(m)$ *given by*

$$f_2(R|A) = \mathcal{C}(A)^{-1}\exp\left\{\mathrm{tr}\left(RA^\top\right)\right\}. \tag{17}$$

Note that (17) is the Fisher matrix distribution, see e.g. Mardia and Jupp (2000), but defined on the special orthogonal group, $SO(m)$, rather than the orthogonal group, $O(m)$.

3.2 The Marginal Likelihood

In this paper we base inference for μ on the size-and-shape information in the observed pre-forms $X_1, \ldots, X_n$. Therefore we should base inference on U_i and Δ_i, $i = 1, \ldots, n$,

where these matrices are extracted from X_i using either (2) or (3), depending on whether we want to retain the reflection information as part of the shape. As above, we continue to focus on (3), in which the orientation information $R_i \in SO(m)$ does not include reflection information. The marginal log-likelihood, ℓ_M, based on the sample (U_i, Δ_i), $i = 1, \ldots, n$, is given by

$$\ell_M(B, \Sigma) = \sum_{i=1}^{n} \log f_1(U_i, \Delta_i; \mu_i, \Sigma), \qquad \mu_i = \sum_{j=1}^{p} z_{ij} B_j, \tag{18}$$

where the marginal density f_1 is defined in (15) and μ_i is defined in the same way as in (6). As we shall see, to calculate ℓ_M a method for numerical evaluation of the normalizing constant (16) is needed; we discuss how to do this in §4.

When maximizing the marginal likelihood, it is important to standardize the B_j as explained in §2.3; see (7).

3.3 The EM Algorithm

An alternative possibility to direct maximization of (18) is to use the EM algorithm for maximizing log-likelihoods when there is missing data; see Dempster et al. (1977) and McLachlan & Krishnan (1997). In the setting of the model (5) and (6), we treat the R_i as missing data. Using (5), and ignoring a constant, the full log-likelihood, ℓ_F, is given by

$$\ell_F(B, \Sigma) = -\frac{nm}{2} \log |\Sigma| - \frac{1}{2} \sum_{i=1}^{n} \operatorname{tr} \left\{ (X_i - \mu_i)^\top \Sigma^{-1} (X_i - \mu_i) \right\}, \tag{19}$$

where μ_i is given by the linear model (6), and

$$B = [B_1, \ldots, B_p]. \tag{20}$$

The implementation of the EM algorithm is summarized in the following updating rule for estimators of B and Σ.

Let $(B^{(r)}, \Sigma^{(r)})_{r\geq 0}$ denote the sequence derived from the EM algorithm for estimating B and Σ in the log-likelihood (18). For $i = 1, \ldots, n$ define

$$\bar{R}_i^{(r)} = E[R_i | U_i, \Delta_i; B^{(r)}, \Sigma^{(r)}], \tag{21}$$

where U_i, Δ_i and R_i are determined using (2) or (3). Write

$$\bar{X}_i^{(r)} = U_i \Delta_i \bar{R}_i^{(r)\top}, \quad i = 1, \ldots, n, \tag{22}$$

and define the $n \times p$ matrix $Z = (z_{ij})$, the $n \times p$ matrix $A = (a_{ji})$ and the $n \times n$ matrix $P = (p_{ij})$ by

$$Z = [z_1, \cdots, z_n]^\top, \quad A = Z(Z^\top Z)^{-1} \quad \text{and} \quad P = Z(Z^\top Z)^{-1} Z^\top. \tag{23}$$

Also, for $r \geq 0$, define the $k \times (mn)$ matrices Y and $\bar{Y}^{(r)}$ and the $k \times (mp)$ matrix $B^{(r)}$ by

$$Y = [X_1, \ldots, X_n], \quad \bar{Y}^{(r)} = \left[\bar{X}_1^{(r)}, \ldots, \bar{X}_n^{(r)}\right], \quad B^{(r)} = [B_1^{(r)}, \ldots, B_p^{(r)}]. \tag{24}$$

The key result which describes the updating rule in the EM algorithm is now stated.

Theorem 2. *Assume that $n \geq p$ and that Z in (23) has full rank p. Then, given a starting value $B^{(0)} = [B_1^{(0)}, \ldots, B_p^{(0)}]$ corresponding to $r = 0$ in (24), the EM updating rule for calculating the sequence $(B^{(r)}, \Sigma^{(r)})_{r\geq 0}$ is given by*

$$B^{(r+1)} = \bar{Y}^{(r)}(A \otimes I_m) \tag{25}$$

and

$$\Sigma^{(r+1)} = \frac{1}{mn}\left\{YY^\top - \bar{Y}^{(r)}(P \otimes I_m)\bar{Y}^{(r)\top}\right\}, \tag{26}$$

where Y and $\bar{Y}^{(r)}$ are defined in (24). Moreover, the updating rules (25) and (26) are equivalent to

$$B_j^{(r+1)} = \sum_{i=1}^{n} a_{ij} \bar{X}_i^{(r)}, \quad j = 1, \ldots, p, \tag{27}$$

and

$$\Sigma^{(r+1)} = \frac{1}{mn}\left\{\left(\sum_{i=1}^{n} X_i X_i^\top\right) - \sum_{i=1}^{n}\sum_{j=1}^{n} p_{ij}\bar{X}_i^{(r)}\bar{X}_j^{(r)\top}\right\}, \tag{28}$$

where the a_{ij} and p_{ij} are, respectively, the components of the matrices A and P defined in (23), and the $\bar{X}_i^{(r)}$ are defined in (22).

Remark 3.1. An important point to note is that the marginal density (15) and the updating rules in Theorem 2 are invariant with respect to arbitrary transformations of the form $X_i \mapsto X_i S_i^\top$, where $S_i \in SO(m)$, $i = 1, \ldots, n$. This is because, although each $\bar{X}_i$ depends on U_i and Δ_i, it does not depend on R_i.

Remark 3.2. In order to remove orientation information, we should standardize the B_j, as explained in §2.3, at each step (25) or (27).

Remark 3.3. If in equations (27) and (28) the updating quantities satisfy $B_j^{(r+1)} = B_j^{(r)} = B_j$ and $\Sigma^{(r+1)} = \Sigma^{(r)} = \Sigma$, then the resulting equations imply that the marginal likelihood function is stationary at B_j and Σ. Such a solution is obtained at the EM convergence point.

