THE OFFSET NORMAL SHAPE DISTRIBUTION FOR DYNAMIC SHAPE ANALYSIS

LARA FONTANELLA* – LUIGI IPPOLITI

University G. d'Annunzio, Chieti-Pescara, Italy

lfontan@unich.it, ippoliti@unich.it

ALFRED KUME

University of Kent, Canterbury, CT2 7NF, UK

(e-mail: a.kume@kent.ac.uk)

September 24, 2018

Abstract. This paper deals with the statistical analysis of landmark data observed at different temporal instants. Statistical analysis of dynamic shapes is a problem with significant challenges due to the difficulty in providing a description of the shape changes over time, across subjects and over groups of subjects. There are several modelling strategies which can be used for dynamic shape analysis. Here, we use the exact distribution theory for the shape of planar correlated Gaussian configurations and derive the induced *offset-normal* shape distribution. Various properties of this distribution are investigated, and some special cases discussed. This work is a natural progression of what has been proposed in Mardia and Dryden (1989), Dryden and Mardia (1991), Mardia and Walder (1994) and Kume and Welling (2010).

Key Words: shape analysis, offset-normal shape distribution, EM algorithm, spatio-temporal correlations

*Corresponding Author: Viale Pindaro 42, 65127 Pescara, ITALY

1 Introduction

This article is concerned with some inferential issues arising from the analysis of dynamic shapes. Describing, measuring and comparing the shape of objects is very popular in a variety of different disciplines (see, for example, Dryden and Mardia, 2016, section 1.2). Much work has been done for static or cross-sectional shape analysis, while considerably less research has focused on dynamic or longitudinal shapes.

If the objects of interest are two-dimensional, their shape features can be represented either by a set of points or by a continuous closed line in the real plane. Here, we suppose that an object can be reliably represented by a configuration of homologous *landmarks* labelled in the same order. It is also assumed that such configurations are realizations from a random matrix, $\mathbf{X}^\dagger = (x_k^\dagger \; y_k^\dagger), k = 1, \ldots, K$, of K landmarks in $\mathbb{R}^2$. Row k of $\mathbf{X}^\dagger$, thus contains the Euclidean coordinates for landmark k.

Shape is typically defined as the geometrical information that remains when location, scale and rotational effects are removed from an object (Dryden and Mardia, 2016). The shape of $\mathbf{X}^\dagger$, denoted as $[\mathbf{X}^\dagger]$, is then the equivalent class of configurations such that, $[\mathbf{X}^\dagger] = \left\{\beta \mathbf{X}^\dagger \mathbf{R} + \boldsymbol{\tau} \mid \beta > 0, \mathbf{R} \in SO(2), \boldsymbol{\tau} \in \mathbb{R}^2\right\}$, where the actions of β, $\mathbf{R}$ and $\boldsymbol{\tau}$ on $\mathbf{X}^\dagger$ represent all the possible rescalings, rotations and translations. The space of these equivalence classes, denoted in the literature as Σ_K^2, is called the shape space and various metrics applied on it determine the type of the shape space constructed. The shape metrics which give rise to geometrical models for shape spaces are induced via the quotient map

$$\begin{aligned} \pi : \quad \mathbb{R}^{K \times 2} &\longrightarrow \Sigma_K^2 \\ \mathbf{X}^\dagger &\longrightarrow [\mathbf{X}^\dagger] \end{aligned}$$

where some original metric on the space of coordinates $\mathbf{X}^\dagger$ is assumed. For example, for the shapes of planar triangles one can construct either positive curved spaces called Kendall's spherical model or a negative curved space called the Bookstein hyperbolic space of triangles (see, e.g., Le and Small, 1999).

When a temporal sequence of landmark data is available, the resulting sequence of shape observations can be seen as generated via the quotient map of a product on configuration spaces as

$$\begin{array}{rcl}\pi : \mathbb{R}^{K\times 2} \times \mathbb{R}^{K\times 2} \cdots \times \mathbb{R}^{K\times 2} & \longrightarrow & \Sigma_K^2 \times \Sigma_K^2 \cdots \times \Sigma_K^2 \\ \left(\mathbf{X}_1^\dagger, \mathbf{X}_2^\dagger, ..., \mathbf{X}_T^\dagger\right) & \longrightarrow & \left([\mathbf{X}_1^\dagger], [\mathbf{X}_2^\dagger], ..., [\mathbf{X}_T^\dagger]\right).\end{array}$$

There are several aspects of shape analysis which merit the attention of statisticians. For example, these include visualization, estimation and testing. Related statistical issues, which are of particular interest for this paper, also concern modelling assumptions and the description of the changes over time in shape.

In practice, at least four modelling strategies can be used for shape analysis. One possibility is to develop models in shape spaces. Some examples on this line are as in Kenobi et al. (2010), Le and Kume (2000) and Kume et al. (2007). However, because of the non-Euclidean nature of the shape space, the definition of natural models and probability distributions, especially in a dynamic setting, is not straightforward. For example, in the static case, a family of shape distributions specified along the lines of Kent distributions on the sphere, is proposed by Kent et al. (2006) as the complex Bingham quartic distribution, where the mean and covariance are analogously defined as in the multivariate Gaussian distribution in Euclidean space. The normalising constant of these distributions however, has to be calculated numerically and the extension of the model to a dynamic setting is difficult.

Another approach, pioneered by Bookstein (1986), is to use an unrestricted multivariate Normal distribution in Bookstein coordinates. This methodology has the advantage of simplicity, but the inference depends on the baseline chosen.

Another possibility is to build models in tangent space to shape space. This approach, which is very common in shape analysis, first requires the estimation of a mean shape at which to take the projection. Procrustes analysis can be used to estimate shapes and Le (1998) and Kent and Mardia (1997) have shown that, for concentrated and isotropic data, the *Procrustes mean shape* is a consistent estimator of the shape of the mean configuration. However, if the distribution for the landmarks is not isotropic or the variability of the data is relatively high, then inconsistencies can arise. Hence, approaches to inference based on tangent space approximation, are valid in datasets with small variability in shape. For a review of statistical tests based on normality assumptions on the tangent space and non-parametric alternatives see, for example, Dryden and Mardia (2016, pg.185). A discussion of using combination-based permutation tests in a dynamic shape analysis setting can also be found in Brombin et al. (2015) and Brombin et al. (2016).

A fourth approach, which probably provides the simplest model for landmarks, assumes a multivariate Normal distribution with mean configuration, $\boldsymbol{\mu}^\dagger$, and covariance matrix, $\boldsymbol{\Sigma}^\dagger$. Various levels of generality can be considered for the covariance matrix, assuming either isotropy or structured correlations between and within landmarks. For the static case, Mardia and Dryden (1989) and Dryden and Mardia (1991) worked out the induced distributions in the shape space under this model.

In this paper, we follow this modelling strategy and extend the structure of the landmark correlations by specifying time series models for shapes. In order to illustrate the natural representation of temporal dependence among landmarks, consider the simple first order stationary Autoregressive (AR) model for configurations,

$$\mathbf{X}_t^\dagger = \phi_1 \mathbf{X}_{t-1}^\dagger + (1-\phi_1)\boldsymbol{\mu}^\dagger + \mathbf{E}_t^\dagger \quad \mathbf{E}_t^\dagger \sim \mathcal{N}_{2K}(\mathbf{0}, \sigma^2 \mathbf{I})$$

where $\mathbf{E}_t^\dagger$ is an isotropic error term and ϕ_1 the autoregressive parameter. By using a marginal approach (Dryden and Mardia, 2016, pg.217), it is easy to show that, given $\mathbf{X}_{t-1}$, the induced shape distribution of $[\mathbf{X}_t^\dagger]$ is isotropic in the shape space with centre at $[\phi_1 \, X_{t-1} + (1-\phi_1)\boldsymbol{\mu}]$. Such conditional dependence of shapes $[\mathbf{X}_t^\dagger]$ is shown in Figure 1, where the dynamic of the shape of $\mathbf{X}_t^\dagger$ is determined by the shape $[\mathbf{X}_{t-1}^\dagger]$ and a pulling effect towards the marginal mean shape $[\boldsymbol{\mu}^\dagger]$. In fact, the shape auto-correlation structure in this case is the one induced by the first order autoregressive representation of configurations onto the shape space via the quotient map. The geometrical interpretation in the shape space of the AR(1) model also suggests that the correlation in this landmark-based model seems natural in the shape space where the pulling towards a mean direction and the distribution of the innovations remain isotropic without any preferred direction, like a zero mean error in multivariate time series models.

Specifying models directly on landmarks has an intuitive appeal and enjoys several advantages. For example, the model is defined in the landmark space where the mean configuration and landmark correlation are naturally defined. This is of practical importance as practitioners want to build models based on assumptions in the configuration space and interpret their estimated quantities, such as mean and correlation, in terms of landmarks.

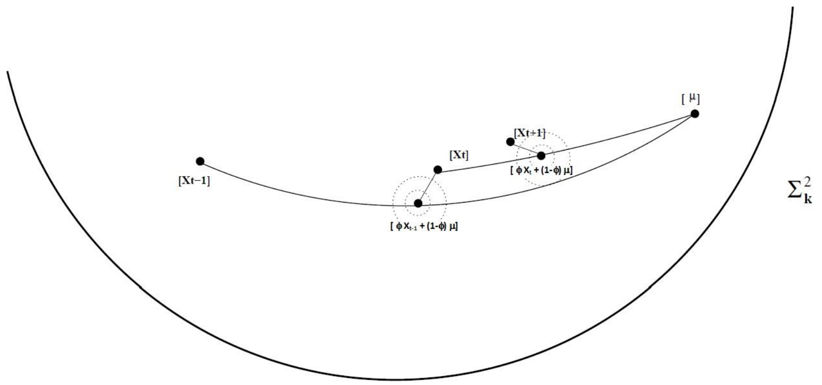


Figure 1: The conditional distribution of $[\mathbf{X}_t^\dagger]$ given $[\mathbf{X}_{t-1}^\dagger]$ is isotropic in the shape space with center at $[\phi_1\ X_{t-1} + (1-\phi_1)\boldsymbol{\mu}]$

Second, the maximum likelihood (ML) estimate of mean shape will always be consistent, provided the parameters of $\mathbf{\Sigma}^\dagger$ are identifiable. Hence, for structured correlations and/or relatively dispersed shape data, inferential results from the likelihood approach could be preferred to those carried out via the Procrustes tangent coordinates, since the tangent space approximation in shape spaces is only appropriate for small regions in the shape space (Dryden and Mardia, 2016, pg.185). Direct evidence of this will be shown by a simulation exercise in Section 6.

Third, the ML approach enables us to perform automatic model selection, construct maximum likelihood ratio tests for a wide range of inference problems and cope with missing data, a feature not immediately available for the Procrustes shape space approach.

Finally, when autoregressive models are used, k-step ahead forecasts can also be obtained for a time series sequence of shape coordinates. To our knowledge, this is the first time an AR process is used to produce forecasts of the mean shape.

The paper is organized as follows. In Section 2, we derive the shape space density induced by a Gaussian distribution on a temporal sequence of landmark configurations. In this section, we also show that our formulation generalizes the results given in Mardia and Dryden (1989) and Dryden and Mardia

(1991) and discuss the difficulties of computing the expectation of a product of quadratic forms, a step needed for the evaluation of the density. In Section 3, we discuss model parameter estimation and give the general update rules of the Expectation-Maximization (EM) algorithm for general $\boldsymbol{\mu}^\dagger$ and $\boldsymbol{\Sigma}^\dagger$ in a dynamic framework. By exploiting the separability structure between "space" and time, a computationally efficient recursive algorithm to estimate AR processes is also introduced. By taking account of the temporal correlation, we show that our approach generalizes the EM algorithm proposed by Kume and Welling (2010). In Section 4 we discuss the difficulties associated with the computation of the expectations required by the E-step and provide a technical result which facilitates the calculations. Then, Section 5 addresses issues related to relabelling invariance of landmarks and Section 6 shows results from a set of simulation studies. In Section 7 we illustrate our method by examining three real data sets and, finally, we conclude the paper in Section 8 with a discussion.

