Maximum-likelihood estimation for the offset normal shape distributions using EM

Alfred Kume

Institute of Mathematics, Statistics and Actuarial Science, University of Kent Canterbury, CT2 7NF,UK

a.kume@kent.ac.uk

Max Welling

School of Information and Computer Science, University of California Irvine, Irvine CA 92697-3425, USA

welling@ics.uci.edu

July 2, 2010

Abstract

The offset-normal shape distribution is defined as the induced shape distribution of a Gaussian distributed random configuration in the plane. Such distributions were introduced in Dryden and Mardia (1991) and represent an important parameterized family of shape distributions for shape analysis. This paper reports a method for performing maximum likelihood estimation of parameters involved. The method consists of an EM algorithm with simple update rules and is shown to be easily applicable in many practical examples. We also show the necessary adjustments needed for using this algorithm for shape regression, missing landmark data and mixtures of offset-normal shape distributions.

Keywords: EM algorithm, shape analysis, offset-normal shape distributions, mean shape.

1 Introduction

Statistical shape analysis has important applications in biology, anatomy, genetics, medicine, archeology, geology, geography, agriculture, image analysis, computer vision, pattern recognition and chemistry (see e.g. 1.2 of Dryden and Mardia 1998). In many situations the object of study is 2-D so its shape features can be explained by the position of a finite collection of points in the plane. These points are called *landmarks*. Assume that the number of not-all-coincident landmarks under study is k with coordinates given by a $k \times 2$ matrix

$$\mathbf{X}^\dagger = \begin{pmatrix} x_1^\dagger & x_2^\dagger & \cdots & x_k^\dagger \\ y_1^\dagger & y_2^\dagger & \cdots & y_k^\dagger \end{pmatrix}^T .$$

In statistical shape analysis, it is of interest to study an iid sample of such planar configurations: $\mathbf{X}_1^\dagger, \cdots, \mathbf{X}_n^\dagger$ generated by some distribution $F(\mathbf{X}^\dagger)$ and observed after each one of those is randomly re-scaled, rotated and translated (c.f. 5.3 of Dryden and Mardia 1998). In other words, our observed data consists of elements

$$s_i(\mathbf{X}_i^\dagger + \mathbf{1}_k \otimes t_i^T)\mathbf{R}_i$$

where $s_i > 0$ is a re-scaling factor, $\mathbf{R}_i$ is an element from $\mathcal{SO}(2)$, the group of rotations in the plane and $\mathbf{1}_k \otimes t_i^T$ with $\mathbf{1}_k$, a k-vector of ones and $\otimes$ the Kronecker product, represents the translation effect by a vector t_i in the plane. Considering s_i, t_i and $\mathbf{R}_i$ as nuisance parameters, the statistical inference based on the underlying distribution $F(\mathbf{X}^\dagger)$ needs to be invariant to location, rotation and scaling for each observed element $s_i(\mathbf{X}_i^\dagger + \mathbf{1}_k \otimes t_i)\mathbf{R}_i$. This is essentially an inference problem based on the shapes of planar configurations $\mathbf{X}_i^\dagger$.

In this paper we will focus on situations where F is Gaussian, namely, $\mathrm{vec}(\mathbf{X}^\dagger)$ has a $2k$ dimensional normal distribution $\mathcal{N}_{2k}(\mathrm{vec}(\mu^\dagger), \Sigma^\dagger)$ where $vec(\mathbf{X}^\dagger)$ is the vector of length $2k$ obtained by concatenating the two columns of $\mathbf{X}^\dagger$. The induced shape distribution, which is the main concern of our paper, is called the Mardia-Dryden offset-normal distribution and is found in Dryden and Mardia (1991). Such distributions given later in equation (4), represent an important family in statistical shape analysis and a considerable amount of research has been done with regard to estimating their parameters. These distributions have appealing practical properties since practitioners want to build models based on the assumptions in configuration space and interpret their estimated quantities, like mean and correlation, in terms of landmarks.

In particular, in Le (1998) and Kent and Mardia (1997) it is shown that if $\Sigma^\dagger$ is a multiple of identity, the *Procrustes mean shape* calculated using the general Procrustes algorithm is a consistent estimator of the shape of $\mu^\dagger$. As a result, the inference carried out via the Procrustes tangent coordinates is based on the induced distribution on the tangent space of the Procrustes

mean (c.f. 7.2 of Dryden and Mardia 1998). In Kent and Mardia (2001) it is shown that if the entries of $\Sigma^{\dagger}$ are small compared to the size of $\mu^{\dagger}$, the inference based on the tangent space of the Procrustes mean is appropriate since the shape variables projected into that space follow approximately normal distributions. However, in the general covariance case and in relatively dispersed shape data this approximation may not be reasonable.

In this paper, we will work directly with offset-normal shape distributions and develop a new method for exact maximum likelihood estimation of parameters involved without making any approximation. Since the distribution is known for these cases, the likelihood function is given in closed form. However, due to its complicated form, direct numerical likelihood optimization based on standard numerical routines is generally difficult and could be unstable, especially when working with full covariance structures and large number of landmarks. For this reason, attempts based on maximum likelihood approach have only been reported for very simple covariance structures and low number of landmarks (c.f. 6.7.4 of Dryden and Mardia 1998). In our method however, dimensionality poses less of a problem since the algorithm runs efficiently based on the Expectation Maximization (EM) algorithm and the update steps take a rather simple form making the implementation straightforward. This enables us to construct maximum likelihood ratio tests for a wide range of inference problems in shape analysis such as two sample problems or shape regression based on Gaussian distributed configurations. The method proposed can also cope with missing data, a feature not immediately available for the Procrustes shape space approach.

The paper is organized as follows. In Section 2 we give an introduction to shape variables and offset-normal shape distributions as well as parameters needed for identifying them. In Section 3 we introduce the EM algorithm for general covariance matrices by establishing the general update rules. Section 4 describes the necessary adjustments of the algorithm for some covariance structures which have applications in statistical shape analysis. Implementation issues related to re-labeling invariance of landmarks and missing data are addressed in Section 5. In Section 6 we consider the extensions of the EM to an estimation approach for shape regression and mixtures of offset shape distributions. In Section 7 we apply the algorithm to real data and conclude the paper with some general remarks about the proposed method and possible future applications of EM algorithm in likelihood based inference for shape analysis.

2 Background and notation

In this section we obtain the shape distribution which is the object of our mle approach. We achieve this by using Bookstein shape variables as in Dryden and Mardia (1991) and discuss the number of parameters involved for estimation.

2.1 Shape variables and the offset-normal shape distribution

For a particular configuration $\mathbf{X}^\dagger$, we can identify its shape as follows. First remove the information about translation by left multiplying $\mathbf{X}^\dagger$ with the $(k-1)\times k$ matrix $\mathbf{L}$ constructed as $(-\mathbf{1}_{k-1}, \mathbf{I}_{k-1})$ where $\mathbf{I}_{k-1}$ is the identity matrix of dimension $(k-1)\times(k-1)$. In fact, the matrix transformation $\mathbf{X}^\dagger \to \mathbf{X} = \mathbf{L}\mathbf{X}^\dagger$ generates the coordinates of the remaining $k-1$ vertices after translating the original configuration $\mathbf{X}^\dagger$ such that its first landmark is mapped to the origin $(0,0)$. Clearly this is a linear projection from $\mathbb{R}^{2k}$ to $\mathbb{R}^{2(k-1)}$. We call $\mathbf{X} = \mathbf{L}\mathbf{X}^\dagger$ the *preform* of configuration $\mathbf{X}^\dagger$ and if we write it as

$$\mathbf{X} = \begin{pmatrix} x_2 & x_3 & \cdots & x_k \\ y_2 & y_3 & \cdots & y_k \end{pmatrix}^T$$

the rotation and scale information can be removed via

$$\mathbf{X} \to \mathbf{X} \begin{pmatrix} x_2 & -y_2 \\ y_2 & x_2 \end{pmatrix} \frac{1}{x_2^2 + y_2^2} = \begin{pmatrix} 1 & u_3 & \cdots & u_k \\ 0 & v_3 & \cdots & v_k \end{pmatrix}^T \tag{1}$$

provided that $x_2^2+y_2^2 > 0$. The shape coordinates of configuration $\mathbf{X}^\dagger$ are $\mathbf{u} = (u_3, \cdots, u_k, v_3, \cdots, v_k)^T$ and are called the Bookstein's shape variables. They are obtained by the coordinates of the remaining landmarks after configuration $\mathbf{X}^\dagger$ is translated, re-scaled and rotated such that its first two landmarks coincide with points $(0,0)$ and $(1,0)$ respectively.

The basic assumption for obtaining $\mathbf{u}$ in this way is $x_2^2 + y_2^2 > 0$, namely, the first two landmarks of $\mathbf{X}^\dagger$ are not coincident. Otherwise, we can choose some other pair of landmarks to define an alternative base line for obtaining shape coordinates. Such a pair exists since the landmarks of $\mathbf{X}^\dagger$ are not-all-coincident.