3.4 Relevant theory for EM

The EM algorithm is a method for maximising the marginal likelihood estimator of B and Σ and, obviously, if the EM and maximum likelihood estimators agree they will have the same asymptotic behaviour and, in particular, the same asymptotic covariance matrix. See Sundberg (1974), Dempster et al. (1977), Wu (1983) and McLachlan and Krishnan (1997) for relevant theoretical results and examples. Specifically, assumptions (5), (6), (7) and (9) in Wu (1983) hold, and consequently the results in Wu (1983) are applicable here. The key result concerning convergence is summarized in the opening paragraph of Section 2.1 of Wu (1983): that the EM algorithm will converge to a point that is not necessarily a global or even a local maximum of the marginal likelihood $L(\theta) = f(x|\theta)$; all we can say with

full confidence is that the EM algorithm terminates at a stationary point of $L(\theta)$. In the numerical calculations performed for this paper, we did not notice a problem of this type occurring, but Wu's result mentioned above does indicate that in general some caution is necessary when using the EM algorithm.

3.5 Standard errors

The common asymptotic covariance matrix can be calculated by finding the Hessian, with respect to the elements of B and Σ, of minus the marginal log-likelihood given in (18). Calculation of this Hessian is fairly long though relatively compact formulae are given in Theorem SM1 in the Supplementary Material. We have not included these formulae in the main body of the paper as in practice it is easier to calculate second numerical partial derivatives; the two methods give results which are numerically close but the numerical partial derivatives version is easier to implement.

4 Procrustes Connections

In this section we investigate the similarities and differences of the EM algorithm described in Theorem 2 and the Procrustes approach to estimation of mean size-and-shape in the simplest case of our model: independent and identically distributed with scalar covariance matrix $\Sigma = \sigma^2 I_k$. In §5.1 and §5.2 we briefly consider the independent and identically distributed models and scalar covariance models, respectively, giving the simplified versions of the updating formulae in each case, as they are of independent interest. Then, in §5.3, we explain the Procrustes connections in this simplified context. Our analysis makes it transparent why Procrustes is consistent when $m = 2$ but not when $m \geq 3$.

4.1 The IID case

In the IID case we have $p = 1$ and $z_{i1} = 1$ for $i = 1, \ldots, n$. Put $\mu = B_1$. The updating formulae (27) and (28) in Theorem 2 in this case simplify to

$$\mu^{(r+1)} = \frac{1}{n} \sum_{i=1}^{n} \bar{X}_i^{(r)} \tag{29}$$

and

$$\Sigma^{(r+1)} = \frac{1}{nm} \left\{ \left(\sum_{i=1}^{n} X_i X_i^\top \right) - n\mu^{(r+1)} \mu^{(r+1)\top} \right\}, \tag{30}$$

where, at each iteration, $\mu^{(r)}$ in (29) should be standardized as explained in §2.3.

4.2 The scalar variance case: $\Sigma = \sigma^2 I_k$

In this case, the updating formula (27) is unchanged, but with $\Sigma^{(r)} = (\sigma^2)^{(r)} I_k$ in the calculation of the $\bar{X}_i^{(r)}$ in (22); while the updating rule (28) simplifies to

$$(\sigma^2)^{(r+1)} = \frac{1}{nmk} \mathrm{tr} \left(\sum_{i=1}^{n} X_i X_i^\top - \sum_{i=1}^{n} \sum_{j=1}^{n} p_{ij} \bar{X}_i^{(r)} \bar{X}_j^{(r)\top} \right). \tag{31}$$

In the independent and identically distributed case with scalar covariance matrix, (31) simplifies further to

$$(\sigma^2)^{(r+1)} = \frac{1}{nmk} \left\{ \mathrm{tr} \left(\sum_{i=1}^{n} X_i X_i^\top \right) - n\mathrm{tr} \left(\mu_i^{(r+1)} \mu_i^{(r+1)\top} \right) \right\}.$$

In some situations, especially when k is relatively large, we may wish to fit a sparse regression model, e.g. using some version of the lasso. In such a situation, in order to simplify the computations, it may be worth considering the case where $\Sigma = \sigma^2 I_k$, a scalar multiple of the identity matrix.

4.3 EM and the Procrustes algorithm

Here, $\Sigma = \sigma^2 I_k$, and the parameters of interest are only the shape of μ and σ^2. One can easily see that expectations (22) imply that (29) and (30) are reduced to

$$\mu^{(r+1)} = \frac{1}{n}\sum_{i=1}^{n} U_i \Delta_i \bar{R}_i^{(r)\top} = \frac{1}{n}\sum_{i=1}^{n} U_i \Delta_i T_{2i}^{(r)} E[\bar{R}|\Phi_i^{(r)}] T_{1i}^{(r)\top} \tag{32}$$

and

$$(\sigma^2)^{(r+1)} = \frac{1}{mnk}\sum_{i=1}^{n} \mathrm{tr}\left(\Delta_i^2 - \mu^{(r+1)}\mu^{(r+1)\top}\right) \tag{33}$$

where

$$T_{1i}^{(r)} \Phi_i^{(r)} T_{2i}^{(r)\top} = \frac{1}{(\sigma^2)^{(r)}} \mu^{(r)\top} U_i \Delta_i, \tag{34}$$

and the left-hand side of (34) is the singular value decomposition of the right-hand side of (34).

Let us now consider the Procrustes estimator of μ. For fixed μ, we find $\Gamma_i \in SO(m)$, $i = 1, \ldots, n$, to minimise

$$\begin{aligned}\sum_{i=1}^{n} ||X_i\Gamma_i - \mu||^2 &= \sum_{i=1}^{n} \mathrm{tr}\left\{(X_i\Gamma_i - \mu)^\top (X_i\Gamma_i - \mu)\right\} \\ &= \sum_{i=1}^{n} \left[\mathrm{tr}(X_i^\top X_i) + \mathrm{tr}(\mu^\top\mu) - 2\mathrm{tr}(\mu^\top U_i \Delta_i R_i^\top \Gamma_i)\right],\end{aligned}$$

where we write $X_i = U_i \Delta_i R_i^\top$ using (3). Note that we do not include location in the above optimization because the X_i have already been Helmertized, so that the effects of location have already been removed. The above expression is minimized over the Γ_i for fixed μ when the $\mathrm{tr}(\mu^\top U_i \Delta_i R_i^\top \Gamma_i)$ are maximized. Using the singular value decomposition for $\mu^\top U_i \Delta_i$ implied by (34), namely

$$\mu^\top U_i \Delta_i = T_{1i}\left(\sigma^2 \Phi_i\right) T_{2i}^\top,$$

it is seen that the optimum choice for Γ_i occurs when

$$T_{21}^\top R_i^\top \Gamma_i T_{1i} = I_m,$$

which yields

$$\Gamma_i = R_i T_{2i} T_{1i}^\top, \quad i = 1, \ldots, n.$$

The alternating iterative Procrustes procedure for estimating μ therefore has the following updating rules:

$$\Gamma_i^{(r+1)} = R_i T_{2i}^{(r)} T_{1i}^{(r)\top}, \quad i = 1, \ldots, n;$$

$$\mu^{(r+1)} = \frac{1}{n}\sum_{i=1}^{n} X_i \Gamma_i^{(r)} = \frac{1}{n}\sum_{i=1}^{n} X_i R_i T_{2i}^{(r)} T_{1i}^{(r)\top} = \frac{1}{n}\sum_{i=1}^{n} U_i \Delta_i T_{2i}^{(r)} T_{1i}^{(r)\top}. \tag{35}$$

Note that if we used for this EM implementation the Helmertized landmarks X_i and if in (32) the expectation $E[R|\sigma^{2(r)}\Phi_i^{(r)}]$ is replaced by I_m, then $X_i R_i T_{2i}^{(r)} T_{1i}^{(r)\top}$ will simply be the pre-form X_i optimally rotated to the current estimate for the mean $\mu^{(r)}$.