2 The offset normal distribution in dynamic shape analysis

In this section we derive the shape space density induced by a Gaussian distribution on a finite set of $K \geq 3$ not-all-coincident labelled landmarks, $\big\{(x^\dagger_{k,t}\ y^\dagger_{k,t}) \in \mathbb{R}^2, k = 1, \ldots, K\big\}$, observed at $t = 1, \ldots, T$ time points. At a given time t, these landmarks are organized in a $(K \times 2)$ configuration matrix, $\mathbf{X}^\dagger_t$, so that the temporal sequence of the T configurations is denoted as $\mathbf{X}^\dagger = \Big(\mathbf{X}^\dagger_1\ \mathbf{X}^\dagger_2\ \ldots\ \mathbf{X}^\dagger_T\Big)$. For such configurations, it is assumed that the distribution of $\mathbf{X}^\dagger$ is Gaussian, i.e. $vec(\mathbf{X}^\dagger) \sim \mathcal{N}_{2KT}\big(vec(\boldsymbol{\mu}^\dagger), \boldsymbol{\Sigma}^\dagger\big)$, where $vec(\mathbf{X}^\dagger)$ and $vec(\boldsymbol{\mu}^\dagger)$ are $(2KT \times 1)$ vectors and $\boldsymbol{\Sigma}^\dagger$ is a $(2KT \times 2KT)$ covariance matrix. As specified above, various levels of generality can be considered for the covariance matrix, including isotropy (the coordinates of all the landmarks are independent with the same variance) or structured correlations $-$ see Section 3. Given the coordinates of the labelled landmarks, the shape variables are obtained by removing the similarity transformations by translating, rotating and scaling the available configurations. In particular, the effect of translation can be removed by linearly projecting the landmark coordinates to the preform space of centered configurations. In practice, if we map the first landmark of each configuration $\mathbf{X}^\dagger_t$ to the origin, the coordinates of the remaining $K - 1$ vertices can be obtained by left multiplying the temporal configuration matrix $\mathbf{X}^\dagger$ by 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 $(K-1)$ and $\mathbf{1}_{K-1}$ is a $(K-1)$-vector of ones. Therefore, we have $\mathbf{X} = \mathbf{L}\mathbf{X}^\dagger = \Big(\mathbf{L}\mathbf{X}_1^\dagger, \ldots, \mathbf{L}\mathbf{X}_T^\dagger\Big) = \Big(\mathbf{X}_1, \ldots, \mathbf{X}_T\Big)$, where $\mathbf{X}_t$ denotes the preform of configuration $\mathbf{X}_t^\dagger$ and $\mathbf{X}$ is the temporal sequence of preforms. In the preform space, we thus have $vec(\mathbf{X}) \sim \mathcal{N}_{2(K-1)T}\big(vec(\boldsymbol{\mu}), \boldsymbol{\Sigma}\big)$, where $vec(\boldsymbol{\mu}) = vec(\mathbf{L}\boldsymbol{\mu}^\dagger) = \big(\mathbf{I}_{2T} \otimes \mathbf{L}\big)vec(\boldsymbol{\mu}^\dagger)$ and $\boldsymbol{\Sigma} = \big(\mathbf{I}_{2T} \otimes \mathbf{L}\big)\boldsymbol{\Sigma}^\dagger\big(\mathbf{I}_{2T} \otimes \mathbf{L}'\big)$.

The shape space of the centred configurations is obtained by removing the information about rotation and scaling. Without loss of generality, we consider here Bookstein shape coordinates $\mathbf{U}_t = (u_{k,t}\ v_{k,t}),\ k = 1, \ldots, K-1$, with $u_{1,t} = 1$ and $v_{1,t} = 0$. At each time t, $\mathbf{U}_t$ can be computed through the mapping $\mathbf{X}_t \to \mathbf{U}_t = \beta_t \mathbf{X}_t \mathbf{R}_t$, where β_t and $\mathbf{R}_t$ are scaling factors and rotation matrices, respectively (see Appendix 1 for details). Then, the temporal sequence of shape coordinates, $\mathbf{U} = \Big(\mathbf{U}_1\, \mathbf{U}_2\, \ldots\, \mathbf{U}_T\Big)$, is defined as the transformation $\mathbf{X} \to \mathbf{U} = \mathbf{X}\mathbf{R}$, where $\mathbf{R} = diag\big(\beta_1\mathbf{R}_1, \beta_2\mathbf{R}_2, \ldots, \beta_T\mathbf{R}_T\big)$.

To find the distribution of the observed "reduced" shape coordinates, $\mathbf{u} = \big\{(u_{k,t}, v_{k,t}),\ k = 2, \ldots, K-1,\ t = 1, \ldots, T\big\}$, we have to integrate out (or marginalize) the scale and the rotation information contained in the $(2T \times 1)$ vector $\mathbf{h} = (\mathbf{h}_1', \mathbf{h}_2', \ldots, \mathbf{h}_T')'$, where $\mathbf{h}_t = (x_{2,t}, y_{2,t})'$. This can be done by considering the transformation $vec(\mathbf{X}) = \mathbf{W}\mathbf{h}$, where $\mathbf{W} = diag\big(\mathbf{W}_1, \ldots, \mathbf{W}_T\big)$, with

$$\mathbf{W}_t = \begin{pmatrix} 1 & u_{3,t} & \ldots & u_{K,t} & 0 & v_{3,t} & \ldots & v_{K,t} \\ 0 & -v_{3,t} & \ldots & -v_{K,t} & 1 & u_{3,t} & \ldots & u_{K,t} \end{pmatrix}',$$

and writing the joint distribution of $(\mathbf{u}, \mathbf{h})$ as

$$f\left(\mathbf{u}, \mathbf{h}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \frac{1}{(2\pi)^{(K-1)T}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} exp\left\{-\frac{\psi(\boldsymbol{\mu}, \boldsymbol{\Sigma})}{2}\right\} |J\big(\mathbf{X} \to (\mathbf{h}, \mathbf{u})\big)|,$$

where $\psi(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \big(\mathbf{W}\mathbf{h} - vec(\boldsymbol{\mu})\big)'\boldsymbol{\Sigma}^{-1}\big(\mathbf{W}\mathbf{h} - vec(\boldsymbol{\mu})\big)$ and $|J\big(\mathbf{X} \to (\mathbf{u}, \mathbf{h})\big)| = \prod_{t=1}^{T} \|\mathbf{h}_t\|^{2(K-2)}$ is the Jacobian of the transformation $\mathbf{X} \to (\mathbf{u}, \mathbf{h})$. Since the quadratic form $\psi(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be expressed as $\psi(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\mathbf{h} - \boldsymbol{\eta})'\boldsymbol{\Gamma}^{-1}(\mathbf{h} - \boldsymbol{\eta}) + g$, where $\boldsymbol{\Gamma}^{-1} = \mathbf{W}'\boldsymbol{\Sigma}^{-1}\mathbf{W}$, $\boldsymbol{\eta} = \boldsymbol{\Gamma}\mathbf{W}'\boldsymbol{\Sigma}^{-1}vec(\boldsymbol{\mu})$ and $g =$

$vec(\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}vec(\boldsymbol{\mu}) - \boldsymbol{\eta}'\boldsymbol{\Gamma}^{-1}\boldsymbol{\eta}$, the joint distribution can be written as

$$f\left(\mathbf{u}, \mathbf{h}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \;=\; \frac{exp(-g/2)}{(2\pi)^{(K-1)T}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} exp\left\{-\frac{(\mathbf{h}-\boldsymbol{\eta})'\boldsymbol{\Gamma}^{-1}(\mathbf{h}-\boldsymbol{\eta})}{2}\right\}\prod_{t=1}^{T}||\mathbf{h}_t||^{2(K-2)}. \qquad (1)$$

2.1 The Dryden-Mardia shape density

In this section we briefly cover some of the key results due to Mardia and Dryden (1989) and Dryden and Mardia (1991) who derived the shape space density induced by the Gaussian distribution for an iid sample of configurations. The Dryden-Mardia shape density, also known as offset-normal shape distrbution, can be obtained by first setting $T = 1$ in equation (1) and then considering the eigen-decomposition of the covariance matrix $\boldsymbol{\Gamma}$. Following Mardia and Dryden (1989), Dryden and Mardia (1991), it can be shown that the marginal density function of $\mathbf{u}$ can be found by integrating out the scale and rotation parameters, so that

$$f\left(\mathbf{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \;=\; \frac{|\boldsymbol{\Gamma}|^{\frac{1}{2}}exp(-g/2)}{(2\pi)^{K-2}|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\sum_{j=0}^{K-2}\binom{K-2}{j}E[l_1^{2j}|\zeta_1, \sigma_1]E[l_2^{2(K-2-j)}|\zeta_2, \sigma_2], \qquad (2)$$

where $E[\cdot|\zeta, \sigma]$ denotes the moments of the univariate Gaussian distribution with parameters (ζ, σ). Although the Dryden-Mardia density appears complicated, the evaluation is relatively easy as these expectations can be computed through the use of generalized Laguerre polynomials (see, Dryden and Mardia, 1991, and Section 2.3 below). For applications of the Dryden-Mardia distribution in the literature see, for example, Dryden and Mardia (2016, pg. 43), Bookstein (2014); Kume and Welling (2010) and Stuart and Ord (1994), Lele and Richtsmeier (1991) and Kendall (1991).

2.2 The offset-normal shape distribution for temporally correlated shapes

For pairs of correlated configurations, Mardia and Walder (1994) have shown that the density function in equation (2) transforms in a rather complicated form and that extending their results to a larger number of correlated configurations (i.e. $T > 2$) is a difficult task. Essentially, this is due to the difficulty of integrating out the dependence of $\mathbf{X}$ on $\mathbf{h}_t$ which, as shown below, appears as the product of T norms

$$f\left(\mathbf{u}|\boldsymbol{\mu},\boldsymbol{\Sigma}\right) \;=\; \frac{exp(-g/2)|\boldsymbol{\Gamma}|^{\frac{1}{2}}}{(2\pi)^{T(K-2)}|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\int\prod_{t=1}^{T}\|\mathbf{h}_t\|^{2(K-2)}f_{\mathcal{N}_{2T}}(\mathbf{h}|\boldsymbol{\eta},\boldsymbol{\Gamma})d\mathbf{h}. \tag{3}$$

By integrating out $\mathbf{h}$, the integral above can be rewritten as $E\left[\prod_{t=1}^{T}(\mathbf{h}'\mathbf{A}_t\mathbf{h})^{(K-2)}\right]$, where, for $\mathbf{0}_t$ a $(t\times t)$ null matrix, $\mathbf{A}_t = diag(\mathbf{0}_{2t-2},\mathbf{I}_2,\mathbf{0}_{2T-2t})$. Hence, evaluating the density (i.e., the off-set normal shape distribution) involves the computation of the moments of a product of quadratic forms in the (noncentral) normal random variable $\mathbf{h}\sim\mathcal{N}_{2T}(\boldsymbol{\eta},\boldsymbol{\Gamma})$.

2.3 Evaluating the shape density

By following Kan (2008), it can be shown that these moments can be computed through the following expansion

$$E\left[\prod_{t=1}^{T}(\mathbf{h}'\mathbf{A}_t\mathbf{h})^{K-2}\right] = \frac{1}{s!}\sum_{v_1=0}^{s_1}\cdots\sum_{v_T=0}^{s_T}(-1)^{\sum_{t=1}^{T}v_t}\binom{s_1}{v_1}\cdots\binom{s_T}{v_T}Q_s(\mathbf{B}_v) \tag{4}$$

where $s_t = (K-2)$ for all t, $s = T(K-2)$, $Q_s(\mathbf{B}_v) = E\big[(\mathbf{h}'\mathbf{B}_v\mathbf{h})^s\big]$, and $\mathbf{B}_v = \sum_{t=1}^{T}\big[s_t/2 - v_t\big]\mathbf{A}_t$. Hence, the moments of a product of quadratic forms can be rewritten as a linear combination of the moments of simpler quadratic forms $Q_s(\mathbf{B}_v)$.

An expression for $Q_s(\mathbf{B}_v)$ which is computationally efficient, is based on the recursive relation between moments and cumulants (Mathai and Provost, 1992, Eq.3.2b.8) and is given by

$$E\left[(\mathbf{h}'\mathbf{B}_v\mathbf{h})^s\right] = s!2^s d_s(\mathbf{B}_v) \tag{5}$$

where $d_s(\mathbf{B}_v) = \frac{1}{2s}\sum_{j=1}^{s}\left[tr(\mathbf{B}_v\boldsymbol{\Gamma})^j + j\boldsymbol{\eta}'(\mathbf{B}_v\boldsymbol{\Gamma})^{j-1}\mathbf{B}_v\boldsymbol{\eta}\right]d_{s-1}(\mathbf{B}_v), \quad d_0(\mathbf{B}_v) = 1$. Although equation (5) does not provide an explicit expression for $Q_s(\mathbf{B}_v)$, it is easy to program.

Finally, we note that both Lemma 2 of Magnus (1986) and Theorem 3.2b.1 of Mathai and Provost (1992), offer alternative solutions for evaluating these moments. However, these solutions are both computationally expensive and are thus not useful in practice.

3 Model parameter estimation

In this section we briefly introduce the EM algorithm for ML parameter estimation and then discuss the maximization and the expectation steps involved by the procedure.