Let us assume now that $\mathrm{vec}(\mathbf{X}^\dagger)$ is distributed as $\mathcal{N}_{2k}(\mathrm{vec}(\mu^\dagger), \Sigma^\dagger)$ and we want to find the distribution of shape variables $\mathbf{u}$. This is achieved by integrating out $\mathbf{h} = (x_2, y_2)^T$ which represents the rotation and scaling information for the preform $\mathbf{X}$. Apart from some zero measurable set, transformation (1) is valid and if we take

$$\mathbf{W} = \begin{pmatrix} 1 & u_3 & \cdots & u_k & 0 & v_3 & \cdots & v_k \\ 0 & -v_3 & \cdots & -v_k & 1 & u_3 & \cdots & u_k \end{pmatrix}^T$$

then $\mathrm{vec}(\mathbf{X}) = \mathbf{W}\mathbf{h}$. Since $\mathbf{X} = \mathbf{L}\mathbf{X}^\dagger$ then $\mathrm{vec}(\mathbf{X}) \sim \mathcal{N}_{2k-2}(\mathrm{vec}(\mu), \Sigma)$ where $\mu = \mathbf{L}\mu^\dagger$ and $\Sigma = (\mathbf{I}_2 \otimes \mathbf{L})\Sigma^\dagger(\mathbf{I}_2 \otimes \mathbf{L}^T)$. Hence, the joint pdf of $(\mathbf{h}^T, \mathbf{u}^T)$ with respect to Lebesgue measure is

$$f(\mathbf{h}, \mathbf{u}; \mu, \Sigma) = \frac{1}{(2\pi)^{k-1}|\Sigma|^{1/2}} \exp\{-\frac{G}{2}\} |J(\mathbf{X} \to (\mathbf{h}, \mathbf{u}))|, \tag{2}$$

where $G = (\mathbf{W}\mathbf{h} - \mathrm{vec}(\mu))^T\Sigma^{-1}(\mathbf{W}\mathbf{h} - \mathrm{vec}(\mu))$ and $|J(\mathbf{X} \to (\mathbf{h}, \mathbf{u}))| = \|\mathbf{h}\|^{2(k-2)}$ is the Jacobian

of the transformation $\mathbf{X} \to (\mathbf{h}, \mathbf{u})$ with $\mathbf{u}$ obtained as in (1). Rewriting G as

$$G = (\mathbf{h} - \nu)^T \Gamma^{-1} (\mathbf{h} - \nu) + g$$

with $$\Gamma^{-1} = \mathbf{W}^T \Sigma^{-1} \mathbf{W}, \quad \nu = \Gamma \mathbf{W}^T \Sigma^{-1} \text{vec}(\mu), \quad g = \mu^T \Sigma^{-1} \text{vec}(\mu) - \nu^T \Gamma^{-1} \nu,$$

we can simplify further (2) by transforming with respect to the eigenbasis of Γ,

$$\Gamma = \Psi \mathbf{D} \Psi^T, \quad \zeta = \Psi^T \nu, \quad \boldsymbol{\ell} = \Psi^T \mathbf{h}$$

where $\mathbf{D} = diag(\sigma_x^2, \sigma_y^2)$. Since the determinant of the Jacobian does not change under orthogonal transformations, the pdf of $(\boldsymbol{\ell}, \mathbf{u})$ is

$$f(\boldsymbol{\ell}, \mathbf{u}; \mu, \Sigma) = \frac{|\Gamma|^{1/2} \exp(g/2)}{(2\pi)^{k-2} |\Sigma|^{1/2}} f_{\mathcal{N}}(\ell_x; \zeta_x, \sigma_x) f_{\mathcal{N}}(\ell_y; \zeta_y, \sigma_y) (\ell_x^2 + \ell_y^2)^{k-2} \tag{3}$$

where $f_{\mathcal{N}}(x; \mu, \sigma)$ denotes the pdf at x of the Gaussian distribution with parameters μ and σ. Using the binomial expansion

$$(\ell_x^2 + \ell_y^2)^{k-2} = \sum_{i=0}^{k-2} \binom{k-2}{i} \ell_x^{2i} \ell_y^{2k-4-2i}$$

we can integrate out $\boldsymbol{\ell} = \Psi^T \mathbf{h}$ (scale and rotation) to obtain the marginal (offset-normal shape) pdf of $\mathbf{U}$

$$f_{\mathbf{U}}(\mathbf{u}; \mu, \Sigma) = \int f(\boldsymbol{\ell}, \mathbf{u}; \mu, \Sigma) d\boldsymbol{\ell} = \frac{|\Gamma|^{1/2} \exp(g/2)}{(2\pi)^{k-2} |\Sigma|^{1/2}} \sum_{i=0}^{k-2} \binom{k-2}{i} \mathbf{E}(\ell_x^{2i} | \zeta_x, \sigma_x) \mathbf{E}(\ell_y^{2k-4-2i} | \zeta_y, \sigma_y) \tag{4}$$

where $\mathbf{E}(l^p | \mu, \sigma)$ denotes the moments of the univariate Gaussian distribution with parameters (μ, σ). These are calculated as (see 3.462/4 and 8.972 in Gradshteyn and Ryzhik 1980).

$$\mathbf{E}(\ell^p | \mu, \sigma) = \left(\frac{\sigma}{\sqrt{-2}} \right)^p H_p \left(\frac{\sqrt{-1}\mu}{\sqrt{2}\sigma} \right) = \begin{cases} (2\sigma^2)^q q! \mathcal{L}_q^{(-1/2)}(\frac{-\mu^2}{2\sigma^2}) & \text{if} \quad p = 2q \\ \mu (2\sigma^2)^q q! \mathcal{L}_q^{(1/2)}(\frac{-\mu^2}{2\sigma^2}) & \text{if} \quad p = 2q+1 \end{cases} \tag{5}$$

where H_p is the Hermite polynomial of order p and

$$\mathcal{L}_q^{(\alpha)}(x) = \sum_{i=1}^{q} \frac{(1+\alpha)_q (-x)^i}{(1+\alpha)_i i! (q-i)!}$$

with $(1+\alpha)_i = (\alpha+1)...(\alpha+i)$, the generalized Laguerre polynomial of order q.

Note that the expression for $f_{\mathbf{U}}$ in (4) is not as complicated as it might first appear since $f_{\mathbf{U}}$ involves only even moments of (5), i.e. $p = 2q$. If $\sigma_x = \sigma_y$, the summation of expectations in (4) simplifies to a simple expression of $\mathcal{L}_{k-1}^1$. This corresponds to the complex covariance case seen later in section 4.1.

2.2 Parameter space

Let us assume that we are given shape observations $\mathbf{u}_1, \cdots, \mathbf{u}_n$ such that they correspond to some unobserved sample $\mathbf{X}_1^\dagger, \cdots \mathbf{X}_n^\dagger$ from $\mathcal{N}_{2k}(\mathrm{vec}(\mu^\dagger), \Sigma^\dagger)$ and we want to estimate parameters $(\mu^\dagger, \Sigma^\dagger)$. Notice that we have in general $2k$ and $k(2k+1)$ parameters for $\mu^\dagger$ and $\Sigma^\dagger$ respectively. Since $\mathrm{vec}(\mathbf{X}) \sim \mathcal{N}_{2k-2}(\mathrm{vec}(\mu), \Sigma)$ where $\mu = \mathbf{L}\mu^\dagger$ and $\Sigma = (\mathbf{I}_2 \otimes \mathbf{L})\Sigma^\dagger(\mathbf{I}_2 \otimes \mathbf{L})^T$, in the preform space, at most $2(k-1) + (2k-1)(k-1)$ parameters could be identified. Due to the shape invariance with respect to scaling and rotating of preforms $\mathbf{X}$, we can only estimate in terms of $\mathbf{u}_i$ those parameters which identify the equivalent class

$$\Theta = \{(s\mu\mathbf{R}, s^2(\mathbf{R}^T \otimes \mathbf{I}_{k-1})\Sigma(\mathbf{R} \otimes \mathbf{I}_{k-1}))|s \in \mathbb{R}^+, \mathbf{R} \in \mathcal{SO}(2)\}. \tag{6}$$

Without loss of generality we can assume that the mean μ is re-scaled and rotated as in transformation (1) such that its first column is $(1, 0)$. So there are at most $2(k-2)$ parameters for the mean and $(2k-1)(k-1)$ for Σ identifying Θ in (6). In fact, Dryden and Mardia (1998) expect that only $(k-2)(2k-3)$ parameters are practically identifiable for Σ (see page 138 there). Therefore, the parameter space has probably a total dimension $2(k-2)+(k-2)(2k-3)$. However, while the parameters for the shape of μ are fully identifiable, in one of the examples considered, we treat the estimation of general covariance as if it has $(2k-1)(k-1)$ identifiable parameters.

Certain conditions on the structure of Σ avoid this identification problem. If for example Σ is that of some complex normal distribution (described in section 4) then it has $(k-1)(k-2)$ parameters and so its entries are fully identifiable up to some re-scaling constant s.

Note that shape coordinates are obtained via a mapping of configurations $\mathbf{X}^\dagger$ to some lower dimensional space of variables $\mathbf{u}$. Therefore there exists a large class of singular and non singular Gaussian distributions in configuration space which induce the same offset-normal shape distribution. However, our estimation method is in fact dealing with only those parameters which identify equivalent classes (6) and not all those identifying $\mu^\dagger$ and $\Sigma^\dagger$ in configuration space.

Alternatively, we could have chosen to filter the translation by simply replacing the matrix $\mathbf{L}$ with some other matrix $\mathbf{K}$ of the same dimension such that its j-th row is given by

$$(-d_j, -d_j, ..., -d_j, jd_j, 0, ..., 0)$$

where $d_j = \{j(j+1)\}^{-\frac{1}{2}}$ is repeated j times. With row vectors orthogonal, $\mathbf{K}$ is in fact the submatrix of the Helmert $k \times k$ matrix. The coordinates of the resulting preform $\mathbf{X}_H = \mathbf{K}\mathbf{X}^\dagger$ are called in 4.1.2 of Dryden and Mardia (1998) *Helmertized landmarks*. If $\mathbf{X}_H$ is then transformed as in (1) the resulting shape variables are called Kendall shape variables. The main algorithms that we describe here are given in terms of preforms $\mathbf{X}$ and Bookstein's shape variables. However, they can be derived in the same way in terms of the Helmertized preforms $\mathbf{X}_H$ and Kendall shape variables.

The only difference is that the covariance matrices in preform space need to appropriately reflect the linear transformation for producing the preforms.