The diagonal matrix $E[R|\sigma^{2(r)}\Phi_i^{(r)}]$ performs some sort of shrinking of the Procrustes update steps, since $I_m - E[R|\sigma^{2(r)}\Phi_i^{(r)}]$ is necessarily non-negative definite. Therefore at convergence, equations (32) and (35) will generate estimates for μ such that the norm of the Procrustes mean is larger that that of the EM mean. As a result, the corresponding estimator for σ^2 determined by (33) will produce a smaller value if the Procrustes mean is used for $\mu^{(r)}$. Table 1 confirms this numerically for $m = 3$. In fact this shrinking action looks different for $m = 2$ and $m \geq 3$ cases. When $m = 2$, $E[R|\sigma^{2(r)}\Phi_i^{(r)}]$ is a scalar multiple of identity; consider (A3) in Appendix A when the matrix on the far right is diagonal. Specifically, for $m = 2$, the shrinking is introducing some constant multiplication for each term in summation (35). When $m \geq 3$, this expectation is not a multiple of identity except in special cases. From standard large sample results, the maximum likelihood estimator provides a consistent estimator of μ when the model is correct, so in general we expect the Procrustes estimator to be different. The only time when they may produce similar results is when the conditional distributions of the R_i are very highly concentrated about the identity, in which case the $||I_m - E[R|\sigma^{2(r)}\Phi_i^{(r)}]||$ will be small.

In this section we have not so far mentioned the need to standardize (32) and (35) at

each iteration. However, this need for standardization is the same in both cases and does not affect our discussion of the Procrustes connections.

4.4 Simulation results

We now present simulation results in Table 1 in which the Procrustes estimator μ is compared with the EM estimator. In the examples considered, $m = 3$, $k = 3$ and $\Sigma = \sigma^2 I_k$ where $\sigma^2 > 0$ is a constant.

It is clear from Table 1 that the EM-based estimator, $\hat{\mu}_{EM}$, is generally more accurate than the Procrustes estimator, $\hat{\mu}_p$, in the size-and-shape distance sense. Specifically, Table 1 shows there is a substantial improvement in the former estimators as the sample size n increases from 20 to 1000 for each fixed σ^2, while there is relatively little change in the Procrustes estimators as n increases, which supports the theoretical statements concerning consistency made in §5.3. Two further tables for different choices of landmark means are given in the Supplementary Material where it is seen that the difference in performance between the Procrustes and EM estimators is typically smaller if the three eigenvalues of the mean configuration are closer to each other and the value of σ is relatively small.

We also illustrate graphically the difference between the EM and Procrustes means for a similar model to that used in Table 1. We consider 40 mean configurations which interpolate in equal steps between $\mu \propto \text{diag}\{60, 50, 1\}$ and $\mu \propto \text{diag}\{60, 50, 45\}$ by only varying the smallest eigenvalue. We initially rescale each of these 40 means so that their configuration size is 1, and then generate $n = 2500$ random samples with $\Sigma = \sigma^2 I_k$ for $\sigma = 0.1, 0.2, 0.3$ and 0.4. The red spheres (the 40 EM means) in Figure 1 mainly block out the green spheres (the 40 exact means), while the blue spheres (the 40 Procrustes means) tend to be further away from the green spheres; the larger σ, the further away from the green spheres the blue spheres tend to be. In Figure 1 a rotation standardization was carried out such that the first three landmarks lie in a plane, with the first landmark at the origin, the second landmark varies along a fixed axis in the plane, the third landmark is

n and σ	$\rho_s(\hat{\mu}_p, \mu)$	$\rho_s(\hat{\mu}_{EM}, \mu)$	$\rho(\hat{\mu}_p, \mu)$	$\rho(\hat{\mu}_{EM}, \mu)$	σ_p	σ_{EM}
n= 20 sig= 0.1	0.0832	0.0831	0.0828	0.0827	0.0741	0.0741
n= 50 sig= 0.1	0.0406	0.0406	0.0386	0.0385	0.0782	0.0783
n= 100 sig= 0.1	0.0412	0.0411	0.0412	0.0411	0.076	0.0761
n= 1000 sig= 0.1	0.038	0.0379	0.0322	0.0321	0.0795	0.0796
n= 20 sig= 0.3	0.4802	0.3511	0.3515	0.2804	0.2088	0.2644
n= 50 sig= 0.3	0.2937	0.1399	0.2671	0.1044	0.2302	0.3091
n= 100 sig= 0.3	0.3627	0.2341	0.3021	0.2242	0.2127	0.2706
n= 1000 sig= 0.3	0.3057	0.0834	0.244	0.0831	0.2317	0.3067
n= 20 sig= 0.8	1.6382	1.2549	0.5048	0.461	0.5372	0.689
n= 50 sig= 0.8	1.339	0.5054	0.6736	0.4609	0.5292	0.7526
n= 100 sig= 0.8	1.328	0.2537	0.6605	0.2544	0.5603	0.8042
n= 1000 sig= 0.8	1.3065	0.1498	0.643	0.1493	0.5735	0.8071

Table 1: The mean values after 1000 runs for EM and Procrustes mean quantities. For each run we simulate data for fixed sample size n and true value of σ as above; ρ_s and ρ represent the Riemmanian distances of size-and-shape and shape space respectively; and the choice of population mean here is $\mu \propto \text{diag}\{60, 10, 1\}$, scaled so that $||\mu|| = 1$.

allowed to move freely in the plane, while the fourth landmark is allowed to move freely in $3D$ space. Further numerical results are presented in part D of the Supplementary Material.