3.1 EM implementation for likelihood optimization

The Expectation-Maximization algorithm (Dempster et al., 1977) is a ML parameter estimation method where part of the data can be considered to be incomplete or "hidden". In the static case, parameter estimation of the Dryden-Mardia distribution through the EM algorithm was first proposed by Kume and Welling (2010), who have discussed the necessary adjustments needed for using this algorithm for shape regression, missing landmark data, and mixtures of offset-normal shape distributions. The approach involves working with the original distribution of the landmark coordinates but treating the rotation and scale as missing/hidden variables. Huang et al. (2015) have recently used the algorithm to consider a mixture of offset-normal shape factor analyzers (MOSFA) and Brombin et al. (2016) have further explored the use of the EM algorithm in a dynamic setting by discussing its limitations when Laguerre polynomials are used to evaluate the offset-normal shape distribution. The methodology has also been extended to 3D shape and size-and-shape analysis by Kume et al. (2017).

In this paper, by taking account of the temporal correlation, we consider an extension of the EM algorithm for the more general distribution given in equation (10). A technical result discussed in Section 4 is proposed to overcome the computational difficulties associated with the E-step procedure (see, Brombin et al., 2016). It will be shown that this result, associated with the recursive algorithm introduced in Section 3.4, enables the estimation of AR processes with no constraints in the number of K and T.

For an iid random sample of N temporal configurations, let $\mathcal{X} = \{\mathbf{X}^{(n)}\}_{n=1,\dots,N}$ denote the full data. Our target is to find the values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ which maximize the log-likelihood function $l(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\, \mathcal{U}) = \sum_{n=1}^{N} log f(\mathbf{u}^{(n)}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\mathcal{U} = \{\mathbf{u}^{(n)}\}_{n=1,\dots,N}$ denotes the observed (shape) data and $f(\mathbf{u}^{(n)}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the off-set normal distribution of shape variables, $\mathbf{u}^{(n)}$, as shown in equation (3).

Under the assumption of Gaussianity, it is easy to show that the EM procedure maximizes the conditional expected log-likelihood

$$\mathcal{Q}_{\boldsymbol{\mu}^{(r)},\boldsymbol{\Sigma}^{(r)}}(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{n=1}^{N} \int log\big(f_{\mathcal{N}}(\mathbf{X}^{(n)}|\boldsymbol{\mu},\boldsymbol{\Sigma})\big) dF(\mathbf{X}^{(n)}|\mathbf{u}^{(n)},\boldsymbol{\mu}^{(r)},\boldsymbol{\Sigma}^{(r)}), \tag{6}$$

where $f_{\mathcal{N}}(\mathbf{X}^{(n)}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ is the pdf of a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, and $dF(\mathbf{X}^{(n)}|\cdot)$ is the conditional distribution of $\mathbf{X}^{(n)}$ evaluated at the current parameters, $\boldsymbol{\mu}^{(r)}$ and $\boldsymbol{\Sigma}^{(r)}$, and its shape $\mathbf{u}^{(n)}$. The EM iteration then alternates between performing an expectation (E) step, which computes the expectation of the log-likelihood evaluated using the current estimate for the parameters, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood computed in the E step. How to evaluate these expectations for different specifications of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ will be shown in the following sections.

3.2 Estimating the mean

The mean of the process can be estimated by considering that, in the M-step, the maximum of $\mathcal{Q}_{\boldsymbol{\mu}^{(r)},\boldsymbol{\Sigma}^{(r)}}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ is obtained at

$$vec(\boldsymbol{\mu}^{(r+1)}) = \frac{1}{N}\sum_{n=1}^{N} \int vec(\mathbf{X}^{(n)}) dF(\mathbf{X}^{(n)}|\mathbf{u}^{(n)},\boldsymbol{\mu}^{(r)},\boldsymbol{\Sigma}^{(r)}). \tag{7}$$

When the description of the changes over time is needed, any regression function can be used. A widely used choice is represented by a polynomial function for which the trend is approximated by a polynomial of low degree capturing the large-scale temporal variability of the process. Assuming $\boldsymbol{\Sigma}_T$ as proportional to the identity matrix, this modelling approach was first proposed in Kume and Welling (2010). Here, we provide a more general formulation which takes care of the presence of the temporal correlation.

Suppose the mean of the process is parameterized by a polynomial function of order M, i.e. $\boldsymbol{\mu}_t^{\dagger} = E[\mathbf{X}_t^{\dagger}] = \sum_{m=0}^{M} \mathbf{B}_m^{\dagger} t^m$, with $\mathbf{B}_p^{\dagger} = \left(\boldsymbol{\beta}_m^{(x)^{\dagger}} \ \boldsymbol{\beta}_m^{(y)^{\dagger}}\right)$, and $\boldsymbol{\beta}_m^{(x)^{\dagger}}$ and $\boldsymbol{\beta}_m^{(y)^{\dagger}}$ K-dimensional vectors of regression coefficients. Accordingly, we write $vec(\mathbf{X}^{\dagger}) \sim \mathcal{N}_{2KT}(\mathbf{D}^{\dagger}\boldsymbol{\beta}^{\dagger}, \boldsymbol{\Sigma}^{\dagger})$, where $\boldsymbol{\beta}^{\dagger} = vec\left(\mathbf{B}_0^{\dagger} \ldots \mathbf{B}_M^{\dagger}\right)$ is a $2K(M+1)$-dimensional vector of regression coefficients and $\mathbf{D}^{\dagger} = (\mathbf{T} \otimes \mathbf{I}_{2K})$ is the $(2KT \times 2K(M+1))$ design matrix $(T > M+1)$ with $\mathbf{T}$ having elements $t_j^m, \ m = 0, \ldots, M$, at each row $j = 1, \ldots, T$.

In the preform space we also have $vec(\mathbf{X}) \sim \mathcal{N}_{2(K-1)T}(\mathbf{D}\boldsymbol{\beta}, \boldsymbol{\Sigma})$, with $\boldsymbol{\beta} = (\mathbf{I}_{2(P+1)} \otimes \mathbf{L})\boldsymbol{\beta}^{\dagger}$, $\mathbf{D} =$

$\mathbf{T} \otimes \mathbf{I}_{2(K-1)}$, and $\boldsymbol{\Sigma} = (\mathbf{I}_{2T} \otimes \mathbf{L})\boldsymbol{\Sigma}^{\dagger}(\mathbf{I}_{2T} \otimes \mathbf{L}')$. Hence, considering the ML estimator of the regression parameters for the complete data (see Appendix 2), it is easy to show that the update rule in the maximization step is given by

$$\boldsymbol{\beta}^{(r+1)} = \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{D}'\boldsymbol{\Omega}^{(r)^{-1}}\mathbf{D}\right)^{-1}\mathbf{D}'\boldsymbol{\Omega}^{(r)^{-1}}\int vec(\mathbf{X}^{(n)})dF\left(\mathbf{X}^{(n)}|\mathbf{u}^{(n)}, \mathbf{D}\boldsymbol{\beta}^{(r)}, \boldsymbol{\Sigma}^{(r)}\right), \qquad (8)$$

where $\boldsymbol{\Omega} = (\boldsymbol{\Sigma}_T^{(r)} \otimes \mathbf{I}_{2K-2})^{-1}$. The expectation step (E-step) is performed by finding the expectations and, given that $vec(\mathbf{X}) = \mathbf{W}\mathbf{h}$ and $dF(\mathbf{X}|\mathbf{u}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = f(\mathbf{h}, \mathbf{u}|\boldsymbol{\mu}, \boldsymbol{\Sigma})d\mathbf{h}/f(\mathbf{u}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we notice that both equations (7) and (8) require the evaluation of the following integral

$$\int vec(\mathbf{X})dF(\mathbf{X}|\mathbf{u}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathbf{W}\frac{\int \mathbf{h}f(\mathbf{h}, \mathbf{u}|\boldsymbol{\mu}, \boldsymbol{\Sigma})d\mathbf{h}}{f(\mathbf{u}|\boldsymbol{\mu}^{(r)}, \boldsymbol{\Sigma}^{(r)})} = \mathbf{W}\ \mathcal{Q}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{W}). \qquad (9)$$

3.3 Estimating the general covariance matrices $\boldsymbol{\Sigma}_S$ and $\boldsymbol{\Sigma}_T$

Due to the large number of parameters, numerical optimization of the full likelihood based on standard numerical routines is difficult, especially when working with full covariance matrices. Separability conditions on the covariance structures has been found useful to overcome most of the difficulties arising in the modelling of complex spatial-temporal dependency structures. In shape analysis, a major advantage of working with separable processes is that the covariance matrix can be decomposed into purely landmark and temporal components.

Assume that the $(2TK \times 2TK)$ covariance matrix in the configuration space can be expressed as $\boldsymbol{\Sigma}^{\dagger} = \boldsymbol{\Sigma}_T \otimes \boldsymbol{\Sigma}_S^{\dagger}$, where $\boldsymbol{\Sigma}_T$ is a $(T \times T)$ covariance matrix between temporal observations and $\boldsymbol{\Sigma}_S^{\dagger}$ is a $(2K \times 2K)$ covariance matrix between landmark coordinates. Under separability conditions, the covariance matrix in the space of preform coordinates can be written as $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_T \otimes \boldsymbol{\Sigma}_S$, where the landmark covariance is given by $\boldsymbol{\Sigma}_S = (\mathbf{I}_2 \otimes \mathbf{L})\boldsymbol{\Sigma}_S^{\dagger}(\mathbf{I}_2 \otimes \mathbf{L}')$. Hence, the induced shape distribution can now be written as

$$f\left(\mathbf{u}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_T \otimes \boldsymbol{\Sigma}_S\right) = \frac{exp(-g/2)|\boldsymbol{\Gamma}|^{\frac{1}{2}}}{(2\pi)^{(K-2)T}|\boldsymbol{\Sigma}_T|^{(K-1)}|\boldsymbol{\Sigma}_S|^{\frac{T}{2}}}E\left[\prod_{t=1}^{T}(\mathbf{h}'\mathbf{A}_t\mathbf{h})^{(K-2)}\right], \qquad (10)$$

where the covariance matrix of the scale and rotation parameters is $\mathbf{\Gamma} = \left(\mathbf{W}'(\mathbf{\Sigma}_T^{-1} \otimes \mathbf{\Sigma}_S^{-1})\mathbf{W}\right)^{-1}$.

The separability assumption has several advantages, including rapid fitting and simple extensions of standard techniques developed in time series and classical spatial statistics (Genton, 2007). In some applications, in fact, the covariance matrices can be assumed to have certain structures and imposing these structures in the estimation typically leads to improved accuracy and robustness (e.g., to small sample effects).

Given the ML estimators $\hat{\mathbf{\Sigma}}_S$ and $\hat{\mathbf{\Sigma}}_T$ (see Appendix 2), the update rules for the covariance matrices in the M-step are defined as

$$\begin{aligned}\mathbf{\Sigma}_S^{(r+1)} =& \frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T}\check{\mathbf{P}}_t^{(r)}\int vec(\mathbf{X}^{(n)})vec(\mathbf{X}^{(n)})'dF\left(\mathbf{X}^{(n)}|\mathbf{u}^{(n)},\boldsymbol{\mu}^{(r)},\mathbf{\Sigma}_T^{(r)}\otimes\mathbf{\Sigma}_S^{(r)}\right)\check{\mathbf{P}}_t^{(r)'} \\ & -\frac{1}{T}\boldsymbol{\mu}^{(r+1)}\mathbf{\Sigma}_T^{-1(r)}\boldsymbol{\mu}^{(r+1)'} \end{aligned} \tag{11}$$

$$\begin{aligned}\mathbf{\Sigma}_T^{(r+1)} =& \frac{1}{N(2K-2)}\sum_{n=1}^{N}\sum_{k=1}^{2K-2}\check{\mathbf{P}}_k^{(r)}\int vec(\mathbf{X}^{(n)})vec(\mathbf{X}^{(n)})'dF\left(\mathbf{X}^{(n)}|\mathbf{u}^{(n)},\boldsymbol{\mu}^{(r)},\mathbf{\Sigma}_T^{(r)}\otimes\mathbf{\Sigma}_S^{(r)}\right)\check{\mathbf{P}}_k^{(r)'}+ \\ & -\frac{1}{2K-2}\boldsymbol{\mu}^{(r+1)'}\mathbf{\Sigma}_S^{-1(r)}\boldsymbol{\mu}^{(r+1)} \end{aligned} \tag{12}$$

with $\check{\mathbf{P}}_k^{(r)} = \mathbf{I}_T \otimes (\mathbf{L}_S^{(r)}\boldsymbol{e}_k)'$ and $\mathbf{\Sigma}_S^{-1(r)} = \mathbf{L}_S^{(r)}\mathbf{L}_S^{(r)'}$, with $\check{\mathbf{P}}_t^{(r)} = (\mathbf{L}_T^{(r)}\boldsymbol{e}_t)' \otimes \mathbf{I}_{2K-2}$, $\mathbf{\Sigma}_T^{-1(r)} = \mathbf{L}_T^{(r)}\mathbf{L}_T^{(r)'}$. Here, $\mathbf{L}_T$ and $\mathbf{L}_S$ are lower triangular matrices and $\boldsymbol{e}_t$ and $\boldsymbol{e}_k$ are, respectively, T-dimensional and $(2K-2)$-dimensional vectors with entries, $e_j(j) = 1$ for $j = t, k$, and zero otherwise. Note that without restrictions, $\mathbf{\Sigma}_S$ and $\mathbf{\Sigma}_T$ are not identifiable. However, this non-identifiability problem can be addressed, for example, by considering $\mathbf{\Sigma}_T$ as a correlation matrix, and not as covariance matrix.