3 EM algorithm for general covariance

The Expectation-Maximization (EM) algorithm is a maximum likelihood parameter estimation method and was originally developed by Dempster et al. (1977) for the cases where part of the data can be considered to be incomplete or "hidden". This paper represents the first attempt to apply this method for shape analysis. We will describe this algorithm in terms of elements in preform space with the hidden/missing data part being the rotation and re-scaling information. Our target is to find the values of μ, Σ identifying equivalent classes in (6) which maximize the log-likelihood function

$$L(\mu, \Sigma) = \sum_{i=1}^{n} \log f_{\mathbf{U}}(\mathbf{u}_i; \mu, \Sigma)$$

where $f_{\mathbf{U}}(\mathbf{u}; \mu, \Sigma)$ is the induced pdf of shape variables $\mathbf{u}$ if $\mathbf{X} \sim \mathcal{N}_{2k-2}(\mu, \Sigma)$ and $\mathbf{u}_i$ are the observed shape data.

The EM algorithm suggests an iterative optimization method such that the current estimate values μ_r, Σ_r are updated with μ_{r+1}, Σ_{r+1} such that $L(\mu_{r+1}, \Sigma_{r+1}) \geq L(\mu_r, \Sigma_r)$ with equality only at some stationary point. In particular, for given μ_r, Σ_r, the values μ_{r+1}, Σ_{r+1} are chosen to maximize the following function with respect to μ and Σ

$$\mathcal{Q}_{\mu_r, \Sigma_r}(\mu, \Sigma) = \sum_{i=1}^{n} \int \log(f_{\mathcal{N}}(\mathbf{X}_i; \mu, \Sigma)) dF(\mathbf{X}_i | \mathbf{u}_i, \mu_r, \Sigma_r)$$

where $f_{\mathcal{N}}(\cdot\ ; \mu, \Sigma)$ is the pdf of Gaussian distribution with mean μ and covariance Σ, and $F(\mathbf{X}_i | \mathbf{u}_i, \mu_r, \Sigma_r)$ is the conditional distribution of $\mathbf{X}_i$ given its shape $\mathbf{u}_i$. The updated values can be calculated once we know how to maximize $\mathcal{Q}_{\mu_r, \Sigma_r}$. This is the M (maximization) step of the EM algorithm. Since the Gaussian distribution is of exponential family form, the algorithm is simplified with the E (expectation) step given in terms of the expectations of the sufficient statistics given the observed data at a current parameter estimate. These are explained in further detail below.

M-step

By interchanging the order of differentiation with expectation and then following the same differentiation rules as in maximum likelihood estimation of multivariate normal distributions (see chapter 15, section 3 Magnus and Neudecker 1988), we see that taking the differential of $\mathcal{Q}_{\mu_r, \Sigma_r}$ with respect to μ and Σ we have

$$d\mathcal{Q}_{\mu_r, \Sigma_r}(\mu, \Sigma) = \frac{1}{2} \mathrm{tr}\left((d\Sigma)\Sigma^{-1}(S - n\Sigma)\Sigma^{-1}\right) + n(d\mathrm{vec}(\mu)'\Sigma^{-1}(M - \mu)) \tag{7}$$

where

$$M = \frac{1}{n}\sum_{i=1}^{n}\int \text{vec}(\mathbf{X}_i)dF(\mathbf{X}_i|\mathbf{u}_i,\mu_r,\Sigma_r)$$

and

$$S = \frac{1}{n}\sum_{i=1}^{n}\int \left(\text{vec}(\mathbf{X}_i)-\text{vec}(\mu)\right)\left(\text{vec}(\mathbf{X}_i)-\text{vec}(\mu)\right)^T dF(\mathbf{X}_i|\mathbf{u}_i,\mu_r,\Sigma_r).$$

Therefore the maximum of $\mathcal{Q}_{\mu_r,\Sigma_r}(\mu,\Sigma)$ is achieved at

$$\text{vec}(\mu_{r+1}) = \frac{1}{n}\sum_{i=1}^{n}\int \text{vec}(\mathbf{X}_i)dF(\mathbf{X}_i|\mathbf{u}_i,\mu_r,\Sigma_r) \tag{8}$$

and

$$\Sigma_{r+1} = \frac{1}{n}\sum_{i=1}^{n}\int \text{vec}(\mathbf{X}_i)\text{vec}(\mathbf{X}_i)^T dF(\mathbf{X}_i|\mathbf{u}_i,\mu_r,\Sigma_r) - \text{vec}(\mu_{r+1})\text{vec}(\mu_{r+1})^T. \tag{9}$$

E-step

The expectation step is performed by finding expectations (8) and (9) which establish the update rules for the parameter estimates μ_r and Σ_r. It is clear that we can calculate them once we know how to calculate entries of $\int \text{vec}(\mathbf{X})dF(\mathbf{X}|\mathbf{u},\mu_r,\Sigma_r)$ and $\int \text{vec}(\mathbf{X})\text{vec}(\mathbf{X})^T dF(\mathbf{X}|\mathbf{u},\mu_r,\Sigma_r)$. These expressions are given in the following

Lemma 1.

$$\int vec(\mathbf{X})dF(\mathbf{X}|\mathbf{u},\mu,\Sigma) = \mathbf{W}\Psi\frac{\int_{\mathbb{R}^2}\boldsymbol{\ell}f(\boldsymbol{\ell},\mathbf{u};\mu,\Sigma)d\boldsymbol{\ell}}{f_{\mathbf{U}}(\mathbf{u};\mu,\Sigma)} \tag{10}$$

and

$$\int vec(\mathbf{X})vec(\mathbf{X})^T dF(\mathbf{X}|\mathbf{u},\mu,\Sigma) = \mathbf{W}\Psi\frac{\int_{\mathbb{R}^2}\boldsymbol{\ell}\boldsymbol{\ell}^T f(\boldsymbol{\ell},\mathbf{u};\mu,\Sigma)d\boldsymbol{\ell}}{f_{\mathbf{U}}(\mathbf{u};\mu,\Sigma)}\Psi^T\mathbf{W}^T \tag{11}$$

where $\mathbf{W}$, Ψ *and* $\boldsymbol{\ell}$ *are defined as in Section 2.1 and for the pairs* $(a,b) \in \{(1,0),(0,1),(2,0),(1,1),(0,2)\}$

$$\frac{\int_{\mathbb{R}^2}\ell_x^a\ell_y^b f(\boldsymbol{\ell},\mathbf{u};\mu,\Sigma)d\boldsymbol{\ell}}{f_{\mathbf{U}}(\mathbf{u};\mu,\Sigma)} = \frac{\sum_{i=0}^{k-2}\binom{k-2}{i}\mathbf{E}(\ell_x^{2i+a}|\zeta_x,\sigma_x)\mathbf{E}(\ell_y^{2k-4-2i+b}|\zeta_y,\sigma_y)}{\sum_{i=0}^{k-2}\binom{k-2}{i}\mathbf{E}(\ell_x^{2i}|\zeta_x,\sigma_x)\mathbf{E}(\ell_y^{2k-4-2i}|\zeta_y,\sigma_y)}$$

The proof of the lemma is found online in the Appendix of the supplemental material.

Since the EM algorithm is a local optimization procedure, running it from different starting points is necessary to increase the chance of finding the global maximum. The integrating measure $dF(\mathbf{X}|\mathbf{u},\mu_r,\Sigma_r)$ is in fact the conditional distribution of pre-form $\mathbf{X}$ (derived by configuration $\mathbf{X}^\dagger$)

given its shape. This measure exists even if the first two landmarks of $\mathbf{X}^\dagger$ are coincident since we can use alternative base lines in order to generate shape coordinates $\mathbf{u}$. This will be addressed in Section 5 but until then we assume that shape variables $\mathbf{u}$ are obtained as in (1).

4 Particular cases

In many situations in shape analysis it is appropriate to reduce the number of parameters by imposing some constraints in the covariance matrix $\Sigma^\dagger$. This is also necessary if we want to carry out maximum likelihood ratio tests and avoid identification problems of parameters. In particular, we will focus on three cases:

- $\Sigma^\dagger$ is that of general complex normal distribution. This covariance type remains invariant under rotations.
- $\Sigma^\dagger$ has a cyclic correlation pattern which is used for large number of landmarks around the boundary of objects (see 6.7.1 in Dryden and Mardia 1998).
- $\Sigma^\dagger$ is simply a multiple of the identity matrix i.e. $\Sigma^\dagger = \sigma^2 \mathbf{I}_{2k}$. In this case, the distribution induced in the shape space is isotropic with center at the shape of μ and concentration depending on the ratio $\|\mu\|/\sigma^2$.

In all these cases the algorithm needs to be adjusted since the corresponding updating rules for Σ_r are shown to be different. In this section, we rely on the complex representation of the shape variables involved and show that the corresponding EM steps of (8) and (9) are easier to calculate since they take a compact form.

4.1 Complex Normal Distributions

One can easily see that shape coordinates $\mathbf{u}$ can be alternatively obtained using complex representation of planar points. If for example the preform $\mathbf{X}$ is rewritten as $\mathbf{Z} = (z_2, z_3, \cdots, z_k)^T \in \mathbb{C}^{k-1}$ such that $z_j = x_j + \sqrt{-1}y_j$ then $\xi = \mathbf{Z}/z_2 = (1, \xi_3, \cdots, \xi_k)^T$ where $\xi_j = u_j + \sqrt{-1}v_j$. Complex normal distributions are particularly important since they correspond to cases when the covariance matrix parameters are fully identifiable and they remain invariant of rotations in preform space. The complex covariance structure for the vector of coordinates $\mathrm{vec}(\mathbf{X}^\dagger)$ corresponds to the restrictions of the form

$$\Sigma^\dagger = \frac{1}{2}\begin{pmatrix} C_1^\dagger & -C_2^\dagger \\ C_2^\dagger & C_1^\dagger \end{pmatrix}$$

where $C_1^\dagger$ is positive definite and $C_2^\dagger$ is skew-symmetric matrix, i.e. $C_2^{\dagger T} = -C_2^\dagger$. The covariance of vec($\mathbf{X}$) has similar form

$$\Sigma = \frac{1}{2}\begin{pmatrix} C_1 & -C_2 \\ C_2 & C_1 \end{pmatrix} \tag{12}$$