5 Applications of shape regression model

In the regression setting suppose that we have labelled landmark observations $X_1, \ldots, X_n$ in the pre-form space and corresponding covariate vectors $z_1, \ldots, z_n$. In this section we look at two datasets; the well-known rat growth data dataset where $m = 2$ (see, for

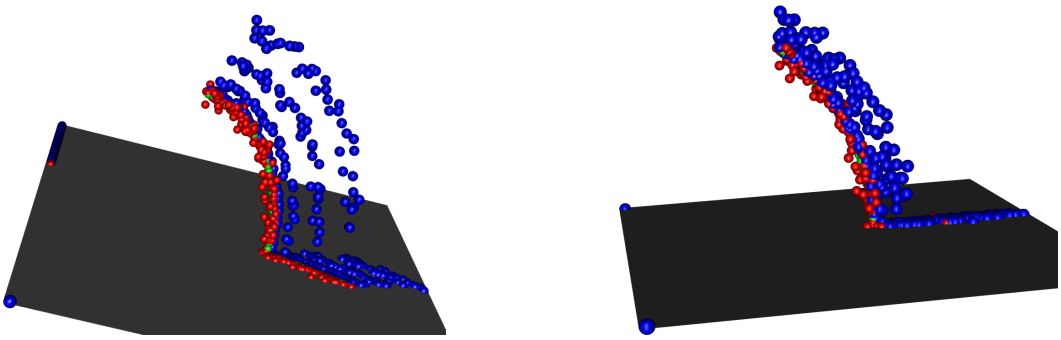

Figure 1: A collection of slowly changing 40 mean simplexes represented by green spheres (mostly covered by the red ones). Each of the means were perturbed by i.i.d. noise and the corresponding EM means (red) and Procrustes means (blue) are shown. The size and shape coordinates (shown on the left) are obtained while the first three landmarks are forced to be on the same plane. The four strands of blue spheres there represent the choices of $\sigma = 0.1, 0.2, 0.3$ and 0.4, with the bias increasing with σ. The shape coordinates (shown on the right) are obtained by standardising additionally for scale so that the first two landmarks are matching the two (blue) points defining the left edge of the shaded area.

example, Dryden and Mardia, 2016, p22) and a human movement dataset where $m = 3$; this experiment was described by Kume et al. (2007) and Alshabani et al. (2007).

5.1 Regression in $m = 2$ Case: Rat Skulls Data

This data set is introduced by Bookstein (1991) and is studied by several other authors in different contexts. For example, Le and Kume (2000b) consider fitting geodesics curves to the corresponding shapes while size is ignored; Kent et al. (2001) consider fitting growth curves for these data in the Procrustes tangent space while size is used as a covariate; Kenobi et al. (2010) consider fitting shape curves defined as projections from the size-and-shape space. We apply our landmark regression model to the size-and-shape response data; size-and-shape is most appropriate to this data as changes in size cannot be treated

separately from the shape. A related model but only using shapes is considered in Kume and Welling(2010).

After removing entries with missing data there are 18 individual rat skulls whose 8 biological landmarks are observed at 8 ages. We have observations at days 7, 14, 21, 30, 40, 60, 90 and 150. Due to the uneven spacing between the days at which the landmarks were recorded we take the logarithm of the observation days.

5.1.1 Model

Let $X_{i,r} \in \mathcal{R}^{7\times 2}$ be the Helmertised configuration for the individual rat $r = 1, \ldots, 18$ at times $i = 1, \ldots, 8$. The design matrix $Z = (z_{ij})$ for the polynomial model of order p is given by

$$\begin{aligned} &Z = (z_{ij} = t_i^{j-1} : i = 1, \ldots, 8; j = 1, \ldots, p) \\ &t_1 = \log(7),\ t_2 = \log(14),\ t_3 = \log(21),\ t_4 = \log(30), \\ &t_5 = \log(40),\ t_6 = \log(60),\ t_7 = \log(90),\ t_8 = \log(150); \end{aligned} \tag{36}$$

and so for the quadratic model, $p = 3$, Z has three columns while the linear model is its submodel that includes only the first two columns of Z. The regression model is

$$X_{i,r}|z_i \sim \mathcal{N}_{7\times 2}\left(\mu_i, \sigma^2 I_2 \otimes I_7\right), \quad i = 1, \ldots, 8; r = 1, \ldots, 18,$$

where $z_i = (1, t_i, \ldots, t_i^{p-1})$. We also generalise the covariance from $\sigma^2 I_2 \otimes I_7$ to $\sigma^2 I_2 \otimes \Sigma$ as in (5), with Σ a general 7×7 covariance matrix. The mean function is the linear combination

$$\mu_i = \sum_{j=1}^{p} z_{ij} B_j, \quad i = 1, \ldots, 8,$$

where the B_j are 7×2 matrices of parameters and the covariance matrix is a scalar, σ^2 times the identity matrix. We fit the linear mean model, $p = 2$, the quadratic mean model, $p = 3$ and the cubic mean model $p = 4$specified by the design matrix (36).

5.1.2 Results

We fit the constant, linear and quadratic models using both the EM algorithm and Newton-Raphson to maximise the marginal likelihood. In Table 2 we display the log-likelihood values at the maximum likelihood estimators. For both estimation procedures it is clear that with 14 degrees of freedom we reject the null hypotheses of the constant mean model, linear mean model and quadratic mean model in favour of the cubic mean model with a separable covariance structure. Figure 2 visually confirms that the cubic mean model is appropriate for this dataset. More specifically, the mean paths represented by black lines and evaluated for the cubic mean model are seen to be in closer agreement with the data than the corresponding paths from the alternative models considered. This difference is more pronounced for the top part of the skull. The sample paths generated for the fitted model vary less for generalized covariance model (see the grey region for the lower landmarks), indicating that that the more general covariance structure does a better job of capturing the landmark dynamics.

	$\sigma^2 I_2 \otimes I_7$	DF	$I_2 \otimes \Sigma$	DF
Constant Model	- 10307.42	13	-4358.03	30
Linear Model	-7170.76	27	-3875.20	54
Quadratic Model	-6807.33	41	-3812.36	68
Cubic Model	**-6710.52**	55	**-3765.37**	82

Table 2: This table displays the value of the maximised log-likelihood from the EM algorithm in column 'EM' for linear, quadratic and cubic models for the rat skull data for both types of covariance structures. Column 'DF' gives the corresponding model degrees of freedom.