The ML estimates of the Kronecker factor matrices are thus obtained through the cyclic optimization scheme of the EM where, in the E-step, the expected values of the complete data sufficient statistics in equations (11) and (12) require the computation of the following integral

$$\int vec(\mathbf{X})vec(\mathbf{X})'dF(\mathbf{X}|\mathbf{u},\boldsymbol{\mu},\mathbf{\Sigma}) = \mathbf{W}\frac{\int \mathbf{h}\mathbf{h}'f(\mathbf{h},\mathbf{u}|\boldsymbol{\mu},\mathbf{\Sigma})d\mathbf{h}}{f(\mathbf{u}|\boldsymbol{\mu}^{(r)},\mathbf{\Sigma}^{(r)})}\mathbf{W}'. \tag{13}$$

We finally note that, although the model can be developed in terms of a general covariance matrix, a useful assumption for practical applications would be to consider the observed configurations as following a *proper* complex Gaussian distribution (Neeser and Massey, 1993) of which, both the isotropic and the cyclic Markov structures are special cases.

3.4 Autoregressive models in configuration space

Assume that, in the preform space, a generic configuration follows the AR(p) model

$$\mathbf{\Phi}(\mathbf{B})\big(vec(\mathbf{X}_t) - vec(\boldsymbol{\mu}_t)\big) = vec(\mathbf{E}_t)$$

where $\mathbf{B}$ is the usual backward shift operator, $\mathbf{\Phi}(\mathbf{B})$ is the matrix polynomial $\mathbf{\Phi}(\mathbf{B}) = \mathbf{I}_{2(K-1)} - \mathbf{\Phi}_1\mathbf{B} - \mathbf{\Phi}_2\mathbf{B}^2 - \ldots - \mathbf{\Phi}_p\mathbf{B}^p$ and $vec(\mathbf{E}_t) \sim \mathcal{N}_{2(K-1)}(\mathbf{0}, \mathbf{\Sigma}_S)$. Separable structures can be obtained from this autoregressive representation by imposing appropriate parameter constraints. These follow from the assumption that the matrix polynomial, $\mathbf{\Phi}(\mathbf{B})$, reduces to a scalar polynomial $\phi(\mathbf{B})$, implying that $\mathbf{\Phi}(\mathbf{B}) = \phi(\mathbf{B})\mathbf{I}_{2(K-1)}$ or, equivalently, that $\mathbf{\Phi}_i\mathbf{B}^i = \phi_i\mathbf{I}_{2(K-1)},\ i = 0, 1, \ldots, p$, with $\phi_0 = 1$.

Consider an AR(p) process with constant mean $vec(\boldsymbol{\mu}_t) = \tilde{\boldsymbol{\mu}}$ for all t and landmark covariance matrix $\mathbf{\Sigma}_S = \sigma^2(\mathbf{I}_2 \otimes \mathbf{L})(\mathbf{I}_2 \otimes \mathbf{L}')$. The model, thus implies that

$$vec(\mathbf{X}_t) = \tilde{\boldsymbol{\mu}}(1 - \phi_1 - \cdots - \phi_p) + \phi_1 vec(\mathbf{X}_{t-1}) + \cdots + \phi_p vec(\mathbf{X}_{t-p}) + vec(\mathbf{E}_t), \tag{14}$$

from which it follows that the conditional distributions are Normal, i.e. $\mathcal{N}_{2(K-1)}\left(\tilde{\boldsymbol{\mu}}_{t|t-p}, \mathbf{\Sigma}_S\right)$, with $\tilde{\boldsymbol{\mu}}_{t|t-p} = \tilde{\boldsymbol{\mu}}(1 - \phi_1 - \cdots - \phi_p) + \phi_1 vec(\mathbf{X}_{t-1}) + \cdots + \phi_p vec(\mathbf{X}_{t-p})$. Then, the following approximation allows for a fast evaluation of the conditional means as well as of the likelihood for autoregressive models.

Result 1. Let the shape coordinates be generated from an AR(p) process and let $vec(\mathbf{X}_0)$ be a given initial configuration. Then, the required expectation of equation (9), *i.e.* $\int vec(\mathbf{X})dF(\mathbf{X}|\mathbf{u}, \boldsymbol{\mu}, \mathbf{\Sigma})$, can be approximated recursively as follows

$$E\big[vec(\mathbf{X}_t)|\mathbf{W}_0, \mathbf{W}_1, \ldots, \mathbf{W}_t\big] \simeq \mathbf{W}_t\ \mathcal{Q}(\tilde{\boldsymbol{\mu}}_{t|t-p}, \mathbf{\Sigma}_S, \mathbf{W}_t), \quad t = 1, \ldots, T.$$

Proof. See Appendix 3.

This result simplifies significantly the computation of the expectation in equation (9) as it only requires the recursive evaluation of T simple conditional expectations for $\mathbf{h}_t$. It is easy to show that the estimate of the marginal mean, $\tilde{\boldsymbol{\mu}}$, involves the computation of the expectation given by equation (9). Then, conditional on $\tilde{\boldsymbol{\mu}}$, the estimate of the autoregressive parameters, $\phi_1, \dots, \phi_p$, can be found by numerical maximization of the log-likelihood. This is a simple optimization problem which can be carried out efficiently through the recursive formula and ensures to check numerically for allowable model parameters. Extending the model to have a non-constant polynomial mean and/or a parameterized landmark covariance matrix, *e.g.* a cyclic Markov structure (Dryden and Mardia, 2016), is also straightforward.

3.5 Baseline invariance

As shown in Section 2, the shape variables $\mathbf{u}$ are calculated after choosing, at each time t, the first two (noncoincident) landmarks as the baseline. This choice, however, does not have to be fixed for each shape observation since, as long as we appropriately rotate the mean and the covariance matrix, the probability distribution turns out to be rotationally invariant (Kume and Welling, 2010). For a more detailed discussion on baseline invariance, see Appendix 4.

4 Computational issues

In this section we discuss some of the problems related to the computation of the expectations

$$\int \mathbf{h} f(\mathbf{h}, \mathbf{u}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{h} = \frac{exp(-g/2)|\boldsymbol{\Gamma}|^{\frac{1}{2}}}{(2\pi)^{(K-2)T}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} E\left[\mathbf{h} \prod_{t=1}^{T} (\mathbf{h}'\mathbf{A}_t\mathbf{h})^{(K-2)}\right] \quad (15)$$

and

$$\int \mathbf{h}\mathbf{h}' f(\mathbf{h}, \mathbf{u}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{h} = \frac{exp(-g/2)|\boldsymbol{\Gamma}|^{\frac{1}{2}}}{(2\pi)^{(K-2)T}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} E\left[\mathbf{h}\mathbf{h}' \prod_{t=1}^{T} (\mathbf{h}'\mathbf{A}_t\mathbf{h})^{(K-2)}\right] \quad (16)$$

involved by equations (9) and (13), respectively. These expectations do not appear exactly in the form of equation (4) and their evaluation is thus difficult to compute analytically.

In the following, a technical result is introduced to overcome the problem. This makes use of an auxiliary Gaussian random variable with mean and variance equal to one, and shows that the expectations in equations (15) and (16) can be rewritten in the form of equation (4) for which results from Section 2 are available.

Result 2. *Let w be an auxiliary Gaussian variable such that $\mathbf{h}_a = (\mathbf{h}', w)' \sim \mathcal{N}_{2T+1}(\boldsymbol{\eta}_a, \boldsymbol{\Gamma}_a)$, with $\boldsymbol{\eta}_a = (\boldsymbol{\eta}'\ 1)'$ and $\boldsymbol{\Gamma}_a = \begin{pmatrix} \boldsymbol{\Gamma} & \mathbf{0} \\ \mathbf{0} & 1 \end{pmatrix}$. Also, let $f_{\mathcal{N}_{2T+1}}(\mathbf{h}_a|\boldsymbol{\eta}_a, \boldsymbol{\Gamma}_a) = f_{\mathcal{N}_{2T}}(\mathbf{h}|\boldsymbol{\eta}, \boldsymbol{\Gamma}) f_{\mathcal{N}}(w|1,1)$. Then, the component-wisely solution of equation (15) is given by*

$$\int h_j \prod_{t=1}^{T} (\mathbf{h}'\mathbf{A}_t\mathbf{h})^{K-2} f_{\mathcal{N}_{2T}}(\mathbf{h}|\boldsymbol{\eta}, \boldsymbol{\Gamma}) d\mathbf{h} = \frac{1}{2}\mathcal{I}_{1_j} + \mathcal{I}_2 - \frac{1}{2}\mathcal{I}_{3_j} \tag{17}$$

where $\mathcal{I}_{1_j} = \int \prod_{t=0}^{T} (\mathbf{h}'\mathbf{A}_t\mathbf{h})^{s_t} f_{\mathcal{N}_{2T}}(\mathbf{h}|\boldsymbol{\eta}, \boldsymbol{\Gamma}) d\mathbf{h}$, $\mathcal{I}_2 = \int \prod_{t=1}^{T} (\mathbf{h}'\mathbf{A}_t\mathbf{h})^{K-2} f_{\mathcal{N}_{2T}}(\mathbf{h}|\boldsymbol{\eta}, \boldsymbol{\Gamma}) d\mathbf{h}$ and, by setting $\tilde{\mathbf{h}}_j = (h_j - w, \mathbf{h}')'$ and $\tilde{\mathbf{A}}_0 = diag(1, \mathbf{0}_{2T})$, $\mathcal{I}_{3_j} = \int \prod_{t=0}^{T} (\tilde{\mathbf{h}}_j'\tilde{\mathbf{A}}_t\tilde{\mathbf{h}}_j)^{s_t} f_{\mathcal{N}_{2T+1}}(\tilde{\mathbf{h}}_j|\tilde{\boldsymbol{\eta}}_j, \tilde{\boldsymbol{\Gamma}}_j) d\tilde{\mathbf{h}}_j$.

Proof. See Appendix 5.

Each single term, $\mathcal{I}_{1_j}$, $\mathcal{I}_2$ and $\mathcal{I}_{3_j}$, can be written as the expectation of the product of quadratic forms which, in turn, can thus be computed by using equations (4) and (5).

The elements of $E\left[\mathbf{h}\mathbf{h}' \prod_{t=1}^{T} (\mathbf{h}'\mathbf{A}_t\mathbf{h})^{(K-2)}\right]$ can be computed as the expectations of product of quadratic forms. The variance components, in fact, are given by $\mathcal{I}_{1_j}$ while the covariance elements, are obtained as

$$\int h_j h_l \prod_{t=1}^{T} (\mathbf{h}'\mathbf{A}_t\mathbf{h})^{K-2} f_{\mathcal{N}_{2T}}(\mathbf{h}|\boldsymbol{\eta}, \boldsymbol{\Gamma}) d\mathbf{h} = \int \prod_{t=0}^{T} (\mathbf{h}'\mathbf{A}_t\mathbf{h})^{s_t} f_{\mathcal{N}_{2T}}(\mathbf{h}|\boldsymbol{\eta}, \boldsymbol{\Gamma}) d\mathbf{h}$$

where $s_0 = 1$, $s_t = K - 2$ for $t = 1, \ldots, T$ and $\mathbf{A}_0$ is a selection matrix with $a_{j,l} = a_{l,j} = 1/2$ and 0 elsewhere.

The computation of the expectations based on the use of Laguerre polynomials is discussed in Brombin et al. (2016). Unfortunately, this is a far less efficient procedure which does not appear useful in practice.

5 Applications

In what follows, we apply the offset Normal shape distribution to three real data examples, with the first two concerning medical imaging studies and the third considering an application in the field of social signal processing (Pantic et al., 2011).

5.1 Landmark shape analysis of Corpus Callosum

Morphologic assessment of brain structures through landmark-based shape analysis has been popular in neuroanatomical research because of its convenience and effectiveness in obtaining shape information. In this example we focus on the Corpus Callosum (CC) which is the major fiber bundle connecting the two hemispheres of the brain. Changes to its shape or structure is the subject of active studies. There is, in fact, an interest in linking the physical changes with neurological impairment and other pathologies such as multiple sclerosis (MS), autism, schizophrenia and Alzheimer.