where $C_1 = \mathbf{L}C_1^\dagger\mathbf{L}^T$ and $C_2 = \mathbf{L}C_2^\dagger\mathbf{L}^T$. If $C = C_1 + \sqrt{-1}C_2$ and we denote by $\mathbf{Z}$ and η the $k-1$ dimensional complex vector representation of $\mathbf{X}$ and μ respectively, the corresponding pdf of $\mathbf{Z}$ is

$$f_{\mathcal{N}}(\mathbf{Z};\eta,C) = \frac{1}{\pi^{k-2}|C|}\exp\{-(\mathbf{Z}-\eta)^* C^{-1}(\mathbf{Z}-\eta)\}$$

where $(\mathbf{Z}-\eta)^*$ represents the conjugate and transpose of $(\mathbf{Z}-\eta)$ (see Mardia et al. 1979). The Jacobian of transformation $\mathbf{Z} \to (z_2, \xi)$ is $\|z_2\|^{k-1}$ and based on the complex calculus one can show that the updated values η_{r+1}, C_{r+1} obtained by optimizing the corresponding $\mathcal{Q}$ function are

$$\eta_{r+1} = \frac{1}{n}\sum_{i=1}^{n}\int \mathbf{Z}_i dF(\mathbf{Z}_i|\xi_i,\eta_r,C_r) = \frac{1}{n}\sum_{i=1}^{n}\frac{\xi_i\int_{\mathbb{C}} z f_{\mathcal{N}}(z\xi_i;\eta_r,C_r)\|z\|^{2(k-2)}dz}{\int_{\mathbb{C}} f_{\mathcal{N}}(z\xi_i;\eta_r,C_r)\|z\|^{2(k-2)}dz} \tag{13}$$

and

$$\begin{aligned} C_{r+1} &= \frac{1}{n}\sum_{i=1}^{n}\int \mathbf{Z}_i\mathbf{Z}_i^* dF(\mathbf{Z}_i|\xi_i,\eta_r,C_r) - \eta_{r+1}\eta_{r+1}^* \\ &= \frac{1}{n}\sum_{i=1}^{n}\xi_i\xi_i^*\frac{\int_{\mathbb{C}}\|z\|^2 f_{\mathcal{N}}(z\xi_i;\eta_r,C_r)\|z\|^{2(k-2)}dz}{\int_{\mathbb{C}} f_{\mathcal{N}}(z\xi_i;\eta_r,C_r)\|z\|^{2(k-2)}dz} - \eta_{r+1}\eta_{r+1}^* \end{aligned} \tag{14}$$

with ratios calculated as in the following Lemma.

Lemma 2.

$$\frac{\int_{\mathbb{C}} z f_{\mathcal{N}}(z\xi;\eta,C)\|z\|^{2(k-2)}dz}{\int_{\mathbb{C}} f_{\mathcal{N}}(z\xi;\eta,C)\|z\|^{2(k-2)}dz} = \omega\frac{(k-1)}{\|b\|}\left(\frac{\mathcal{L}_{k-1}(-\|b\|^2/a)}{\mathcal{L}_{k-2}(-\|b\|^2/a)} - 1\right)$$

$$\frac{\int_{\mathbb{C}} f_{\mathcal{N}}(z\xi;\eta,C)\|z\|^{2(k-1)}dz}{\int_{\mathbb{C}} f_{\mathcal{N}}(z\xi;\eta,C)\|z\|^{2(k-2)}dz} = \frac{(k-1)}{a}\frac{\mathcal{L}_{k-1}(-\|b\|^2/a)}{\mathcal{L}_{k-2}(-\|b\|^2/a)}$$

where $a = \xi^* C^{-1}\xi$, $b = \xi^* C^{-1}\eta$ *and* $\omega = e^{\sqrt{-1}\theta}$ *such that* $\bar{\omega}\xi^* C^{-1}\eta$ *is a positive number.*

The proof of the lemma is found online in the Appendix of the supplemental material.

Remark

If the covariance matrix C is known the EM needs to perform only updates (13) where $C_r = C$. The rotation invariance of the complex covariance structures implies from (6) that the estimates for the remaining parameters η are now obtained only modulo rotations since the scale is fixed by the known values of the covariance.

4.2 Cyclic Markov Covariance

If the original covariance matrix $\Sigma^\dagger$ has the form $\Sigma^\dagger = \sigma^2 \mathbf{I}_2 \otimes \Gamma$ where

$$\Gamma(i,j) = (\gamma^{|i-j|} + \gamma^{k-|i-j|})/(1-\gamma^k) \quad 1 \le i,j \le k \quad \text{and} \quad 0 \le \gamma < 1$$

we say that $\Sigma^\dagger$ has a cyclic Markov structure and use it when the number of landmarks is large. It can be shown that for $i \ge j$

$$\Gamma^{-1}(i,j) = \begin{cases} (1+\gamma^2)/(1-\gamma^2) & \text{if} \quad 1 \le i = j \le k \\ -\gamma/(1-\gamma^2) & \text{if} \quad 2 \le j = i+1 \le k \text{ or } \ i = 1, j = k, \\ 0 & \text{otherwise.} \end{cases}$$

This is a special case of the general Complex Normal distribution corresponding to situations where the $C_2^\dagger$ component is zero and $C_1^\dagger = 2\sigma^2\Gamma$ (c.f. 6.7 in Dryden and Mardia 1998). Since the estimation is based on identifying elements from (6) then without loss of generality we can assume that $\sigma^2 = 1/2$ and as a result $C = \mathbf{L}\Gamma\mathbf{L}^T$. The optimal point for η in $\mathcal{Q}$ does not depend on the covariance structure and so the updated value η_{r+1} is calculated as in (13). Replacing η with η_{r+1} in $\mathcal{Q}$ and noting that

$$(\mathbf{Z}_i - \eta)^* C^{-1} (\mathbf{Z}_i - \eta) = \mathrm{Tr}(C^{-1}(\mathbf{Z}_i - \eta)(\mathbf{Z}_i - \eta)^*)$$

we are left to find the value γ_{r+1} which maximizes

$$\mathcal{Q}_{\eta_r, C_r}(\eta_{r+1}, C) = -n \ln |C| - \mathrm{Tr}\left(C^{-1} \left(\sum_{i=1}^{n} \int_{\mathbb{C}} \mathbf{Z}_i \mathbf{Z}_i^* dF(\mathbf{Z}_i | \xi_i, \eta_r, C_r) - n\eta_{r+1}\eta_{r+1}^* \right) \right)$$

Since the values $\int_{\mathbb{C}} \mathbf{Z}_i \mathbf{Z}_i^\dagger dF(\mathbf{Z}_i|\xi_i, \eta_r, C_r)$ are obtained as in Lemma 2, this is clearly a simple univariate optimization problem and can be carried out numerically. In fact, we do not even need to find the exact maximizing value of this function as long as we find a value γ_{r+1} such that $\mathcal{Q}_{\eta_r, C_r}(\eta_{r+1}, C_{r+1}) > \mathcal{Q}_{\eta_{r+1}, C_r}(\eta_{r+1}, C_r)$ which then implies $L(\eta_{r+1}, C_{r+1}) \ge L(\eta_r, C_r)$. This updating procedure is in the form of Generalized EM algorithm (c.f. McLachlan and Krishnan 1997). If we are to apply the algorithm in terms of the Helmertized preforms $\mathbf{X}_H$ then the algorithm runs in the same way as before. The computation time can be significantly reduced since the covariance structure in the preform space will now be $C_H = \mathbf{K}\Gamma\mathbf{K}^T$ with $C_H^{-1} = \mathbf{K}\Gamma^{-1}\mathbf{K}^T$ and $|C_H| = (1-\gamma)^{k+1}(1+\gamma)^{k-1}/(1-\gamma^k)^2$ (c.f. 6.7.1 of Dryden and Mardia 1998).

4.3 Isotropic case

This corresponds to situations when the covariance between landmark is $\Sigma^{\dagger} = \sigma^2 \mathbf{I}_{2k}$. It suffices for the EM algorithm now to calculate only η_{r+1} since as we can see from (6) without loss of generality we can fix σ to 1 and so the covariance matrix in preform space is known. Note that in such cases, the covariance in the space of Helmertised preforms $\mathbf{X}_H$ is $\sigma^2 \mathbf{I}_{2k-2}$ (isotropic) and the EM algorithm now resembles the *ordinary Procrustes algorithm* which consists of subsequent rotation matching of ξ_i with the current proposed value η_r. This is achieved by multiplications with ω_i for each observation. The algorithm here differs only by the presence of a re-scaling factor $\frac{(k-1)}{\|b\|}\left(\frac{\mathcal{L}_{k-1}(-\|b\|^2/a)}{\mathcal{L}_{k-2}(-\|b\|^2/a)} - 1\right)$ where $\|b\| = \|\mathbf{Z}\|\,\|\eta\| \cos\rho/2$ and $\|b\|^2/a = \|\eta\|^2 \cos^2\rho/2$ where ρ is the Kendall shape distance between $\mathbf{Z}$ and η. It can easily be seen that the estimated value of $\|\eta\|$ leads to estimation of the concentration parameter (see 6.6.2 in Dryden and Mardia 1998) for such shape distributions.

5 Incomplete data and base line invariance

In this section we show that the algorithm can be easily adjusted for missing data. In order to establish that we need to deal first with the base line invariance since the missing data for one individual can contain those landmarks used as base line for another.
Recall that the shape variables $\mathbf{u}$ given in (1) are calculated after the base line for configuration $\mathbf{X}^{\dagger}$ is defined by the first two landmarks. These landmarks are assumed non coincident but the algorithm can run even if this is not the case as long as some other non coincident pair exists. The choice of the baseline however does not have to be fixed for each shape observation. In the following we show how to implement the algorithm for alternative choices of baselines.