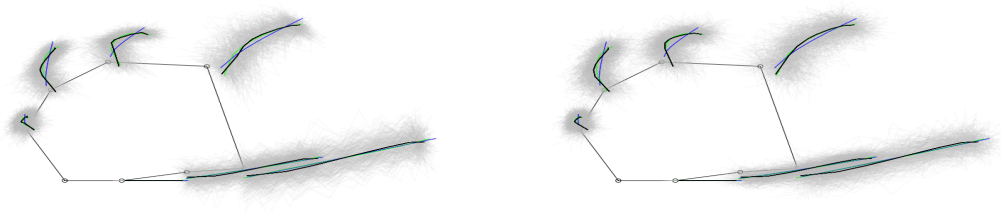

Figure 2: The fitted cubic mean model for isotropic covariance (left) and the fitted cubic mean model with a general covariance (right) are represented by the solid (black) lines and observations are in blue. Simulated data consisting of 500 sample paths from the respective models are shown in grey. The size-and-shapes are standardized using Bookstein coordinates, i.e. one landmark is fixed to zero while another one is allowed to vary only along one side of the horizontal axis. The mean paths for the corresponding linear and quadratic mean models are shown in blue and green respectively.

5.2 Regression in $m = 3$ Case: Human Movement Data

The data set considered here contains the records of 14 individuals performing a pointing action using the index finger to a particular target point and then back to the original position. In fact we have only four landmark locations observed during the pointing action and back. These landmarks observed in 1440 equally spaced time intervals are the shoulder, elbow, index finger and the lower back. The shape of the corresponding tetrahedra is changing over time. During these 1440 time observation points there is some inconsistency across the subjects as to when and how long the pointing action took place. In order to simplify the analysis we have chosen to manually assign the start and the finishing times of the action and consider for each individual only 200 intermediate observations. As a result we have for 14 subjects the size-and-shape observations in 200 equally spaced intervals. We study the landmarks for a particular individual and it can be seen by plotting the observations that each landmark follows a nearly closed curved trajectory.

5.2.1 Model

Let $X_i \in \mathbb{R}^{3\times 3}$ be the Helmertized configuration for $i = 1, \ldots, 200$ and $z_{i,j}$ is a vector of covariates associated with the configuration X_i and model j, $j = 1, 2, 3$. The covariates are described for each proposed model below. Let

$$X_i | z_{i,j} \sim \mathcal{N}_{3\times 3}\left(\mu_{i,j}, \sigma^2 I_3 \otimes I_3\right), \quad i = 1, \ldots, 200$$

and $j = 1, 2, 3$ represent three nested models. Define the mean function in each case for $i = 1, \ldots, 200$ by

$$\mu_{i,1} = B_1 + t_i B_2, \tag{37}$$

$$\mu_{i,2} = B_1 + t_i B_2 + t_i^2 B_3, \tag{38}$$

$$\mu_{i,3} = B_1 + t_i B_2 + t_i^2 B_3 + t_i^3 B_4. \tag{39}$$

The unknown parameter matrices B_r are real 3×3 matrices for $r = 1, \ldots, 4$, but with B_1 standardized so that the 3 elements in the upper triangle are zero. The 3×3 covariance matrix Σ is a function of 6 unknown parameters. Numerical studies not reported here suggest assuming that the variance is a scalar multiple of the identity is too restrictive. Finally, t_{200} is the time of the final observation in the pointing loop. From here we are going to refer to the three models as the linear (37), quadratic (38) and cubic (39) models.

	EM	DF
Constant Model	-134363.4	6
Linear Model	-128991.9	15
Quadratic Model	-119266.1	24
Cubic Model	-118972.5	33

Table 3: This table displays the value of the maximised marginal log-likelihood using the EM algorithm. The models fitted are the constant mean, linear mean, quadratic mean and cubic mean models for the human movement data.

5.2.2 Results

We fit the three models (linear, quadratic and cubic) using the EM algorithm. The maximised log-likelihood values are displayed in Table 3. Application of the standard large-sample log-likelihood ratio test indicates that, we should reject the simpler constant, linear and quadratic models in favour of the cubic model. Figure 3 shows the fitted values of the individual trajectories of the original data superimposed on the trajectory of the cubic model after a convenient standardization is imposed. Specifically, location is standardized by fixing landmark 1 to the origin of coordinates; and rotation standardization is carried out on the remaining 3 landmarks by using a Gram-Schmidt construction similar to that

adopted in Section 2.3, such that landmark two is allowed to vary along one side of the x-axis, and landmark three is varying in the x,y plane, which is highlighted in Figure 3.

Figure 3 compares the fitted values obtained from the quadratic and cubic mean models, and especially the latter, does a reasonably good job of representing the original data since the observed data paths are closed curves in size-and-shape space.

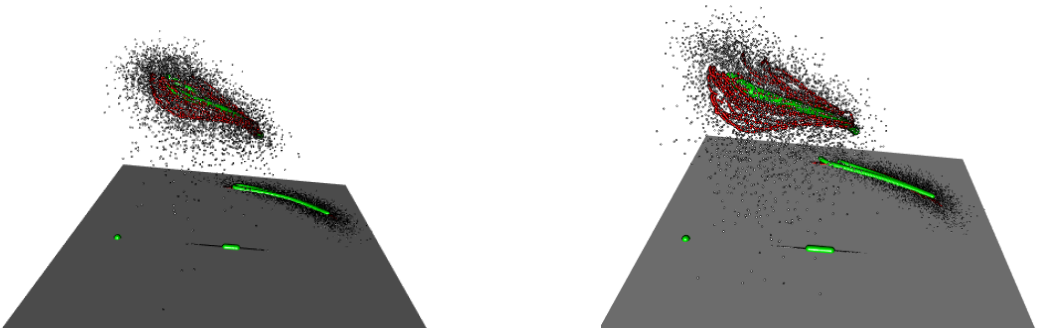

Figure 3: The fitted polynomial mean paths (qubic-left and quadratic-right) in green, observations are in red; the rotation standartisation is obtained by fixing landmark 1 to the origin, landmark 2 is allowed to vary only along a chosen axis and landmark 3 is varying only in the standardizing plane (the shaded region), landmark 4 is allowed to freely vary in 3-d space. Simulated data from the fitted models are shown in black.

5.3 A Goodness of fit analysis

For the dataset considered in Section 5.1 and the dataset considered in Section 5.2, we selected a fitted model in each case using maximum likelihood. In order to test the goodness-of-fit for each model we considered two approaches. First, we implemented a *parametric bootstrap* method (see Appendix B for practical details and theoretical justification). In each case, we simulated 1000 random paths and for each of them we found the corre-

sponding MLE estimates. The 1000 optimal likelihood values at such estimates were then compared to those of the fitted parameters to the real data. For the 3-d real data example, we observed that 46.2% of the values were above the observed MLE value, corresponding to a two-sided p-value of 92.4%. This suggests a good fit; and for the 2-d data example the corresponding two-sided p-value is about 5.8%.