We have used the Open Access Series of Imaging Studies (OASIS) longitudinal database (www.oasis-brains.org) to investigate shape changes of the CC for two groups of 34 non-demented (nd) and 15 demented (d) people. The subjects have similar age (average 76.5 and 77.2, for nd and d respectively), are all right-handed and include 23 men and 26 women. For each subject, we used MRI images available from three visits (i.e. $T = 3$). A midsagittal slice was extracted from each volumetric image, and the CC contours were then extracted by automatic segmentation for each subject. Finally, nine anatomical landmarks were identified as described in He et al. (2010) on the boundary of each CC. Most of these landmarks refer to extreme points or terminals and maxima of curvature and can be identified mathematically with good accuray. A pictorial representation of a typical midsagittal plane image and the chosen set of landmarks is given in Figure 2.

In order to identify possible shape differences of the CC, the generalized likelihood ratio test (GLRT) is applied to test whether the mean paths of demented and non-demented groups differ from each other

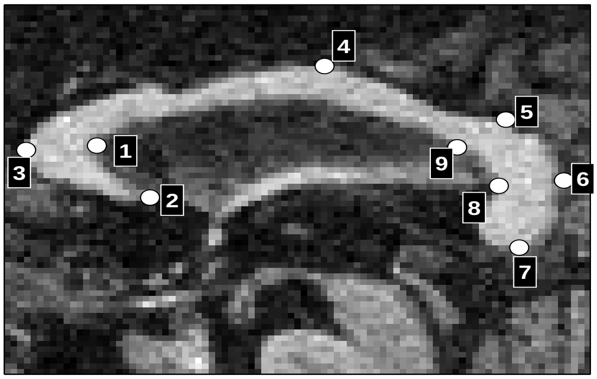


Figure 2: Midsagittal plane image and landmarks of the CC: 1) *Interior angle of genu*, 2) *Tip of genu*, 3) *Anterior most of CC*, 4) *Topmost of CC*, 5) *Splenium topmost point*, 6) *Posterior most of CC*, 7) *Bottommost of splenium*, 8) *Interior notch of splenium*, 9) *CC-fornix junction*

only by some rotation. We thus consider the hypothesis test $H_0 : \tilde{\boldsymbol{\mu}}_d = \tilde{\boldsymbol{\mu}}_{nd}$ mod(rot), $\boldsymbol{\Sigma}_d = \boldsymbol{\Sigma}_{nd}$, versus $H_0 : \tilde{\boldsymbol{\mu}}_d \neq \tilde{\boldsymbol{\mu}}_{nd}$ mod(rot), $\boldsymbol{\Sigma}_d = \boldsymbol{\Sigma}_{nd}$. Because of the limited number of visits and the small number of landmarks, the means are estimated by using equation (7) and the covariance matrices by using equations (11) and (12). To speed up the computation, shape coordinates are obtained using the complex representation of planar points. The GLRT statistic (27.8 on 30 df) first suggests that a model with a constant mean (in time) and general complex covariance matrix $\boldsymbol{\Sigma}$ is preferable to the full model with time varying mean estimated by using equation (7), and same covariance structure. No further model simplification, for example associated with the specification of restricted models with either $\boldsymbol{\Sigma}_S = \mathbf{I}$ or $\boldsymbol{\Sigma}_T = \mathbf{I}$, seems to be possible. The covariance matrices, $\hat{\boldsymbol{\Sigma}}_S$ and $\hat{\boldsymbol{\Sigma}}_T$, obtained through equations (11) and (12), are shown in Appendix 7. Clearly, the general covariance case allows to represent the second order non-stationary features of the process.

To test the mean shape differences between the two groups, the selected model is first run for the pooled sample, giving a log-likelihood value of 7315.4. The likelihood values for the alternative hypothesis, equal to 5113.9 and 2207.7 for non-demented and demented groups, can be obtained by running the EM separately for each group, while keeping the entries of $\boldsymbol{\Sigma}_d = \boldsymbol{\Sigma}_{nd}$ the same as those of shown in Appendix 7. Since the p-value for the GLRT statistic (12.44 on 15 df) is 0.65, there is strong evidence that, modulo rotations, there are no differences in the mean shape paths of the two groups considered in

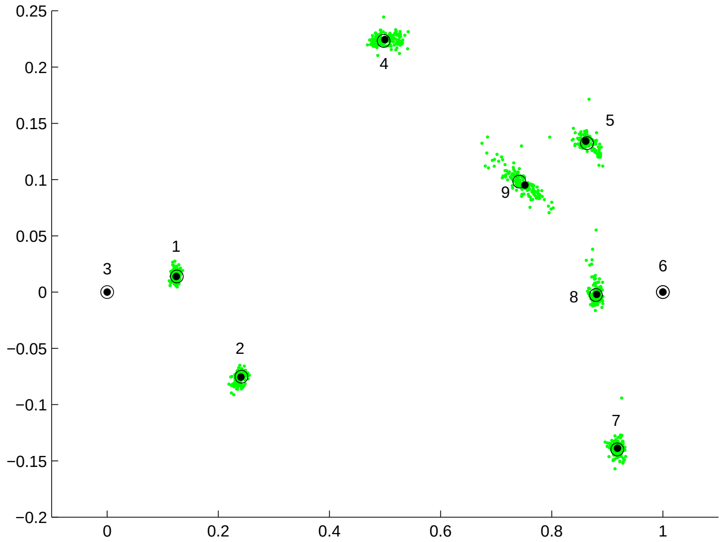


Figure 3: CC in Bookstein shape coordinates with baseline defined by points (0,0) and (1,0). The green dots represent the observed configuration landmarks while the black dots and circles represent the landmarks of the estimated means for demented and non-demented groups

this study. This is clearly shown in Figure 3 where the landmarks of the two mean shapes are very close to each other. The representation is in Bookstein shape coordinates with baseline given by landmarks 3 and 6.

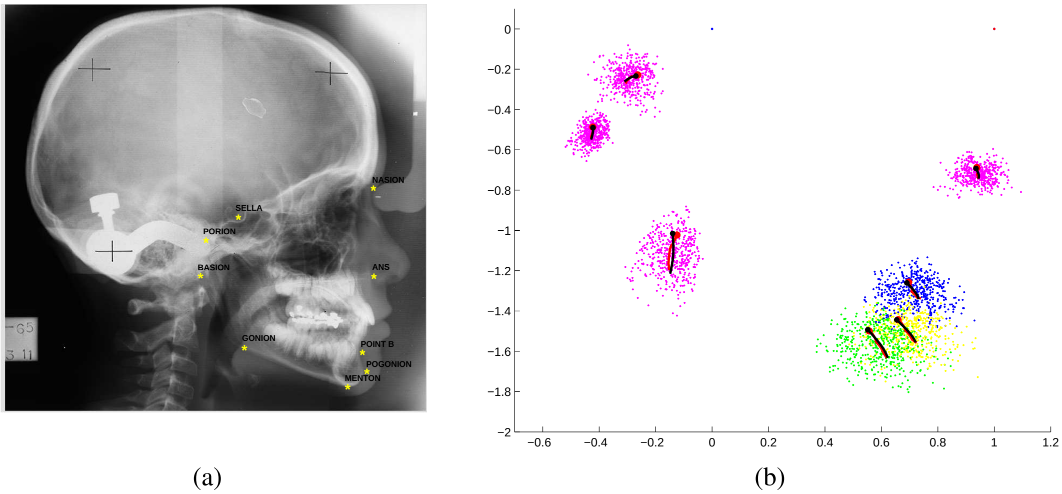
5.2 Shape changes in the craniofacial complex

Sagittal malocclusions are highly prevalent and have functional, esthetic, and social implications that make them a public health issue. Precise descriptions of how and when these abnormalities emerge and change during childhood and adolescence can inform our understanding of their underlying of their underlying biology and facilitate diagnosis from craniofacial shape. Many studies have described growth extensively (see, for example, Sridhar (2011) and references therein), while the shape changes of the craniofacial complex have been less investigated. The dynamics of the cranial and facial structures, in fact, are of a very intricate nature and quantifying their variation, and detecting the locations where the shape change is most active during different stages of development, still represent challenging problems for the orthodontic treatment planning.

The data for this study were obtained from the American Association of Orthodontists Foundation (AAOF) craniofacial legacy growth collection files. The data are available on the internet and refer to a digital repository of records from 9 craniofacial growth study collections in the United States and Canada. Important details on the use of materials from the AAOF Craniofacial Growth Legacy Collection in the orthodontic literature can be found in (Al-Jewair et al., 2018). Here, lateral cephalograms at 10 different maturational stages in cervical vertebrae are used for the analysis. These stages, roughly correspond to the chronological age classes of $[7, 8], [8, 9], \dots, [16, 17]$ years. The data refer to a sample of 47 subjects presenting normal occlusion (Class I molar and canine relationships, normal overbite and overjet), with no vertical or sagittal skeletal discrepancies and with a well-balanced facial profile. The part of the craniofacial complex considered here is represented by a set of 9 anatomical landmarks and, as a representative example, Figure 4(a) shows a typical cephalograph image with the landmark configuration. Starting from the left, we find the *posterior border of the ramus* which is straight and continuous with the *inferior border of the body* of the bone. At its junction with the posterior border is the *angle of the mandible* which is important for the attachment of the Masseter and the Pterygoideus muscles. Moving toward the right, we then find the *menton* which is the most inferior part of the mandibular symphysis. This ridge, which sometimes presents a centrally depressed area, represents the median line of the external surface of the mandible. Note that the *Sella* (the center of the hypophyseal fossa) and the *Nasion* (the junction of the nasal and frontal bones at the most posterior point on the curvature of the bridge of the nose) act as reference landmarks and will be considered as baseline for Bookstein registration.

Traditionally, craniofacial analysis has relied on simple distances and angles between anatomical landmarks (Oyonarte et al., 2016), which give only a limited representation of the phenomenon under study (Moyers and Bookstein, 1979; Franchi et al., 2001). Here to describe the dynamics of the craniofacial complex, dynamic shape regression models are fitted to the data. Furthermore, since forecasting the shape changes is an important part of the analysis, the last age is excluded from the estimation procedure and is used only for comparative purposes to test the predictive performance of the model.

For a model with $\mathcal{P}$ parameters, model selection is based on the AIC $= -2l(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathcal{U}) + 2\,\mathcal{P}$ statistic. Results from the analysis suggest that the best AIC value (-21825.24) among the fitted models is obtained



(a)

(b)

Figure 4: a) Cephalometric landmarks (yellow points) superimposed on the cephalogram of a subject at the first maturational level. The two anatomical landmarks of *sella* and *nasion* are used as baseline for Bookstein registration. b) The Bookstein shape coordinates with baseline defined by points (0,0) and (1,0) and the fitted paths are from the second order shape regression model with AR(1) errors (red line) and the functional model (black line). Different colors are used to differentiate the patterns of Point B, Pogonion and Menton.

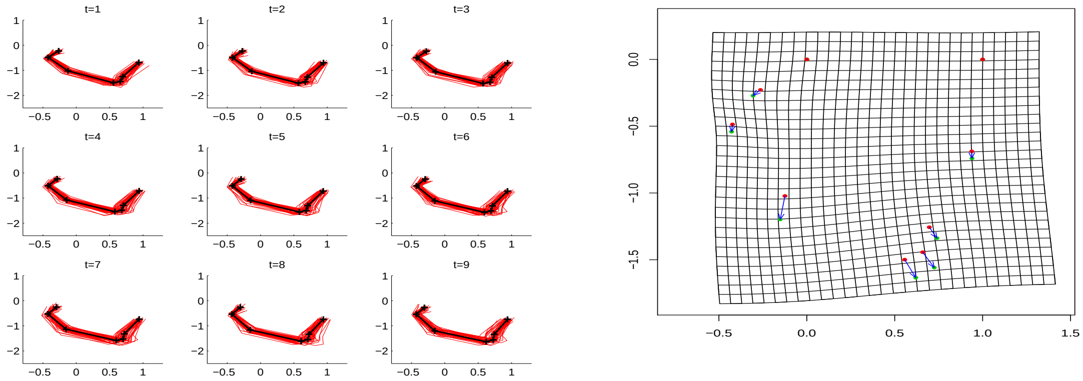
for the second order polynomial shape regression with AR(1) errors ($\hat{\phi}_1 = 0.39$) and isotropic $\mathbf{\Sigma}_S$. For the EM algorithm, the starting value for $\boldsymbol{\beta}$ are taken such that $\mathbf{B}_0$ is the configuration at time $t = 1$ while $\mathbf{B}_m,\ m > 0$, are chosen at random. For nested models, previous estimates are used as starting values for the fit of subsequent models. This procedure usually facilitates the convergence of the estimation process.

Results of the estimation procedure are shown in Figure 4(b). For comparison purposes, we also consider the fit of a functional model based on the construction, and combination, of roughness penalties on functions in space and time. Full details on the model are given in Kent et al. (2001) and we refer to them for known results. The figure shows the fit of this model (in red) using the “special” parametrization (see, Kent et al., 2001), which captures the curvature in the paths trough a term which is second order in time and linear in space. As it can be noticed, both models are able to provide a good description of the dynamics of the shape data and the fits are very similar. However, one drawback of the functional model is that the associated model parameters are not easy to interpret and, in general, it is not designed to produce $k-$step ahead forecasts of the shape coordinates.