Lemma 3. *If* $\mathbf{X}$ *and* $\hat{\mathbf{X}}$ *are the corresponding preforms of configuration* $\mathbf{X}^{\dagger}$ *obtained by the baseline choices of the first two landmarks and* i *and* j *respectively, then*

$$\hat{\mathbf{X}} = \mathbf{L}\mathbf{P}_{ij}\begin{pmatrix} 0 \cdots 0 \\ \mathbf{I}_{k-1} \end{pmatrix}\mathbf{X}$$

where $\mathbf{P}_{ij}$ *is a permutation matrix which rearranges the rows of* $\mathbf{X}^{\dagger}$ *such that the first two are exchanged with those in positions* i *and* j*.*

The proof of this Lemma is straightforward since without loss of generality we assume that the first landmark of $\mathbf{X}^{\dagger}$ is at $(0,0)$ i.e. $\mathbf{X}^{\dagger} = \begin{pmatrix} 0\ 0 \\ \mathbf{X} \end{pmatrix}$. This implies $\mathbf{L}\mathbf{P}_{ij}\begin{pmatrix} 0\ 0 \\ \mathbf{X} \end{pmatrix} = \begin{pmatrix} 0\ 0 \\ \hat{\mathbf{X}} \end{pmatrix}$ and so the stated relationship follows.
The matrix $\mathbf{A}_{ij} = \mathbf{L}\mathbf{P}_{ij}\begin{pmatrix} 0 \cdots 0 \\ \mathbf{I}_{k-1} \end{pmatrix}$ is clearly square of dimension $k-1$ and using the properties

of Kronecker products (see e.g. A.3.2 in Mardia et al. 1979) we have $\text{vec}(\hat{\mathbf{X}}) = (\mathbf{I}_2 \otimes \mathbf{A}_{ij})\text{vec}(\mathbf{X})$. As a result, if $\mathbf{X} \sim \mathcal{N}_{2(k-1)}(\text{vec}(\mu), \Sigma)$ then $\hat{\mathbf{X}} \sim \mathcal{N}_{2(k-1)}(\text{vec}(\hat{\mu}), \hat{\Sigma})$ where $\hat{\mu} = \mathbf{A}_{ij}\mu$ and $\hat{\Sigma} = (\mathbf{I}_2 \otimes \mathbf{A}_{ij})\Sigma(\mathbf{I}_2 \otimes \mathbf{A}_{ij}^T)$. If in particular $\hat{\mathbf{u}}$ contains the corresponding shape coordinates of $\mathbf{X}^\dagger$ with respect to the alternative baseline then $dF(\mathbf{X}|\mathbf{u}, \mu, \Sigma) = dF(\hat{\mathbf{X}}|\hat{\mathbf{u}}, \hat{\mu}, \hat{\Sigma})$ and therefore

$$\int \text{vec}(\mathbf{X}) dF(\mathbf{X}|\mathbf{u}, \mu, \Sigma) = (\mathbf{I}_2 \otimes \mathbf{A}_{ij}^{-1}) \int \text{vec}(\hat{\mathbf{X}}) dF(\hat{\mathbf{X}}|\hat{\mathbf{u}}, \hat{\mu}, \hat{\Sigma})$$

$$\int \text{vec}(\mathbf{X})\text{vec}(\mathbf{X})^T F(\mathbf{X}|\mathbf{u}, \mu, \Sigma) = (\mathbf{I}_2 \otimes \mathbf{A}_{ij}^{-1}) \int \text{vec}(\hat{\mathbf{X}})\text{vec}(\hat{\mathbf{X}})^T dF(\hat{\mathbf{X}}|\hat{\mathbf{u}}, \hat{\mu}, \hat{\Sigma}) \, (\mathbf{I}_2 \otimes \mathbf{A}_{ij}^{-T}).$$

This implies that the choice of the baseline is not important as long as we appropriately transform the values μ and Σ to $\hat{\mu}$ and $\hat{\Sigma}$.

We return now to the missing data problem and by assuming that for some particular observation $\mathbf{X}_i$ not all the landmarks are given. Applying Lemma 3 for an appropriate permutation matrix, without loss of generality we assume that the last $p < k-2$ landmarks are unobserved and the base line is defined by the first two of those observed. Denote by $\mathbf{X}$ the preform of this particular observation $\mathbf{X}_i$ and write it as $\mathbf{X} = \left(\hat{\mathbf{X}}, \tilde{\mathbf{X}}\right)^T$, where $\hat{\mathbf{X}}$ and $\tilde{\mathbf{X}}$ correspond to the observed and unobserved set of landmarks respectively. The square matrix $\mathcal{I}$ defined as

$$\mathcal{I} = \begin{pmatrix} \mathbf{I}_{k-1-p} & 0\cdots 0 & 0\cdots 0 & 0\cdots 0 \\ 0\cdots 0 & 0\cdots 0 & \mathbf{I}_{k-1-p} & 0\cdots 0 \\ 0\cdots 0 & \mathbf{I}_p & 0\cdots 0 & 0\cdots 0 \\ 0\cdots 0 & 0\cdots 0 & 0\cdots 0 & \mathbf{I}_p \end{pmatrix}$$

has the property

$$\text{vec}(\mathbf{X}) = \mathcal{I}^T \begin{pmatrix} \text{vec}(\hat{\mathbf{X}}) \\ \text{vec}(\tilde{\mathbf{X}}) \end{pmatrix} \tag{15}$$

and

$$\text{vec}(\mathbf{X})\text{vec}(\mathbf{X})^T = \mathcal{I}^T \begin{pmatrix} \text{vec}(\hat{\mathbf{X}})\text{vec}(\hat{\mathbf{X}})^T & \text{vec}(\hat{\mathbf{X}})\text{vec}(\tilde{\mathbf{X}})^T \\ \text{vec}(\tilde{\mathbf{X}})\text{vec}(\hat{\mathbf{X}})^T & \text{vec}(\tilde{\mathbf{X}})\text{vec}(\tilde{\mathbf{X}})^T \end{pmatrix} \mathcal{I}. \tag{16}$$

Now, the only shape information observed in this case is $\hat{\mathbf{u}}$ which is obtained after applying transformation (1) to submatrix $\hat{\mathbf{X}}$. Since the unknown information is $\mathbf{h}$ and $\tilde{\mathbf{X}}$ the contribution on the EM algorithm of this particular observation is carried out by taking expectations with respect to the following distribution

$$dF(\mathbf{X}|\hat{\mathbf{u}}, \mu, \Sigma) = dF(\mathbf{X}|\hat{\mathbf{X}}, \mu, \Sigma) dF(\hat{\mathbf{X}}|\hat{\mathbf{u}}, \mu, \Sigma) \tag{17}$$

where $dF(\mathbf{X}|\hat{\mathbf{X}}, \mu, \Sigma)$ is the conditional distribution of $\mathbf{X}$ given its sub-matrix $\hat{\mathbf{X}}$. Applying standard results on conditional Gaussian distributions (see e.g. Theo. 3.2.4 in Mardia et al. 1979)

$$dF(\mathbf{X}|\hat{\mathbf{X}},\mu,\Sigma) = f_{\mathcal{N}}(\tilde{\mathbf{X}};\Omega,\Sigma_{22.1})d\tilde{\mathbf{X}}$$

where

$$\Omega = \text{vec}(\tilde{\mu}) + \Sigma_{21}\hat{\Sigma}^{-1}(\text{vec}(\hat{\mathbf{X}}) - \text{vec}(\hat{\mu})) \quad \mathcal{I}\Sigma\mathcal{I}^T = \begin{pmatrix} \hat{\Sigma} & \Sigma_{12} \\ \Sigma_{21} & \tilde{\Sigma} \end{pmatrix}$$

$\hat{\Sigma} = cov(\text{vec}(\hat{\mathbf{X}}))$, $\tilde{\Sigma} = cov(\text{vec}(\tilde{\mathbf{X}}))$, $\Sigma_{12} = \Sigma_{21}^T = \mathbf{E}(\text{vec}(\hat{\mathbf{X}} - \hat{\mu})\text{vec}(\tilde{\mathbf{X}} - \tilde{\mu})^T)$ and $\Sigma_{22.1} = \tilde{\Sigma} - \Sigma_{21}\hat{\Sigma}^{-1}\Sigma_{12}$. The next term in (17) $dF(\hat{\mathbf{X}}|\hat{\mathbf{u}},\mu,\Sigma)$ is the same as $dF(\hat{\mathbf{X}}|\hat{\mathbf{u}},\hat{\mu},\hat{\Sigma})$ which represents the conditional probability measure applied to the marginal distribution of $\hat{\mathbf{X}}$ with parameters $\hat{\mu}$ and $\hat{\Sigma}$.

Integrating out both (15) and (16) with respect to $dF(\mathbf{X}|\hat{\mathbf{X}},\mu,\Sigma)$ followed by $dF(\hat{\mathbf{X}}|\hat{\mathbf{u}},\hat{\mu},\hat{\Sigma})$ the corresponding expressions (10) and (11) are actually

$$\int \text{vec}(\mathbf{X})dF(\mathbf{X}|\hat{\mathbf{u}},\mu,\Sigma) = \mathcal{I}^T \int \begin{pmatrix} \text{vec}(\hat{\mathbf{X}}) \\ \Omega \end{pmatrix} dF(\hat{\mathbf{X}}|\hat{\mathbf{u}},\hat{\mu},\hat{\Sigma}) \tag{18}$$

and

$$\int \text{vec}(\mathbf{X})\text{vec}(\mathbf{X})^T dF(\mathbf{X}|\hat{\mathbf{u}},\mu,\Sigma) = \mathcal{I}^T \int \begin{pmatrix} \text{vec}(\hat{\mathbf{X}})\text{vec}(\hat{\mathbf{X}})^T & \text{vec}(\hat{\mathbf{X}})\Omega^T \\ \Omega\text{vec}(\hat{\mathbf{X}})^T & \Sigma_{22.1} + \Omega\Omega^T \end{pmatrix} dF(\hat{\mathbf{X}}|\hat{\mathbf{u}},\hat{\mu},\hat{\Sigma})\mathcal{I}. \tag{19}$$

These expectations are given in terms of $\int \text{vec}(\hat{\mathbf{X}})dF(\hat{\mathbf{X}}|\hat{\mathbf{u}},\hat{\mu},\hat{\Sigma})$ and $\int \text{vec}(\hat{\mathbf{X}})\text{vec}(\hat{\mathbf{X}})^T dF(\hat{\mathbf{X}}|\hat{\mathbf{u}},\hat{\mu},\hat{\Sigma})$ which coincide with EM expressions for configurations with $k-p$ landmarks.