The second approach was to graphically assess the goodness of fit by generating some random data from the models and then by visually inspecting the sample variation from the fitted models with that of the data. In particular, for both models of Figure 2, the grey cloud of the paths represent the 500 simulated paths from the fitted models. In Figure 2, the plot on the left corresponds to an isotropic covariance matrix while the plot on the right corresponds to a general covariance matrix. The observed data do not seem out of line with the general sample variation from the fitted model. A similar graphical display is also present in Figure 3 for the 3-d data example.

6 Concluding remarks

The development of regression models in the analysis of shape and size-and-shape data is an important problem in object data analysis. In the paper we have developed the following approach: use labelled landmarks to describe the size-and-shape of an object; start with a Gaussian model for landmarks; project the size-and-shape configurations (see §2 for definitions) onto the relevant size-and-shape space; determine the induced regression model on the size-and-shape space and use this to perform estimation and inference. The last step is technically quite challenging but useful progress can be made under the assumption of Gaussian landmarks; see Theorem 1. As an alternative to direct maximisation of the marginal likelihood, we develop an EM approach in which all information in each configuration which is not relevant to size-and-shape is discarded and treated as missing data. The resulting EM procedure is described in Theorem 2. From a practical point of view we

prefer to use the EM procedure rather than direct maximization of the marginal likelihood.

This regression model, fitted using the EM algorithm, has been used in various numerical studies. In §4 we study the Procrustes approach (see Dryden and Mardia, 2016, p.72) and use a simple case of our model to provide new insights into why the Procrustes approach typically fails when variability in the configurations increases. Moreover, it is shown in §5, through the analysis of a rat skull dataset and a human movement data set, that our approach provides a valuable and tractable methodology for regression modelling of real size-and-shape data in what is a challenging and highly nonlinear setting.

Acknowledgements

This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/K022547/1]. The authors are extremely grateful to the Associate Editor and the reviewers for their constructive suggestions.

Appendix A: Numerical computation of $\mathcal{C}$ and its derivatives

We now discuss how to perform the most challenging computational steps when determining the maximum likelihood estimators of $B_1, \ldots, B_p$ and Σ and the observed information matrix: namely, the evaluation of $\mathcal{C}(A)$ defined in (16) and its first partial derivatives with respect to elements of A. In this section we consider $R \in SO(m)$ when $m = 2$ and $m = 3$, corresponding to objects in 2 and 3 dimensions. Similar but slightly more complicated formulae are given in Section SM1 of the Supplementary Material for the case where $R \in O(m)$.

A.1: Calculations when $m = 2$

When $m = 2$, using a standard parametrisation of $SO(2)$, namely

$$R = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$$

we have

$$\mathrm{tr}\left(RA^\top\right) = (a_{11} + a_{22})\cos(\theta) + (a_{12} - a_{21})\sin(\theta).$$

Therefore in this case (17) is a von Mises distribution on the circle, see e.g. Mardia and Jupp (2000), and hence the normalising constant and its first partial derivatives can be expressed in terms of the modified Bessel function of the first kind, $\mathcal{I}_\nu$; see e.g. Abramowitz and Stegun (1972). Specifically,

$$\mathcal{C}(A) = \int_{\theta=0}^{2\pi} \exp\left\{(a_{11} + a_{22})\cos(\theta) + (a_{12} - a_{21})\sin(\theta)\right\} d\theta = 2\pi\mathcal{I}_0(\rho), \tag{A1}$$

where $\rho = \left\{(a_{11} + a_{22})^2 + (a_{12} - a_{21})^2\right\}^{1/2}$ and $\mathcal{I}_0$ is the modified Bessel function of the first kind of degree zero.

To perform the updates (25)-(28) it is necessary to calculate, at iteration r, $\bar{X}_i = U_i\Delta_i\bar{R}_i^{(r)\top}$, $i = 1, \ldots, n$, using (21) to calculate $\bar{R}_i^{(r)}$.

The required result in the case $m = 2$ is summarized in the following lemma.

Lemma A.1 *Suppose* $m = 2$ *and define the* 2×2 *matrix*

$$M = (m_{ij})_{i,j=1}^{2} = \mu^{\top}\Sigma^{-1}U\Delta. \tag{A2}$$

Then

$$\bar{R} = E_{\mu,\Sigma}[R|U,\Delta] = \int_{R\in SO(2)} Rf_2(R|U,\Delta;\mu,\Sigma)(dR) = \mathcal{A}(\rho)\begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}, \tag{A3}$$

where $\cos\alpha = (m_{11} + m_{22})/\rho$, $\sin\alpha = (m_{12} - m_{21})/\rho$,

$$\rho = \sqrt{(m_{11} + m_{22})^2 + (m_{12} - m_{21})^2}$$

and $\mathcal{A}(\rho) = \mathcal{I}_1(\rho)/\mathcal{I}_0(\rho)$.

To calculate a particular $\bar{R}_i^{(r)}$, substitute $\mu = \mu_i = \sum_{i=1}^{n} z_{ij}B_j^{(r)}$, $\Sigma = \Sigma^{(r)}$, $U = U_i$ and $\Delta = \Delta_i$ in (A2) and then use (A3).

A.2: Calculations when $m = 3$.

The case $m = 3$ is more challenging though numerical procedures for doing the computations approximately are available. We make use of the relationship between the normalizing constant $\mathcal{C}(A)$ in (16) and the normalizing constant of the Bingham distribution on the sphere; see Prentice (1984) and Wood (1993). Define the Bingham normalization constant on the unit sphere $\mathcal{S}^q = \{x \in \mathbb{R}^q : x^{\top}x = 1\}$ by

$$\mathcal{C}_q(\Lambda) = \int_{x\in\mathcal{S}^{q-1}} \exp\{x^{\top}\Lambda x\}[dx], \tag{A4}$$

where $[dx]$ denotes unnormalized geometric measure on the unit sphere $\mathcal{S}^{q-1}$. The relevant cases here are $q = 4$ and $q = 6$. The case $q = 4$ gives the normalising constant itself. Given a 4×4 matrix $\Xi = \mathrm{diag}\{\xi_1, \xi_2, \xi_3, \xi_4\}$, we define the 6×6 matrices

$$\Xi_j = \mathrm{diag}\{\xi_1, \xi_2, \xi_3, \xi_4, \xi_j, \xi_j\}, \quad j = 1, 2, 3, 4. \tag{A5}$$

We note the following useful fact: since the density function $f_2(R|A)$ of the Fisher matrix distribution in (17) is natural exponential family, it follows that the first partial derivatives of $\log \mathcal{C}(A)$ with respect to components of A are given by the first moments of R. Moreover, using a result of Kume and Wood (2007), these derivatives can be expressed in terms of Bingham normalising constants of higher dimension. Before moving on, we briefly discuss how to calculate these normalizing constants numerically.