The same Figure 4(b) clearly suggests that the shape changes are localized and that these mainly occur at the inferior border of the body, at the angle as well as at the mandibular symphysis. The analysis of overall morphologic changes can also be completed by means of a transformation grid (Dryden and Mardia, 2016). For example, at each time point, the subplots in Figure 5(a) show both the estimated conditional means and the Bookstein shape coordinates for the 47 subjects. The grid is thus obtained by using a pair of thin-plate splines which map specific points of the estimated conditional mean at time t (i.e., $\tilde{\boldsymbol{\mu}}_{t|t-p}$) onto homologous points of the target configuration at time $t' > t$ (i.e., $\tilde{\boldsymbol{\mu}}_{t'|t'-p}$). The idea is that each point of the mean configuration gives a local indication of shape change, and that one can use the whole set of points of the estimated conditional means for finding global and local shape changes. Considering for example the first and the ninth maturational stages, Figure 5(b) reveals a closure of the angle associated with an upward-forward direction of the shape changes at the *posterior border of the ramus* and with an upward direction of the shape changes at the symphysis.

We conclude the analysis by showing forecasting results in Figure 6 where the one-step ahead forecast is compared with the Bookstein shape coordinates at the tenth stage. Specifically, the upper panels show the best (left) and worse (right) predictions corresponding to the minimum (0.039) and maximum (0.149) values of the distribution of the root mean squared prediction errors obtained for the 47 subjects. The forecast corresponding to the $75°$ percentile (0.092) is also represented in the middle panel. The bottom panels emphasize the magnitude of the prediction error at each landmark, with the length of each arrow representing the distance between predicted and true coordinate shapes. In general, results suggest that good shape predictions at the last age can be obtained for most of the analysed subjects.

Note that a similar analysis can also been found in Appendix 9 where we describe the shape changes of eight biological landmarks of the skulls of 18 different rats observed at 8 different ages. The data have been analysed for several purposes and a description can be found, for example, in Kent et al. (2001), Kume and Welling (2010) and Kenobi et al. (2010). It is shown that our model improves the fit obtained by Kume and Welling (2010) who already improved the fit provided by Le and Kume (2000).



(a)

(b)

Figure 5: a) Craniofacial configurations in Bookstein shape coordinates with baseline defined by points (0,0) and (1,0). At each time, the subplots shows the 47 subjects shape coordinates with the fit provided by the estimated conditional mean (tick line) of a second order polynomial model with AR(1) errors. b) Thin-plate spline transformation grid between $\tilde{\boldsymbol{\mu}}_{t=1|\cdot}$ and $\tilde{\boldsymbol{\mu}}_{t=9|\cdot}$.

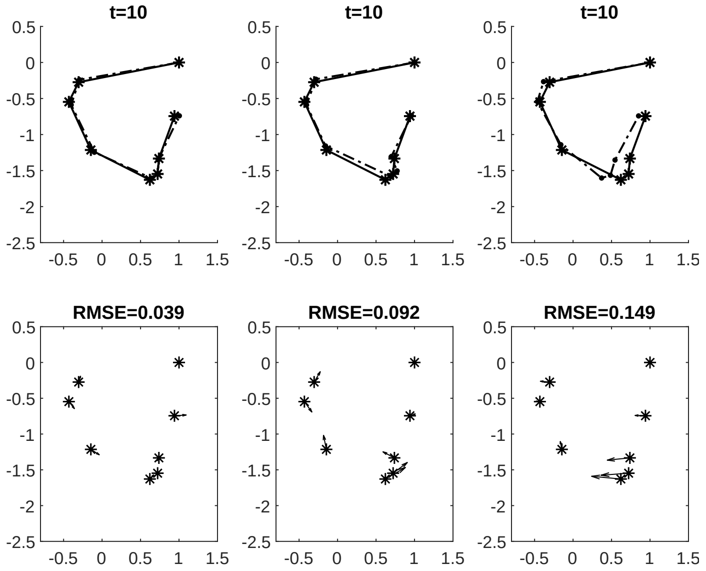
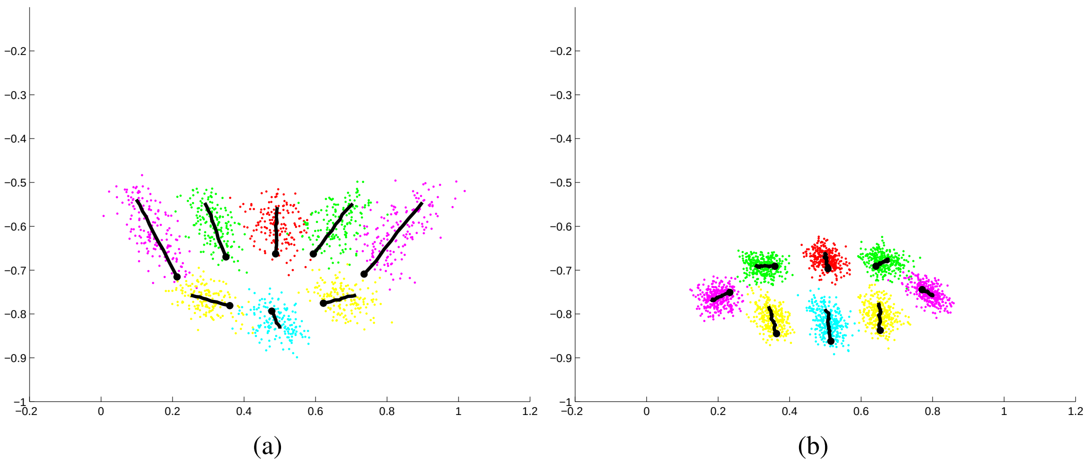


Figure 6: Forecasts (continuous line) of shape coordinates at the tenth maturational level. The configurations are represented using Bookstein shape coordinates with baseline defined by points (0,0) and (1,0). The upper panels show the best (left) and worse (right) predictions corresponding to the minimum (0.039) and maximum (0.149) root mean squared prediction errors computed among the 47 subjects. The forecast corresponding to the $75°$ percentile (0.092) is also represented in the middle panel. The bottom panels emphasize the magnitude of the prediction error at each landmark, with the length of each arrow representing the distance between predicted and true coordinate shapes.

5.3 Featuring facial expressions

One of the non-verbal communication method by which one understands the mood/mental state of a person is the expression of face. Hence, due to the important role of facial expressions in human interaction, the ability to perform Facial Expression Recognition (FER) automatically enables a range of novel applications in fields such as human-computer interaction and data analytics (see, for example, Wang et al., 2018 and references therein). From a physiological perspective, facial expressions result from the deformations of some facial features caused by an emotion. Since each emotion corresponds to a typical stimulation of the face muscles, the aim of this section is to evaluate the possibility of using our shape regression models to recognize basic emotions by only considering the deformations of some facial permanent features such as eyes, eyebrows and mouth. It is assumed that the mean shape estimated by our model can provide a *pattern* which encodes the expression as sufficiently as possible. For the purposes of this paper we shall ignore any differences between the single individuals. However, if desired, a classification of the facial expressions could be performed, for example, by simply minimizing the Euclidean distance between the estimated *pattern* and each landmark facial configuration or by using a finite mixture of Gaussian distributions within the proposed EM algorithm.

We consider data from the FG-NET (Face and Gesture Recognition Research Network) database with facial expressions and emotions from the Technical University Munich (Wallhoff, 2006). The database has been generated in an attempt to assist researchers who investigate the effects of different facial expressions as part of the European Union project. Here, we focus on the happiness and sadness expressions trying to emphasize the main features of their dynamics. We work on landmark data as described in Brombin et al. (2015) and consider the material gathered from 16 different individuals. The complete set of landmarks on the face is shown in Appendix 8. The landmarks have been manually placed on the first frame (representing the neutral expression) of the video sequences available for each subject and then tracked automatically, frame-by-frame, by using the Kanade-Lucas-Tomasi algorithm (Boda and Priyadarsini, 2016). An exploratory analysis of the data shows that the dynamic of both expressions is mainly concentrated at the mouth. Hence, in the following, we limit the analysis only to this region (i.e. landmarks 27-34) and we further consider a downsampling which uses 10 and 20 frames (times)



(a)

(b)

Figure 7: Representation of the mouth for happiness (left) and sadness (right) expressions. The dots represent the Bookstein shape coordinates while the mle fitted paths are from the first order shape regression model with AR(1) errors. The closed circles on the estimated paths represent the position of a landmark at the neutral state.

for happyness and sadness expressions, respectively. Note that the last frame ends with the apex of the expression. Figure 7 shows results from fitting a simple linear model with first order autoregressive structure ($\hat{\phi}_1 = 0.20$ and $\hat{\phi}_1 = 0.40$ for happiness and sadness, respectively). Though simple, this model represents an extension of the regression model with independent errors used in Brombin et al. (2016). The representation is in Bookstein shape coordinates with baseline given by landmarks 11 and 19. As it can be noted, the estimated mean paths are consistent with the definition of the *Action Units* (AUs) described in Ekman et al. (2002) and which code the fundamental actions of individual or groups of muscles for both expressions. In fact, the left subplot clearly shows that happiness is characterised by a raising of the lip corners describing an upward curving of mouth and expansion on vertical and horizontal direction (AU 12+25). On the other hand, the subplot on the right shows that for sadness the lips are stretched horizontally with the lower lip pushed up and the lip corners turned down (AU 15+17). The chin muscle below the lower lip pushes the lower lip upward and is clearly raised in sadness, increasing the size of the lower lip by curling it forward.

6 Discussion

In this paper we have studied the problem of modelling the temporal dynamics of a sequence of shapes. By extending the results proposed in Mardia and Dryden (1989) and Dryden and Mardia (1991) in a space time setting, and with the idea that models specified in the configuration space are flexible and easy to interpret, we have derived a closed form expression for the offset-normal distribution in shape space induced by temporally correlated Gaussian processes specified in the configuration space.

Evaluating the exact likelihood in a dynamic framework is a computationally demanding task since combinatorial complexities appear in the calculation of the product of quadratic forms. For general covariance matrices, and for longitudinal studies with small values of T, we have shown that one possibility to overcome this difficulty is to rely on the technical result of section 4. For longer time series, more structured temporal covariance matrices for autoregressive processes can be considered and an efficient recursive algorithm can then be used as described in section 3.4.

In the application section, we have shown that the proposed models are able to describe the dynamics of several real data and that they can be useful to improve the scientific understanding of the underlying mechanisms of shape changes. Results from section 5.2, have also shown that these models warrant consideration when the interest is in producing forecasts of shape data.

Simulation studies are always important tool for the evaluation of new methods. Hence, to better describe the performance of the EM algorithm, we have also carried out a set of simulations under different scenarios. The study was designed to: a) give an indication of the computational difficulties in evaluating the exact likelihood, b) investigate the statistical properties of the EM algorithm and c) compare the performance with competing methods.

In general, the procedure does not seem to show numerical instabilities and the algorithm tend to converge quickly to the true parameter values. However, in real data applications, running the algorithm from different starting points in the parameter space can increase the chance of finding the global maximum of the likelihood function. Results have also shown that, in the general covariance case, applying the methodology proposed in section 4 may appear to be limited as the result only fully applies in practice for small T and K. However, under separability conditions, the same methodology can be applied to

AR models and results show that with the use of the recursive algorithm Gaussian maximum likelihood is feasible on much larger data sets than was possible previously. Focusing on bias, standard error and mean square error for the parameters, further results also suggest that the proposed estimator shows good statistical properties and that, for temporally correlated shapes, it is preferable to approaches based on tangent space approximations. There is also the hint that the estimator seems to be robust to plausible deviations from normality. Of course, a much more in depth simulation study is required to fully assess the robustness of the estimator as we have only assumed here that the noise distribution follows a (symmetric) Student's t-distribution with known degrees of freedom. Detailed results of the simulations can be found in Appendix 6.

We have discussed the use of autoregressive models which come from separability assumptions of $\Sigma^{\dagger}$. Working with full covariance structures, and avoiding separability, is possible in theory. However, in practice, especially for a large number of landmarks, not all the parameters are identifiable and the likelihood optimization, based on standard numerical routines, is generally difficult and unstable. The development of vector AR process (VAR) appears more feasible in practice and the use of VAR models with a richer parameter structure needs to be explored, also with the availability of a larger data set. This will be a topic for future works.

Finally, we note that the procedure we have described here was given in terms of preforms $\mathbf{X}$ and Bookstein's shape variables. However, as discussed in Kume and Welling (2010), the algorithms can also be derived in terms of the Helmertized preforms 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. If there are missing values, the methods in Kume and Welling (2010) can be used.