6 Extensions

6.1 Shape regression

Recent attempts to study the shape change in time are based on the Procrustes tangent coordinates or spherical splines in Kendall shape spaces (see Kent et al. 2001, Kume et al. 2007). In this part we will focus on a particular model based on the offset shape distributions.

Let us assume that our observations are shapes of

$$\mathbf{X}_i^\dagger = \mathbf{A}_0^\dagger + \mathbf{A}_1^\dagger t_i + \mathbf{E}_i^\dagger \quad i = 1,\ldots,n$$

where $\mathbf{A}_0^\dagger$ and $\mathbf{A}_1^\dagger$ are matrices of the same dimension as $\mathbf{X}_i^\dagger$, t_i are observation time points and $\mathbf{E}_i^\dagger$ are errors from $\mathcal{N}_{2k}(\mathbf{0}_{2k},\sigma^2\mathbf{I}_{2k})$. In the preform space of Helmertized landmarks $\mathbf{X}^H = \mathbf{K}\mathbf{X}^\dagger$ we now have

$$\mathbf{X}_i^H = \mathbf{A}_0 + \mathbf{A}_1 t_i + \mathbf{E}_i \quad i = 1,\ldots,n$$

where $\mathbf{A}_0 = \mathbf{K}\mathbf{A}_0^\dagger$, $\mathbf{A}_1 = \mathbf{K}\mathbf{A}_1^\dagger$ and $\mathbf{E}_i = \mathbf{K}\mathbf{E}_i^\dagger$. Since the rows of $\mathbf{K}$ are orthogonal then one can easily see that $\mathbf{E}_i = \mathbf{K}\mathbf{E}_i^\dagger$ are generated from $\mathcal{N}_{2k-2}(\mathbf{0}_{2k-2}, \sigma^2\mathbf{I}_{2k-2})$. The equation above can be written as

$$\mathbf{X}_i^H = A\mathbf{T}_i + \mathbf{E}_i \quad i = 1, \ldots, n$$

where $\mathbf{A} = (\mathbf{A}_0, \mathbf{A}_1)$ and $\mathbf{T}_i = \mathbf{I}_2 \otimes (1, t_i)^T$. If $\mathbf{u}_i$ are the shape coordinates of $\mathbf{X}_i^H$, the corresponding log-likelihood function is

$$L(\mathbf{A}, \sigma) = \sum_{i=1}^{n} \log f_{\mathbf{U}}(\mathbf{u}_i; \mathbf{A}\mathbf{T}_i, \sigma).$$

In the following, we address the question of likelihood estimation for the parameter matrix $\mathbf{A}$ without dealing with the identification issues. The method here consists of simple EM update rules which produce stationary points of the corresponding likelihood function.

Without loss of generality we can fix σ to 1 and so the parameters to estimate are only elements of $\mathbf{A}$ modulo the rotation effect. We can deal with rotation invariance by making sure that after each iteration, the *intercept* $\mathbf{A}_0$ is a configuration whose first landmark is a point on the real axis of coordinates, namely, the first row of $\mathbf{A}_0$ has the second component 0. The corresponding $\mathcal{Q}$ function for the EM is now

$$\mathcal{Q}_{\mathbf{A}_r}(\mathbf{A}) = \sum_{i=1}^{n} \int \log(f_{\mathcal{N}}(\mathbf{X}_i; \mathbf{A}\mathbf{T}_i, \mathbf{I}_{2k-2}))dF(\mathbf{X}_i|\mathbf{u}_i, \mathbf{A}_r\mathbf{T}_i, \mathbf{I}_{2k-2}).$$

Proceeding in the same way as in linear regression theory, it can be seen that the update rules for the parameters in $\mathbf{A}$ are

$$\mathbf{A}_{r+1} = \sum_{i=1}^{n} \left(\int \mathbf{X}_i dF(\mathbf{X}_i|\mathbf{u}_i, \mathbf{A}_r\mathbf{T}_i, \mathbf{I}_{2k-2})\mathbf{T}_i^T \right) B^{-1} \tag{20}$$

where $B = \sum_{i=1}^{n} \mathbf{T}_i\mathbf{T}_i^T$. The E-step in this case is completed based on Lemma 2.

If however, our shape observations are obtained from a collection of m paths such that

$$\mathbf{X}_{ij}^H = A\mathbf{T}_i + \mathbf{E}_{ij}, \quad i = 1, \ldots, n \quad \text{and} \quad j = 1, \ldots, m \tag{21}$$

where $\mathbf{E}_{ij}$ are generated from $\mathcal{N}_{2k-2}(\mathrm{vec}(\mathbf{0}_{(k-1)\times 2}), \sigma^2\mathbf{I}_{2k-2})$, then the update rules are

$$\mathbf{A}_{r+1} = \sum_{j=1}^{m}\sum_{i=1}^{n} \left(\int \mathbf{X}_{ij} dF(\mathbf{X}_{ij}|\mathbf{u}_{ij}, \mathbf{A}_r\mathbf{T}_i, \mathbf{I}_{2k-2})\mathbf{T}_i^T \right) B^{-1}. \tag{22}$$

We shall apply this method to the second example in Section 7.

6.2 Mixture distributions

In practice it is possible that shape data is better explained by a mixture of offset normal shape distributions (see e.g Burl 1997). The EM can still be applied with relative ease. Assume for

example that $\mathbf{X}_1^\dagger, \cdots, \mathbf{X}_n^\dagger$ is a sample from some mixture of Gaussian distributions with pdf $\mathcal{F}^\dagger(\mathbf{x}^\dagger) = \sum_{\alpha=1}^M f_{\mathcal{N}}(\mathbf{x}^\dagger; \mu_\alpha^\dagger, \Sigma_\alpha) p_\alpha$ where $p_\alpha > 0$ such that $\sum_{\alpha=1}^M p_\alpha = 1$. Clearly the induced pdf of preform $\mathbf{X}$ is $\mathcal{F}(\mathbf{x}) = \sum_{\alpha=1}^M f_{\mathcal{N}}(\mathbf{x}; \mu_\alpha, \Sigma_\alpha) p_\alpha$ and the induced distribution of the shape variables $\mathbf{u}$ will be $\mathcal{F}_{\mathbf{U}}(\mathbf{x}) = \sum_{\alpha=1}^M f_{\mathbf{U}}(\mathbf{u}; \mu_\alpha, \Sigma_\alpha) p_\alpha$ where $f_{\mathbf{U}}(\mathbf{u}; \mu_\alpha, \Sigma_\alpha)$ is defined as in section 2. The parameters that we need to estimate here are μ_α,Σ_α and p_α where $\alpha = 1, \cdots, M$. The EM algorithm in these cases can be applied by considering α and variables $\mathbf{h}$ as *hidden* and constructed in exactly the same way as that given in section 2.7.2 of McLachlan and Krishnan (1997) for finite mixtures of multivariate Gaussian component densities. Taking derivatives with respect to variables and equating them to zero we find the update rules

$$
\begin{aligned}
p_\alpha^{r+1} &= \frac{1}{n} \sum_{i=1}^n P(\alpha|\mathbf{u}_i) \\
\mu_\alpha^{r+1} &= \frac{1}{\sum_{j=1}^n P(\alpha|\mathbf{u}_j)} \sum_{i=1}^n P(\alpha|\mathbf{u}_i) \int \mathrm{vec}(\mathbf{X}_i) dF(\mathbf{X}_i|\mathbf{u}_i; \mu_\alpha^r, \Sigma_\alpha^r) \\
\Sigma_\alpha^{r+1} &= \frac{1}{\sum_{j=1}^n P(\alpha|\mathbf{u}_j)} \sum_{i=1}^n P(\alpha|\mathbf{u}_i) \int \mathrm{vec}(\mathbf{X}_i)\mathrm{vec}(\mathbf{X}_i)^T dF(\mathbf{X}_i|\mathbf{u}_i; \mu_\alpha^r, \Sigma_\alpha^r) - \mu_\alpha^{r+1}\mu_\alpha^{r+1^T}
\end{aligned}
$$

where

$$
P(\alpha|\mathbf{u}_i) = \frac{f_{\mathbf{U}}(\mathbf{u}_i; \mu_\alpha^r, \Sigma_\alpha^r) p_\alpha^r}{\sum_{\beta=1}^M f_{\mathbf{U}}(\mathbf{u}_i; \mu_\beta^r, \Sigma_\beta^r) p_\beta^r}.
$$

These update rules are similar to those in Section 2. In the mixture case however, the influence of every data point on μ_α^{r+1} and Σ_α^{r+1} is weighted by a factor of $\frac{P(\alpha|\mathbf{u}_i)}{\sum_{j=1}^n P(\alpha|\mathbf{u}_j)}$. For the complex normal cases the algorithm has a similar form involving expressions as in Lemma 2.

7 Applications

For illustrative purposes we choose two data sets which have previously been studied in the shape analysis literature. The first is chosen in order to illustrate the estimation of the mean and covariance of the offset normal distributions for different covariance structures. The second example is chosen to demonstrate the application of EM for the shape regression model introduced in the previous section. In addition we have also used synthetic datasets to evaluate the full covariance model.

If the variation of landmarks is much smaller than the mean baseline (as is typically the case), taking as parameter starting values those derived from the normal approximations given in 6.59 of Dryden and Mardia (1998) seem to work well. For simple covariance structures, the Procrustes mean and identity matrix as starting values for μ and C seems to work well in all the following

examples that we have explored. Moderate variations from those starting points do not seem to matter. Except for the last experiment on synthetic data, all calculations were carried out on a modern PC with processor speed 3Mhz, and the calculation time for running each EM took no longer than 10 minutes.