Two useful options for numerical calculation of the Bingham normalizing constant in general dimensions are the saddlepoint approximations of Kume and Wood (2005) and the Holonomic Gradient method of Sei and Kume (2015). Here, we focus on the former as it is faster and easier to implement, though it is typically less accurate.

We now present a result which expresses the first moments of components of R in terms of Bingham normalizing constants.

Proposition A.1. *Suppose that the* 3×3 *matrix* $A = \mu^\top \Sigma^{-1} U \Delta$ *in (17) has singular value decomposition*

$$A = T_1 \Phi T_2^\top, \tag{A6}$$

where $\Phi = \text{diag}\{\phi_1, \phi_2, \phi_3\}$. Define $\Xi = \text{diag}\{\xi_1, \xi_2, \xi_3, \xi_4\}$ *where*

$$\xi_4 = \phi_1 + \phi_2 + \phi_3 \quad \text{and} \quad \xi_i = 2\phi_i - \xi_4, \;\; i = 1, 2, 3. \tag{A7}$$

Then

$$\bar{R} = E[R|A] = \int_{R \in SO(3)} R f_2(R|A)(dR) = T_1 \Omega T_2^\top, \tag{A8}$$

where f_2 *is the conditional density defined in (17),* $\Omega = diag\{\omega_1, \omega_2, \omega_3\}$ *and, in terms of (A4) and (A5),*

$$\omega_j = 1 - \frac{\mathcal{C}_6(\Xi_k) + \mathcal{C}_6(\Xi_\ell)}{\pi \mathcal{C}_4(\Xi)}, \quad j, k, \ell \in \{1, 2, 3\}, \quad j \neq k \neq \ell \neq j. \tag{A9}$$

It may appear at first glance that the right hand side of (A8) could depend on the particular version of the singular value decomposition used, i.e. on whether or not we

insist that T_1 and T_2 are both in $SO(3)$. Reassuringly, it turns out that this is not the case: the result is invariant with respect to whether we use (2) or (3).

Appendix B: A parametric bootstrap test of goodness-of-fit

Let $y_i \in \mathcal{M}$ denote a response vector associated with sample unit i, $1 \leq i \leq n$, where $\mathcal{M}$ denotes a manifold embedded in $\mathbb{R}^d$, i.e. $\mathcal{M} \subseteq \mathbb{R}^d$. Let $x_i \in \mathbb{R}^p$ denotes a covariate vector associated with y_i. Consider the parametric model with joint density g where

$$g(y_1, \ldots, y_n | x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(y_i | x_i, \theta), \tag{B1}$$

where $\theta \in \Theta \subseteq \mathbb{R}^q$ is a parameter vector. Define

$$\ell_i(\theta) = \log f(y_i | x_i, \theta) \quad \text{and} \quad \bar{\ell}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell_i(\theta), \tag{B2}$$

and define the maximum likelihood estimator

$$\hat{\theta} = \text{argmin}_{\theta \in \Theta} \, \bar{\ell}(\theta). \tag{B3}$$

The proposed test statistic is

$$T = \bar{\ell}(\hat{\theta}). \tag{B4}$$

Under mild regularity conditions, stated below, the following result holds: if the parametric model is correct and $\hat{\theta}$ is a consistent estimator of θ_0, the statistic T is asymptotically normally distributed after suitable centering and rescaling. A brief sketch of the proof of this result is given at the end of the section.

Rather than use the asymptotic distribution directly, which entails fundamentally straightforward but cumbersome calculations to estimate the asymptotic variance of T, we prefer to obtain a p-value using a parametric bootstrap, where the parametric bootstrap samples

are generated from the fitted model with joint density $g(y_1, \ldots, y_n | x_1, \ldots, x_n; \hat{\theta})$. Moreover, bootstrap theory indicates that there are potential practical benefits in using the bootstrap rather than the asymptotic distribution directly; see Hall (1992). Regarding notation, we write the responses for bootstrap sample b as $y_1^{(b)}, \ldots, y_n^{(b)}$, $1 \leq b \leq B$, where $y_i^{(b)} \sim f(y_i | x_i, \hat{\theta})$, and define

$$\bar{\ell}^{(b)}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log f(y_i^{(b)} | x_i, \theta), \tag{B5}$$

with the associated maximum likelihood estimator

$$\hat{\theta}^{(b)} = \text{argmin}_{\theta \in \Theta} \, \bar{\ell}^{(b)}(\theta), \tag{B6}$$

and define

$$T^{(b)} = \bar{\ell}^{(b)}(\hat{\theta}^{(b)}). \tag{B7}$$

We consider the following bootstrap goodness-of-fit test.

Algorithm: Parametric Bootstrap Test of Goodness-of-Fit

Step 1: Calculate T in (B4) using the definitions in (B1)-(B3).

Step 2: For each b, $1 \leq b \leq B$, simulate the bootstrap samples $y_i^{(b)}$, $1 \leq i \leq n$, and calculate $T^{(1)}, \ldots, T^{(B)}$ as defined in (B7), using (B5) and (B6).

Step 3: Calculate the (two-sided) bootstrap p-value by

$$p = 2 \min(p^*, 1 - p^*) \quad \text{and} \quad p^* = \frac{1}{B} \sum_{b=1}^{B} I(T^{(b)} > T).$$

Finally, we consider the asymptotic distribution of T. Assume

(i) $\hat{\theta} \xrightarrow{p} \theta_0$, where for each n, $\hat{\theta}$ is defined in (B3);

(ii) Suppose $\ell_i(\theta_0)$, $1 \leq i \leq n$, satisfies a central limit theorem in the sense that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\ell_i(\theta_0) - E_0\left\{\ell_i(\theta_0)\right\}\right] \to N(0, \sigma^2),$$

where $\sigma^2 > 0$ is the limiting variance, assumed to exist; and

(iii) For some open neighbourhood $\mathcal{N} \subset \mathcal{M}$ of $\theta_0 \in \mathcal{M}$ (open in the topology of $\mathcal{M}$ rather than in the topology of the ambient space), and some fixed family of matrices $\{K(\theta) : \theta \in \mathcal{N}\}$ which are positive definite and continuous over $\theta \in \mathcal{N}$, we have, for any given $\epsilon > 0$

$$\lim_{n\to\infty} P_0\left[\sup_{\theta\in\mathcal{N}} \left|\left|\nabla\nabla^\top \bar{\ell}(\theta) - K(\theta)\right|\right| > \epsilon\right] = 0,$$

where $||.||$ is the Frobenious matrix norm, and P_0 denotes probability calculated the population distribution corresponding to $\theta = \theta_0$.