Acknowledgments

The authors would like to thank the Editor, the Associate Editor, and a referee for helpful comments and suggestions which have significantly improved the quality of the paper. The authors are also very grateful to J.T. Kent for comments on preliminary versions.

References

T. Al-Jewair, E. Stellrecht, L. Lewandowski, and R. Chakakic. American association of orthodontists foundation craniofacial growth legacy collection in the orthodontic literature use and trends: A systematic review. *American Journal of Orthodontics and Dentofacial Orthopedics*, 153:15–25, 2018.

R. Boda and M.J.P. Priyadarsini. Face detection and tracking using KLT and Viola Jones. *ARPN Journal of Engineering and Applied Sciences*, 11:13472–13476, 2016.

F. L. Bookstein. Size and shape spaces for landmark data in two dimensions. *Statistical Science*, 1: 181–242, 1986.

F. L. Bookstein. *Measuring and Reasoning: Numerical Inference in the Sciences*. Cambridge University Press, Cambridge, 2014.

C. Brombin, L. Salmaso, L. Fontanella, and L. Ippoliti. Nonparametric combination-based tests in dynamic shape analysis. *Journal of Nonparametric Statistics*, 27(4):460–484, 2015.

C. Brombin, L. Salmaso, L. Fontanella, L. Ippoliti, and C. Fusilli. *Parametric and nonparametric Inference for statistical dynamic shape analysis with applications*. Springer, SpringerBriefs in Statistics, 2016.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 34:1–38, 1977.

I. L. Dryden and K. V. Mardia. General shape distributions in the plane. *Advances in Applied Probability*, 23:259–276, 1991.

I.L. Dryden and K. V. Mardia. *Statistical shape analysis with Applications in R*. Wiley series in probability and statistics. Wiley, Chichester, 2016.

P. Ekman, Friesen W.V., and J.C. Hager. *Facial Action Coding System Investigator's Guide*. Consultant Pschologists Press, Salt Lake City, UT, 2002.

L. Franchi, T. Baccetti, and J. A. McNamara. Thin-plate spline analysis of mandibular growth. *Angle Orthodontist*, 71:83–89, 2001.

M.G. Genton. Separable approximations of space-time covariance matrices. *Environmetrics*, 18:681–695, 2007.

Q. He, Y. Duan, K. Karsch, and J. Miles. Detecting corpus callosum abnormalities in autism based on anatomical landmarks. *Psychiatry Research: Neuroimaging*, (183):126–132, 2010.

C. Huang, M. Styner, and H. Zhu. Clustering high-dimensional landmark-based two-dimensional shape data. *Journal of the American Statistical Association*, 110:946–961, 2015.

R. Kan. From moments of sum to moments of product. *Journal of Multivariate Analysis*, 99:542–554, 2008.

D. G. Kendall. The Mardia-Dryden distribution for triangles: a stochastic calculus approach. *Journal of Applied Probability*, 28:239–247, 1991.

K. Kenobi, I.L. Dryden, and H. Le. Shape curves and geodesic modeling. *Biometrika*, 97:567–584, 2010.

J. T. Kent and K. V. Mardia. Consistency of procrustes estimators. *Journal of the Royal Statistical Society, Series B*, (59):281–290, 1997.

J.T. Kent, K.V. Mardia, R.J. Morris, and R.G. Aykroyd. Functional models of growth for landmark data. In K.V. Mardia, J.T. Kent, and R.G. Aykroyd, editors, *Proceedings in Functional and Spatial Data Analysis*, pages 109–115. Springer, 2001.

J.T. Kent, K.V. Mardia, and P. McDonnell. The complex bingham quartic distribution and shape analysis. *Journal of the Royal Statistical Society, Series B*, 68:747–765, 2006.

A. Kume and M. Welling. Maximum likelihood estimation for the offset-normal shape distributions using EM. *Journal of Computational and Graphical Statistics*, 19:702–723, 2010.

A. Kume, I.L. Dryden, and H. Le. Shape space smoothing splines for planar landmark data. *Biometrika*, 94:513–528, 2007.

A. Kume, I.L. Dryden, and A.T.A. Wood. Shape inference based on multivariate normal matrix distributions. *Technical Report, University of Kent. Submitted*, 2017.

H. Le. On the consistency of procrustean mean shapes. *Advances in Applied Probability*, 30:53–63, 1998.

H. Le and A. Kume. Detection of shape changes in biological features. *Journal of Microscopy*, 2:140–147, 2000.

H. Le and C. G. Small. Multidimensional scaling of simplex shapes. *Pattern recognition*, 32:1601–1613, 1999.

S. Lele and J. T. Richtsmeier. Euclidean distance matrix analysis: a coordinate-free approach for comparing biological shapes using landmark data. *American Journal of Physical Anthropology*, 86:239–359, 1991.

J. Magnus. The exact moments of a ratio of quadratic forms in normal variables. *Annales d'Economie et de Statistique*, 4:95–109, 1986.

K. V. Mardia and I. L. Dryden. The statistical analysis of shape data. *Biometrika*, 76:271–281, 1989.

K. V. Mardia and A. N. Walder. Shape analysis of paired landmark data. *Biometrika*, 81:185–196, 1994.

A.M. Mathai and S.B. Provost. *Quadratic Forms in Random Variables: Theory and Applications*. Marcel Dekker, New York, 1992.

R.E. Moyers and F.L. Bookstein. The inappropriateness of conventional cephalometrics. *American Journal of Orthodontics*, 75:599–617, 1979.

F.D. Neeser and J.L. Massey. Proper complex random processes with applications to information theory. *IEEE Transaction on Information Theory*, 39:1293–1302, 1993.

R. Oyonarte, M. Hurtado, and M.V. Castro. Evolution of ANB and SN-GoGn angles during craniofacial growth: a retrospective longitudinal study. *APOS Trends in Orthodontics*, 6, 2016.

M. Pantic, R. Cowie, F. Drrico, D. Heylen, M. Mehu, and P. Pelachaud. Social signal processing: the research agenda. In *Visual Analysis of Humans*, pages 511–538. Springer, London, 2011.

P. K. Sridhar. *Textbook of Craniofacial Growth*. Jaypee Brothers Medical Publishers, New Delhi, 2011.

A. Stuart and K. Ord. Kendalls advanced theory of statistics. In *Distribution Theory*. Arnold, London, 1994.

F. Wallhoff. Facial expressions and emotion database. http://www.mmk.ei.tum.de/ waf/fgnet/feedtum.html, 2006.

N. Wang, G. Xinbo, T. Dacheng, Y. Heng, and L. Xuelong. Facial feature point detection: a comprehensive survey. *Neurocomputing*, 275:50–65, 2018.

SUPPLEMENTARY MATERIAL

Appendix 1: Bookstein coordinates

Bookstein shape coordinates are defined as $\mathbf{U}_t = (u_{k,t}\, v_{k,t}),\ k = 1, \ldots, K-1$, with $u_{1,t} = 1$ and $v_{1,t} = 0$. $\mathbf{U}_t$ can be computed through the mapping $\mathbf{X}_t \to \mathbf{U}_t = \beta_t \mathbf{X}_t \mathbf{R}_t$, where at each time t, the scaling factors and the rotation matrices are given by (Dryden and Mardia, 2016, pg. 43)

$$\beta_t = (x_{2,t}^2 + y_{2,t}^2)^{-1} \quad \text{and} \quad \mathbf{R}_t = \begin{pmatrix} x_{2,t} & -y_{2,t} \\ y_{2,t} & x_{2,t} \end{pmatrix}, \qquad t = 1, \ldots, T.$$

The temporal sequence of shape coordinates, $\mathbf{U} = \Big(\mathbf{U}_1\, \mathbf{U}_2\, \ldots\, \mathbf{U}_T\Big)$, is thus obtained through the transformation $\mathbf{X} \to \mathbf{U} = \mathbf{X}\mathbf{R}$, where $\mathbf{R} = diag\big(\beta_1 \mathbf{R}_1, \beta_2 \mathbf{R}_2, \ldots, \beta_T \mathbf{R}_T\big)$.

The "reduced" shape coordinates are defined as $\mathbf{u} = \big\{(u_{k,t}, v_{k,t}),\ k = 2, \ldots, K-1, t = 1, \ldots, T\big\}$; i.e. they do not include the first landmark with coordinates $u_{1,t} = 1$ and $v_{1,t} = 0$.

Appendix 2: ML estimators for the regression parameters and covariance matrices

ML estimator of the regression parameters for the complete data is given by

$$\hat{\boldsymbol{\beta}} = \frac{1}{N} \sum_{n=1}^{N} \big(\mathbf{D}'\boldsymbol{\Omega}^{-1}\mathbf{D}\big)^{-1} \mathbf{D}'\boldsymbol{\Omega}^{-1} vec(\mathbf{X}^{(n)}),$$

In order to derive the update rules for the covariance matrices, we write their ML estimators as

$$\hat{\boldsymbol{\Sigma}}_S = \frac{1}{NT} \sum_{n=1}^{N} \sum_{t=1}^{T} \check{\mathbf{P}}_t\, vec(\mathbf{X}^{(n)}) vec(\mathbf{X}^{(n)})'\, \check{\mathbf{P}}_t' - \frac{1}{T} \hat{\boldsymbol{\mu}}\, \boldsymbol{\Sigma}_T^{-1}\, \hat{\boldsymbol{\mu}}'$$

and

$$\hat{\boldsymbol{\Sigma}}_T = \frac{1}{N(2K-2)} \sum_{n=1}^{N} \sum_{k=1}^{2K-2} \check{\mathbf{P}}_k\, vec(\mathbf{X}^{(n)}) vec(\mathbf{X}^{(n)})'\, \check{\mathbf{P}}_k' - \frac{1}{2K-2} \hat{\boldsymbol{\mu}}'\, \boldsymbol{\Sigma}_S^{-1}\, \hat{\boldsymbol{\mu}},$$

where, from the Cholesky decompositions $\hat{\boldsymbol{\Sigma}}_T^{-1} = \mathbf{L}_T \mathbf{L}_T'$ and $\hat{\boldsymbol{\Sigma}}_S^{-1} = \mathbf{L}_S \mathbf{L}_S'$, we have $\check{\mathbf{P}}_t = (\mathbf{L}_T \boldsymbol{e}_t)' \otimes \mathbf{I}_{2K-2}$ and $\check{\mathbf{P}}_k = \mathbf{I}_T \otimes (\mathbf{L}_S \boldsymbol{e}_k)'$; $\boldsymbol{e}_t$ and $\boldsymbol{e}_k$ are, respectively, T-dimensional and $(2K-2)$-dimensional vectors with entries, $e_j(j) = 1$ for $j = t, k$, and zero otherwise.

Appendix 3: Proof of Result 1 in section 3.4

To save space, the proof is given for an AR(1) process. Extension to higher order is straightforward. We use the following approximation

$$E\big[vec(\mathbf{X}_t)|\mathbf{W}_0, \mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_t, \mathbf{W}_{t+1}, \ldots, \mathbf{W}_T\big] \simeq E\big[vec(\mathbf{X}_t)|\mathbf{W}_0, \mathbf{W}_1, \ldots, \mathbf{W}_t\big].$$

and implement this iteratively using the following identities based on the iterated expectations.