7.1 Gorilla skulls

The first application is based on careful measurements of locations of 8 landmarks chosen in the middle plane of 29 male and 28 female gorilla skulls. This data set is studied in detail in (see e.g. 7.1.2 in Dryden and Mardia 1998) where some inference results are reported based on Procrustes tangent space coordinates. In the following, we will use the likelihood approach for checking whether there is any sex shape difference between these groups and explore the covariance structure between landmarks.

A schematic representation of male and female skulls is provided by the polygons shown in Figure 1 where the first two vertices are (0,0) and (1,0) respectively. Bookstein shape coordinates are those of the remaining vertices.

The EM algorithm applied to male skulls produces the log-likelihood values at the mle estimates 1048.48, 981.415 and 874.971 for general, complex normal and isotropic covariance structures respectively. The shapes of the means for each model do not differ much since the Kendall shape distance ρ between any two of them (including the Procrustes mean) is less than 0.006. Analogous log-likelihood values for the female group are 1110.5, 1054.28 and 954.01 with the interpoint shape distance ρ between mean shapes not exceeding 0.005. However, the Kendall shape distances between the means of males with those of females are around 0.056.

Figure 1 about here

If we are to test $H_0 : \Theta \in \Omega_0$ versus $H_1 : \Theta \in \Omega_1$ where $\Omega_0 \subset \Omega_1$ then for large samples and under regularity conditions, the likelihood ratio test is based on

$$-2 \log \Delta = 2(\sup_{H_1} \log L(\Theta) - \sup_{H_0} \log L(\Theta)) \sim \chi^2_r$$

where $r = \dim(\Omega_1) - \dim(\Omega_0)$.

Let us assume that the preforms of males and female skulls are generated from complex normal distributions $\mathcal{CN}_7(\eta_{\mathbf{m}}, C_m)$ and $\mathcal{CN}_7(\eta_{\mathbf{f}}, C_f)$ respectively. If we want to test whether these means differ from each other only by some rotation we consider the hypothesis test

$$H_0 : \eta_{\mathbf{m}} = \eta_{\mathbf{f}} \mathrm{mod(rot)}, C_m = C_f \quad \text{versus} \quad H_1 : \eta_{\mathbf{m}} \neq \eta_{\mathbf{f}} \mathrm{mod(rot)}, C_m = C_f.$$

The degrees of freedom, i.e. the difference between the dimensionality of parameter spaces in this case is 13. If we run the EM for the pooled sample, the log-likelihood value at the mle estimates is 1940.39 where the estimated covariance matrix C is given in Table 1. Based on the remark at the end of subsection 4.1, the likelihood values for the alternative hypothesis can be obtained by

running the EM separatively for each group while keeping the entries of $C_m = C_f = C$ the same as those of Table 1. We obtain the maximum log-likelihood values 954.4 and 1026.6 for the groups of males and females. Hence, $-2\log\Delta$ distributed as χ^2_{13} under H_0 is 80.2. Since $P(\chi^2_{13} > 80.02)$ is almost zero there is a strong evidence that modulo rotations, $\eta_{\mathbf{m}}$ and $\eta_{\mathbf{f}}$ are different.
These test results are consistent with those in example 7.2 of Dryden and Mardia (1998) where procrustes tangent coordinates are used to suggest shape gender difference.

Table 1 about here

From the covariance matrix estimates we can infer some results about the correlation between landmarks. For example, the complex covariance matrix for the pooled sample given in Table 1, indicates that the imaginary component has relatively small values. This suggests that the covariance between x-coordinates and y-coordinates is relatively small. The value of the fifth diagonal element of C (which is the variance of the length distance between landmark six and one) is the smallest, suggesting that with respect to the others, the sixth landmark varies the least. This could be easily confirmed visually from Figure 1.

7.2 Shape Regression for Rat skulls

The data set considered here consists of the position of 8 biological landmarks of the skulls of 18 different rats. This data set is studied in several publications (e.g. Bookstein 1991, Mardia and Walder 1994, Le and Kume 2000b). The data are obtained by X-ray images of each of these rats which are observed at 8 different time points when they are 7, 14, 21, 30, 40, 60, 90, and 150 days old.

We apply to this data the regression model (21) where the update rules for maximizing the likelihood are as in (22). The EM algorithm in this case converges reasonably quickly where the starting value for $\mathbf{A}$ is taken such that $\mathbf{A}_0$ is the first observation and $\mathbf{A}_1$ is the matrix of zeros. However, different choices of the starting values for parameters do not seem to alter the solution. The resulting values of the estimated parameter matrices $\mathbf{A}_0$ and $\mathbf{A}_1$ defined in subsection 6.1 are given in Table 2. One entry of $\mathbf{A}_0$ is fixed at zero to ensure rotation invariance. Based on these parameter values, the Bookstein shape coordinates obtained via equation (1) for both fitted and observed configurations are shown in Figure 2. In Le and Kume (2000b) the mean shapes for each time point shapes are shown to develop closely to a geodesic line in shape spaces. The geodesic line which starts from the mean shape at time 7 and ends at time 150 is included in Figure 2 dot-dashed lines. Our fitted regression shapes in solid lines seems to fit to the data better.

Figure 2 about here

Table 2 about here

7.3 The Full Covariance Model

Parameters are difficult to identify in the full covariance EM-algorithm. To test the validity of this method for estimating the probability density we conducted the following experiment. We generated 20 training datasets for each of the following sizes: $n = \{25, 50, 100, 250, 500, 1000, 1500\}$ (i.e. we generated $20 \times 7 = 140$ training datasets in total). For each of these datasets we also generated a separate test set from the same density of 500 data-cases.

Each dataset was generated from the same offset normal shape distribution using three landmarks with means $\mu_1 = [1, 0]^T, \mu_2 = [0, 2]^T, \mu_3 = [0, 0]^T$ and variances $S_1 = \text{diag}(1, 0.5), S_2 = \text{diag}(2, 1), S_3 = \text{diag}(3, 1.5)$, which were subsequently being transformed into Bookstein shape coordinates.

The algorithm of section 3 was run on each of the 140 training datasets with random initializations for the mean and covariance estimates. After each iteration of EM the mean of the first landmark was realigned with the x-axis (this has no effect on the offset-normal density). EM learning was terminated whenever the log-likelihood per data-point increased less than $1E-5$ or when 2000 updates were performed, whichever came first. Standard errors were computed for each value of dataset size n across the 20 repetitions of the experiment.

Results are shown in figures 3 and 4. In terms of the root mean square error for the parameters μ and S, we clearly observe they do not converge to 0 as we increase the datasets size[1]. In contrast, the log-likelihood curves for train data and test data converge smoothly to the same value. Moreover, they also converge to the same value as the log-likelihood computed using the true parameter values. This strongly suggests that our estimate of the probability density does converge to the true offset normal density from which the data was generated. To show the details of the behavior for large datasets we separately plot the log-likelihood curves for a restricted subset in Figure 4 (right).

Figures 3,4 about here

Concluding remarks

It follows from an early result of Fisher (see equation 1 Jamshidian and Jennrich 2000) that

$$dL(\mu, \Sigma)) = \{d\mathcal{Q}_{\mu_r, \Sigma_r}(\mu, \Sigma)\}_{\mu_r=\mu, \Sigma_r=\Sigma}.$$

[1]We compared parameters only after applying a linear coordinate transformation to both the true and estimated mean and covariance. This transformation removes a redundancy in the parametrization and has no effect on the density, see section 2.2.

In particular, the values of gradient function of the likelihood can be obtained by straightforward substitutions of (7) and the results of Lemma 1. Therefore, the results of this paper can be also used to develop alternative, possibly quicker, gradient likelihood optimization methods (see Jamshidian and Jennrich (1997) for various EM acceleration methods). However, our EM approach consists of algebraically convenient update expressions which ensure that the likelihood function is always increasing and, for a particular covariance structure, it resembles the ordinary Procrustes algorithm.

The algorithm that we propose in this paper runs without problem on various data sets that we have tested. The EM is efficient here since the amount of missing data is relatively small, it has simple update rules and does not have numerical instabilities in the sense that the likelihood function always increases.

In the complex normal case the consistency of the estimators obtained here is automatically implied from the general likelihood theory. In particular, simulation results applied to the isotropic case considered in Subsection 4.3 show that the algorithm converges quickly to the true parameter values including the concentration parameter of the shape distribution. For the general covariance case we show in section 7.3 that we have convergence of the probability density but not consistency. We think that this is probably due to the non-identifiability of parameters.

In general, one drawback of EM algorithms is that sometimes the successive steps in parameter space can be small and therefore we might decide to stop the algorithm too early. Since the likelihood function and its gradient is known explicitly, equation 5.1 in Jamshidian and Jennrich (1997) could be adopted here as a stoping rule.

Using the likelihood function one can obtain numerically the Hessian matrix to give estimated standard errors of the mle estimators. They can also be obtained from second order differentials of the log-likelihood function $L(\mu, \Sigma)$. These expressions can in principle be calculated explicitly (not shown here) but they will be cumbersome and numerical problems are likely to appear due to high dimensional matrix inversions. This is a problem which needs further investigation since the calculation of second order derivatives of the likelihood could lead to quicker optimization methods. Jamshidian and Jennrich (2000) consider various methods for obtaining the Hessian of the likelihood using the gradient.

Running the algorithm from different starting points in the parameter space can increase the chance of finding the global maximum. However, more theoretical work is needed to explore the existence of multiple modes in the likelihood function. In particular, Dryden (1989) identifies situations where singular Gaussian distributions in configuration space can lead to non-degenerate offset shape distributions. Note that for the widely used general Procrustes algorithm, unless the sample points are within a *regular ball*, the Procrustes mean is not necessarily unique (c.f. 9.1 in Kendall et al. 1999).

One possible generalization of this EM approach is to work with the induced affine shape distributions of Gaussian distributed planar configurations (see e.g. Leung et.al. 1998). Clearly, for the shapes of 3 dimensional configurations the update rules are the same as those in (8) and (9), but the corresponding expectations are much more complicated and a possible approach will be to calculate these integrals numerically.