Conditions (i) and (iii) can be checked in particular cases using results in van der Vaart (2000), for example, whereas (ii) involves checking classical conditions (see e.g. Chung, 2001).

We now sketch a proof of the asymptotic normality of T under conditions (i)-(iii) above. Condition (iii) permits a second-order Taylor expansion

$$\begin{aligned} T \equiv \bar{\ell}(\hat{\theta}) &= \bar{\ell}(\theta_0) + (\hat{\theta} - \theta_0)^\top \nabla\bar{\ell}(\theta_0) + \frac{1}{2}(\hat{\theta} - \theta_0)^\top \nabla\nabla^\top \bar{\ell}(\theta_0)(\hat{\theta} - \theta_0) + o_p(n^{-1}) \\ &= E_0[\bar{\ell}(\theta_0)] + \left[\bar{\ell}(\theta_0) - E\left\{\bar{\ell}(\theta_0)\right\}\right] - \frac{1}{2}(\hat{\theta} - \theta_0)^\top \nabla\nabla^\top \bar{\ell}(\theta_0)(\hat{\theta} - \theta_0) + o_p(n^{-1}). \end{aligned} \tag{B8}$$

Note that the first term on the RHS of (B8) is a constant, the second term will be $O_p(n^{-1/2})$ but after multiplication by $n^{1/2}$ it will be asymptotically Gaussian $N(0, \sigma^2)$, whereas the third term is $O_p(n^{-1})$, but after multiplication by n it is asymptotically χ^2_q where q is the dimension of θ.

With some further calculations along the lines of Hall (1992) and Hall and Wilson (1991), it may be shown that as $n \to \infty$, and assuming $B \to \infty$, the bootstrap p-value is asymptotically uniform on $[0, 1]$ with distributional error of size $O_p(n^{-1})$, if the parametric model

is correct. Note that the error is $O_p(n^{-1})$ rather than $O_p(n^{-2})$ due to the fact that T is not a pivotal statistic, even asymptotically.

References

[1] Abramowitz, M. & Stegun, I.A. (1972). *Handbook of Mathematical Functions*. 9th ed. Dover, New York.

[2] Alshabani, A. K. S. & Dryden, I. L. & Litton, C. D. & Richardson, J. (2007). Bayesian analysis of human movement curves, *J. Roy. Statist. Soc. Ser. C* , **4**, 415–428.

[3] Bookstein, F.L. (1991). *Morphometric Tools for Landmark Data: Geometry and Biology*. Cambridge University Press, Cambridge.

[4] Chung, K.L. (2001). *A Course in Probability Theory*. 3rd Edition. Academic Press, San Diego.

[5] Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. of the Royal Statistical Society, Series B*, **34**, 1-38.

[6] Diaz-Garcia, J.A. & Jaimez, G. R & Mardia, K. V. (1997). Wishart and Pseudo-Wishart Distributions and Some Applications in Shape Theory. *J. Mult. Anal.* **63**, 73-87.

[7] Dryden, I.L. & Mardia, K.V. (2016). *Statistical Shape Analysis, with Applications in R, 2nd edition*. John Wiley, Chichester.

[8] Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer, New York.

[9] Hall, P. and Wilson, S.R. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics*, **47**, 757-762.

[10] Kendall, D.G., Barden, D., Carne, T.K. & Le, H. (1999). *Shape and Shape Theory.* John Wiley, New York.

[11] Kenobi, K., Dryden, I. L., & Le, H. (2010). Shape curves and geodesic modelling. *Biometrika*, **97**(3): 567584.

[12] Kent J.T. & Mardia K.V. (1997). Consistency of Procrustes estimators *J. R. Statist. Soc. Ser. B*, **59**, 281-290.

[13] Kent J.T. & Mardia K.V. (2001) Shape, Procrustes tangent projections and bilateral symmetry. *Biometrika* **88**, 469-485.

[14] Kent, J. T., Mardia, K. V., Morris, R. J., & Aykroyd, R. G. (2001). Functional models of growth for landmark data. *In Proceedings in Functional and Spatial Data Analysis*, 109115. Leeds University Press.

[15] Kume, A. & Dryden, I. L. & Le, H. (2007). Shape-space smoothing splines for planar landmark data. *Biometrika* **94**, 513-528.

[16] Kume, A. & Welling, M. (2010) Maximum Likelihood Estimation for the Offset-Normal Shape Distributions Using EM, *J. Comp. and Graph. Stat.*, **19**, No. 3: 702 -723.

[17] Kume, A. & Wood A.T.A. (2005) Saddlepoint approximations for the Bingham and Fisher-Bingham normalising constants. *Biometrika* **92**, 465-476.

[18] Kume, A. & Wood, A.T.A. (2007). On the derivatives of the normalising constant of the Bingham distribution. *Statistics & Probability Letters*, **77**, 832-837.

[19] Le, H. (1998). On the consistency of Procrustean mean shapes *Adv. in Appl. Probab.* **30**, 5363.

[20] Le, H. & Kume, A. (2000a). The Fréchet mean and the shape of the means. *Adv. in Appl. Probab.* **32**, 101-114.

[21] Le, H. & Kume, A. (2000b). Detection of shape changes in biological features. *J. Microscopy* **200**, 140-147.

[22] Magnus, J.R. & Neudecker, H. (1988). *Matrix differential calculus with applications in statistics and econometrics*. John Wiley & Sons, New York.

[23] Mardia, K.V. & Jupp, P.E. (2000). *Directional Statistics.* John Wiley, New York.

[24] McLachlan, G.J. & Krishnan, T. (1997). *The EM Algorithm and Extensions*. John Wiley, New York.

[25] Muirhead, R.J. (1982) *Aspects of Multivariate Statistical Theory*, John Wiley & Sons, New York.

[26] Prentice, M. J. (1986). Orientation statistics without parametric assumptions. *J. R. Statist. Soc. Ser. B*, **48**, 214-222.

[27] Sei, T & Kume, A. (2015). Calculating the normalising constant of the Bingham distribution on the sphere using the holonomic gradient method. *Statistics and Computing*, **25**, 321-332.

[28] Sundberg, R. (1974). Maximum Likelihood Theory for Incomplete Data from an Exponential Family. *Scand. J. Stat* **1**, 49-58.

[29] van der Vaart, A. W. (2000). *Asymptotic Statistics.* Cambridge University Press, Cambridge.

[30] Wood, A.T. A. (1993). Estimation of the concentration parameters of the Fisher Matrix distribution on SO(3) and the Bingham distribution on S_q, $q \geq 2$. *Australian Jou. Stat*, **31**, 69-79.

[31] Wu, C.F.J. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.* **11**, 96-103.