8 Supplemental Materials

Apendix The proofs of Lemmas 1 and 2 (Appendix.pdf).

Data and R code The relevant datasets and the basic code for running the EM for complex covariance and shape regression can be found in a zipped format. The readme file there contains starting instructions [2] (dataRcode.zip).

Aknowledgement

We would like to thank Ian Dryden for general discussions about this work and the anonymous referees for their suggestions.

References

Bookstein, F.L. (1986). Size and shape spaces for landmark data in two dimensions. *Stat. Sci.* **1**, 181-242.

Bookstein, F.L. (1991). *Morphometric tools for landmark data*. Cambridge University Press.

Brignell, C.J., Browne, W.J., & Dryden, I.L. (2005). *Covariance weighted Procrustes analysis*. In S. Barber, P.D. Baxter, K.V.Mardia, & R.E. Walls (Eds.), Quantitative Biology, Shape Analysis, and Wavelets, 107-110. Leeds, Leeds University Press.

Burl, M.C. (1997). *Recognition of visual object classes*, Phd Thesis. California Institute of Technology, Pasadena, CA.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc.*, Series B, **34**, 1-38.

Dryden, I.L. (1989). *The statistical analysis of shape data*. Phd Thesis, University of Leeds, U.K.

[2] For more information please contact the authors directly.

Dryden, I.L. & Mardia, K.V. (1991). General shape distributions in a plane. *Adv. Appl. Prob.* **23**, 259-276.

Dryden, I.L. & Mardia, K.V. (1998). *Statistical Shape Analysis*. John Wiley, Chichester.

Gradshteyn, I.S. & Ryzhik, I.M. (1980). *Table of Integrals, Series, and Products*. Academic Press.

Jamshidian, M. & Jennrich, R. I. (2000). Standard Errors for EM Estimation.*J. Roy. Stat. Soc.*, Series B. **62**, 257-270.

Jamshidian, M. & Jennrich, R. I. (1997). Acceleration of the EM Algorithm by Using Quasi-Newton Methods.*J. Roy. Stat. Soc.* , Series B. **59**, 569–587.

Kendall, D.G., Barden, D., Carne, T.K. & Le, H. (1999). *Shape and Shape Theory*. John Wiley & Sons.

Kent, J.T. & Mardia, K.V. (1997). Consistency of Procrustes Estimators. *J. Roy. Stat. Soc*, Series B **59**, 281-290.

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

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*, 109-115. Leeds University Press.

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

Le, H. (2003). Unrolling shape curves. *J. London Math. Soc.* **68**(2), 511 - 526.

Le, H. (1998). On the consistency of procrustean mean shapes *Adv. Appl. Probab.* **30**, 53-63.

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

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

Leung, T. K., Burl M. C. & Perona, P., (1998). Probabilistic affine invariants for recognition, *In Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn*, 678–684.

Table 1: MLE estimates (re-scaled by 10^2) for the complex covariance matrix of the pooled sample

$10^2 C$						
3.24+0.00i	2.53+0.50i	2.05+0.45i	1.41+0.29i	0.36+0.01i	1.39−0.62i	2.54−0.77i
2.53−0.50i	2.14+0.00i	1.74+0.07i	1.18+0.03i	0.30−0.04i	1.03−0.66i	1.87−0.96i
2.05−0.45i	1.74−0.07i	1.48+0.00i	1.00+0.02i	0.25−0.02i	0.87−0.55i	1.55−0.79i
1.41−0.29i	1.18−0.03i	1.00−0.02i	0.73+0.00i	0.19−0.02i	0.62−0.39i	1.10−0.54i
0.36−0.01i	0.30+0.04i	0.25+0.02i	0.19+0.02i	0.07+0.00i	0.17−0.08i	0.31−0.09i
1.39+0.62i	1.03+0.66i	0.87+0.55i	0.62+0.39i	0.17+0.08i	0.82+0.00i	1.31+0.18i
2.54+0.77i	1.87+0.96i	1.55+0.79i	1.10+0.54i	0.31+0.09i	1.31−0.18i	2.30+0.00i

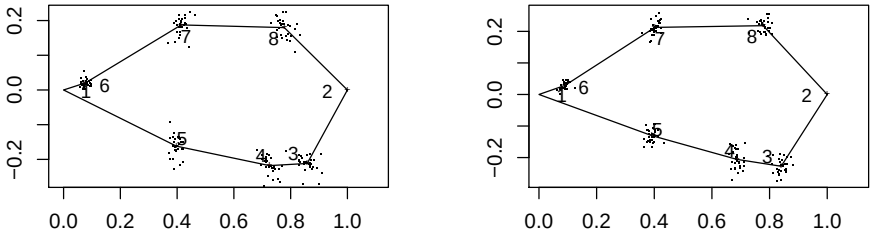

Figure 1: Bookstein shape coordinates for male (left) and female (right) gorilla skulls and a schematic representation of their mean shape with the first two landmarks standartized.

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

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

Mardia, K.V. & Walder, A.N. (1994). Shape analysis of paired landmark data. *Biometrika* **81**, 185-196.

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

M. Welling (2005). An Expectation Maximization Algorithm for Inferring Offset-Normal Shape Distributions. *Tenth Internat. Work. Artif. Intell. and Stat.*, **10**, Babados.

Table 2: Parameter estimates for the simple linear regression model.

$\mathbf{A}_0$		$\mathbf{A}_1$	
7.875	0.000	0.059	−0.084
8.855	−4.721	0.030	−0.184
6.135	−12.777	−0.132	−0.179
−3.688	−21.964	−0.320	−0.224
−19.768	−13.784	−0.524	−0.078
−13.093	−3.732	−0.277	0.030
−8.333	4.701	−0.025	0.163

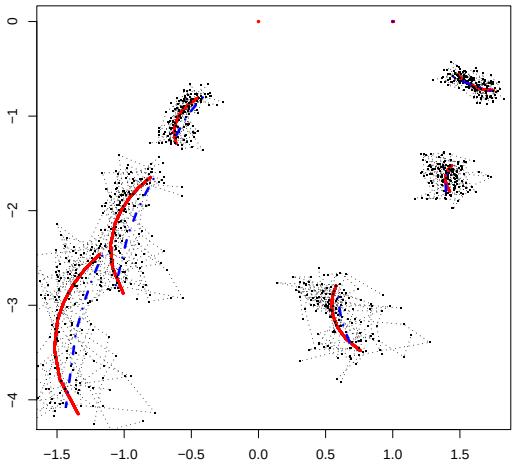

Figure 2: The change in time of Bookstein shape coordinates. Each line shows the path of a particular landmark in time. Observed configuration landmarks are in dashed lines, geodesic fitted configurations are in dot dashed lines and those of the mle fitted path are in solid lines. The baseline is defined by points $(0, 0)$ and $(1, 0)$

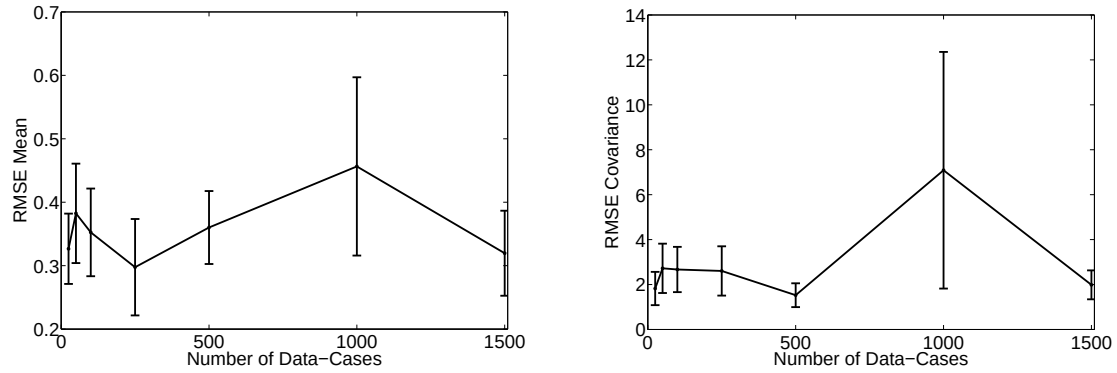

Figure 3: The RMSE for the mean (left) and covariance (right) as a function of the number of training data-points.

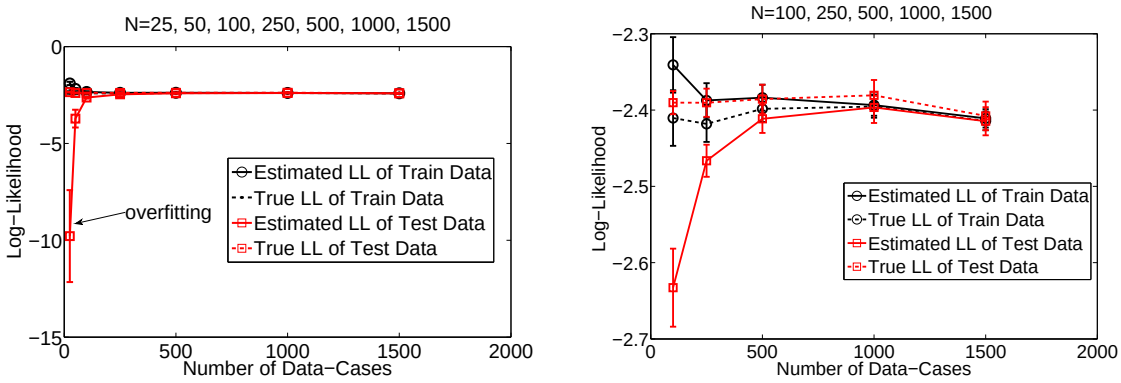

Figure 4: The log-likelihood (LL) as a function of the number of data-cases. The right figures display the same lines but with $N = 25$ and $N = 50$ removed to better show the convergence for large values of N. We compare A) the LL with parameters estimated from training data (solid black line), B) the LL with true parameters (dashed black line), C) the LL of test data with parameters estimated on training data (solid red line), D) the LL of test data with true parameters (dashed red line).