For $t = 1$, $vec(\mathbf{X}_1) = \tilde{\boldsymbol{\mu}}(1-\phi_1) + \phi_1 vec(\mathbf{X}_0) + vec(\mathbf{E}_1)$ and since $\mathbf{X}_0$ is assumed given, we have $E\big[vec(\mathbf{X}_1)|\mathbf{W}_0, \mathbf{W}_1\big] = E\big[vec(\mathbf{X}_1)|\mathbf{W}_1\big] = \mathbf{W}_1 Q(\tilde{\boldsymbol{\mu}}_{1|0}, \boldsymbol{\Sigma}_S, \mathbf{W}_1)$, with $\tilde{\boldsymbol{\mu}}_{1|0} = \tilde{\boldsymbol{\mu}}(1-\phi_1) + \phi_1 vec(\mathbf{X}_0)$. For $t = 2, 3, \ldots, T$ we also have $vec(\mathbf{X}_t) = \tilde{\boldsymbol{\mu}}(1-\phi_1) + \phi_1 vec(\mathbf{X}_{t-1}) + vec(\mathbf{E}_t)$ with

$$\begin{aligned}
E\big[vec(\mathbf{X}_t)|\mathbf{W}_0, \mathbf{W}_1, \ldots, \mathbf{W}_t\big] &= E\big[\tilde{\boldsymbol{\mu}}(1-\phi_1) + \phi_1 vec(\mathbf{X}_{t-1}) + vec(\mathbf{E}_t)|\mathbf{W}_0, \mathbf{W}_1, \ldots, \mathbf{W}_t\big] \\
&= E\big[E\big[\tilde{\boldsymbol{\mu}}(1-\phi_1) + \phi_1 vec(\mathbf{X}_{t-1}) + vec(\mathbf{E}_t)|\mathbf{W}_0, \mathbf{W}_1, \ldots, \mathbf{W}_{t-1}\big]|\mathbf{W}_t\big] \\
&= E\big[\tilde{\boldsymbol{\mu}}(1-\phi_1) + \phi_1 E\big[vec(\mathbf{X}_{t-1})|\mathbf{W}_0, \mathbf{W}_1, \ldots, \mathbf{W}_{t-1}\big] + vec(\mathbf{E}_t)|\mathbf{W}_t\big] \\
&= E\big[\tilde{\boldsymbol{\mu}}_{t|t-1} + vec(\mathbf{E}_t)|\mathbf{W}_t\big] \\
&= \mathbf{W}_t Q(\tilde{\boldsymbol{\mu}}_{t|t-1}, \boldsymbol{\Sigma}_S, \mathbf{W}_t), \quad t = 1, \ldots, T. \quad \Box
\end{aligned}$$

where $E\big[vec(\mathbf{X}_{t-1})|\mathbf{W}_0, \mathbf{W}_1, \ldots, \mathbf{W}_{t-1}\big]$ is implemented in the previous step, so that $\tilde{\boldsymbol{\mu}}_{t|t-1}$ acts as the forecast mean of observation $vec(\mathbf{X}_t)$. This also shows that, given the past, the conditional likelihood can be computed recursively and, for an iid sample of temporally correlated configurations, the update rule for the mean in the M-step is obtained as

$$\begin{aligned}
\tilde{\boldsymbol{\mu}}^{(r+1)} = \frac{1}{\alpha}\sum_{n=1}^{N}\sum_{t=2}^{T}\int vec\left(\mathbf{X}_t^{(n)}\right) dF\left(\mathbf{X}^{(n)}|\mathbf{u}^{(n)}, \tilde{\boldsymbol{\mu}}^{(r)}, \phi_1^{(r)}, \boldsymbol{\Sigma}_S\right) - \\
\frac{\phi_1^{(r)}}{\alpha}\sum_{n=1}^{N}\sum_{t=2}^{T}\int vec\left(\mathbf{X}_{t-1}^{(n)}\right) dF\left(\mathbf{X}^{(n)}|\mathbf{u}^{(n)}, \tilde{\boldsymbol{\mu}}^{(r)}, \phi_1^{(r)}, \boldsymbol{\Sigma}_S\right)
\end{aligned}$$

where $\alpha = N(T-1)(1-\phi_1^{(r)})$.

Appendix 4: Baseline invariance

By standardizing along a different pair of landmarks, we essentially re-parametrize the problem and, accordingly, the MLE solutions represent the same objects. For example, the preform of $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, obtained after a choice of baseline landmarks, say 1 and 2, is actually linearly related via a matrix $\boldsymbol{P}^{12,kl}$ to the preform of $\boldsymbol{X}$ but with baseline landmarks choice k and l (see Lemma 3 in Kume and Welling, 2010 for explicit expression of $\boldsymbol{P}^{12,kl}$). Namely, the preform with baseline k and l coming from $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the same in distribution as the preform coming from $\hat{\mathbf{X}} = \boldsymbol{P}^{12,kl}\boldsymbol{X} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ with baseline 1 and 2 where $\hat{\boldsymbol{\mu}} = \boldsymbol{P}^{12,kl}\boldsymbol{\mu}$ and $\hat{\boldsymbol{\Sigma}} = \boldsymbol{P}^{12,kl}\boldsymbol{\Sigma}\boldsymbol{P}^{12,kl\prime}$.

Therefore, following a similar discussion as in Kume and Welling (2010), the elementary conditional expectations of first and second conditional moments need to reflect the appropriate linear relations. For example, if we are to evaluate our elementary expectations, $E(\boldsymbol{X}_t|\boldsymbol{U}_t^{k,l}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $E(\boldsymbol{X}_t\boldsymbol{X}_t'|\boldsymbol{U}_t^{k,l}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, in terms of baseline coordinates k, l, we have

$$E(\boldsymbol{X}_t|\boldsymbol{U}_t^{k,l}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \boldsymbol{P}^{12,kl^{-1}}E(\boldsymbol{X}_t|\hat{\boldsymbol{U}}_t^{1,2}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$$

$$E(\boldsymbol{X}_t\boldsymbol{X}_t'|\boldsymbol{U}_t^{k,l}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \boldsymbol{P}^{12,kl^{-1}}E(\boldsymbol{X}_t\boldsymbol{X}_t^t; \hat{\boldsymbol{U}}_t^{1,2}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})\boldsymbol{P}^{12,kl^{-t}},$$

where the expectations on the right are the same as if one chooses the baseline on landmarks 1 and 2 while rearranging the mean and variance components accordingly.

Appendix 5: Proof of Result 2 in section 4

The expected value of $h_j \prod_{t=1}^T (\mathbf{h}'\mathbf{A}_t\mathbf{h})^{K-2}$, for $j = 1, \ldots, 2T$, can be obtained considering the following development for the expectation of $(h_j - w)^2 \prod_{t=1}^T (\mathbf{h}'\mathbf{A}_t\mathbf{h})^{K-2}$

$$\begin{aligned}
&\int (h_j - w)^2 \prod_{t=1}^T (\mathbf{h}'\mathbf{A}_t\mathbf{h})^{K-2} f_{\mathcal{N}_{2T}}(\mathbf{h}|\boldsymbol{\eta}, \boldsymbol{\Gamma}) f_N(w|1,1) d\mathbf{h}dw = \\
&= \int h_j^2 \prod_{t=1}^T (\mathbf{h}'\mathbf{A}_t\mathbf{h})^{K-2} f_{\mathcal{N}_{2T}}(\mathbf{h}|\boldsymbol{\eta}, \boldsymbol{\Gamma}) d\mathbf{h} \int f_{\mathcal{N}}(w|1,1)dw \\
&- 2\int h_j \prod_{t=1}^T (\mathbf{h}'\mathbf{A}_t\mathbf{h})^{K-2} f_{\mathcal{N}_{2T}}(\mathbf{h}|\boldsymbol{\eta}, \boldsymbol{\Gamma}) d\mathbf{h} \int w f_{\mathcal{N}}(w|1,1)dw \\
&+ \int \prod_{t=1}^T (\mathbf{h}'\mathbf{A}_t\mathbf{h})^{K-2} f_{\mathcal{N}_{2T}}(\mathbf{h}|\boldsymbol{\eta}, \boldsymbol{\Gamma}) d\mathbf{h} \int w^2 f_{\mathcal{N}}(w|1,1)dw
\end{aligned}$$

Since

$$\int f_{\mathcal{N}}(w|1,1)dw = 1, \quad \int w f_{\mathcal{N}}(w|1,1)dw = 1, \quad \int w^2 f_{\mathcal{N}}(w|1,1)dw = 2$$

we have

$$\begin{aligned}
&\int (h_j - w)^2 \prod_{t=1}^T (\mathbf{h}'\mathbf{A}_t\mathbf{h})^{K-2} f_{\mathcal{N}_{2T}}(\mathbf{h}|\boldsymbol{\eta}, \boldsymbol{\Gamma}) f_{\mathcal{N}}(w|1,1) d\mathbf{h}dw = \int h_j^2 \prod_{t=1}^T (\mathbf{h}'\mathbf{A}_t\mathbf{h})^{K-2} f_{\mathcal{N}_{2T}}(\mathbf{h}|\boldsymbol{\eta}, \boldsymbol{\Gamma}) d\mathbf{h} + \\
&- 2\int h_j \prod_{t=1}^T (\mathbf{h}'\mathbf{A}_t\mathbf{h})^{K-2} f_{\mathcal{N}_{2T}}(\mathbf{h}|\boldsymbol{\eta}, \boldsymbol{\Gamma}) d\mathbf{h} + 2\int \prod_{t=1}^T (\mathbf{h}'\mathbf{A}_t\mathbf{h})^{K-2} f_{\mathcal{N}_{2T}}(\mathbf{h}|\boldsymbol{\eta}, \boldsymbol{\Gamma}) d\mathbf{h}
\end{aligned}$$

Given

$$\begin{aligned}
\mathcal{I}_{3_j} &= \int (h_j - w)^2 \prod_{t=1}^{T} (\mathbf{h}'\mathbf{A}_t\mathbf{h})^{K-2} f_{\mathcal{N}_{2T}}(\mathbf{h}|\boldsymbol{\eta}, \boldsymbol{\Gamma}) f_{\mathcal{N}}(w|1,1) d\mathbf{h} dw \\
\mathcal{I}_{1_j} &= \int h_j^2 \prod_{t=1}^{T} (\mathbf{h}'\mathbf{A}_t\mathbf{h})^{K-2} f_{\mathcal{N}_{2T}}(\mathbf{h}|\boldsymbol{\eta}, \boldsymbol{\Gamma}) d\mathbf{h} \\
\mathcal{I}_j &= \int h_j \prod_{t=1}^{T} (\mathbf{h}'\mathbf{A}_t\mathbf{h})^{K-2} f_{\mathcal{N}_{2T}}(\mathbf{h}|\boldsymbol{\eta}, \boldsymbol{\Gamma}) d\mathbf{h} \\
\mathcal{I}_2 &= \int \prod_{t=1}^{T} (\mathbf{h}'\mathbf{A}_t\mathbf{h})^{K-2} f_{\mathcal{N}_{2T}}(\mathbf{h}|\boldsymbol{\eta}, \boldsymbol{\Gamma}) d\mathbf{h}
\end{aligned}$$

and $\mathcal{I}_{3_j} = \mathcal{I}_{1_j} - 2\mathcal{I}_j + 2\mathcal{I}_2$, the integral of interest can be obtained as $\mathcal{I}_j = \frac{1}{2}\mathcal{I}_{1_j} + \mathcal{I}_2 - \frac{1}{2}\mathcal{I}_{3_j}$
The solution is

$$\begin{aligned}
&\int h_j \prod_{t=1}^{T} (\mathbf{h}'\mathbf{A}_t\mathbf{h})^{K-2} f_{\mathcal{N}_{2T}}(\mathbf{h}|\boldsymbol{\eta}, \boldsymbol{\Gamma}) d\mathbf{h} = 0.5 \int h_j^2 \prod_{t=1}^{T} (\mathbf{h}'\mathbf{A}_t\mathbf{h})^{K-2} f_{\mathcal{N}_{2T}}(\mathbf{h}|\boldsymbol{\eta}, \boldsymbol{\Gamma}) d\mathbf{h} \\
&+ \int \prod_{t=1}^{T} (\mathbf{h}'\mathbf{A}_t\mathbf{h})^{K-2} f_{\mathcal{N}_{2T}}(\mathbf{h}|\boldsymbol{\eta}, \boldsymbol{\Gamma}) d\mathbf{h} - 0.5 \int (h_j - w)^2 \prod_{t=1}^{T} (\mathbf{h}'\mathbf{A}_t\mathbf{h})^{K-2} f_{\mathcal{N}_{2T}}(\mathbf{h}|\boldsymbol{\eta}, \boldsymbol{\Gamma}) f_N(w|1,1) d\mathbf{h} dw
\end{aligned}$$

Appendix 6: Simulation results

This section provides a discussion of a set of simulations we have carried out to investigate the performance of the EM approach under different scenarios. We begin by investigating how the estimation procedure performs under the results proposed in sections 3 and 4. The study is designed to give an indication of the computational burden of the EM algorithm, especially for different values of K, T and N.

The first set of simulations gives an idea of the computational difficulties of working with the exact likelihood. Table 1 compares the CPU times[1] (in seconds) required to compute the expectations in equations (9) and (13) using both Laguerre polynomials and equations (4) and (5). The analysis is carried out using a "minimal model specification setting" with $K = 3$ (*i.e.* a triangle), $T = 4, 5, 6$, $N = 10, 20, 30$ and $\boldsymbol{\Sigma}_T$ equal to the correlation matrix of a first order autoregressive process with $\phi_1 = 0.5$.

Results suggest that is difficult to work with Laguerre polynomials with $K = 3$ and $T > 6$ and that, in general, despite an explicit expression for $Q_s(\mathbf{B}_v)$ is available (see, for example, the discussion in Brombin et al., 2016), it is impossible to use Laguerre polynomials for K and T greater than 4. Kan's formulation and results from section 4 allow for a more efficient procedure. However, estimation remains infeasible for K and T greater than 6.

The second set of simulations is thus carried out using results from section 3.4 where the expectation of equation (9) is computed recursively. By using $K = 8$, $T = 8, 15, 30, 50$ and $N = 1, 10, 20$, Table 2 shows the results from 50 simulations of a first order autoregressive process with parameter values $\phi_1 = 0.25, 0.5, 0.75$, constant mean $\tilde{\boldsymbol{\mu}}$

[1] Results are obtained in Matlab with an Intel(R) Core(TM) i7-4558U CPU 2.80 GHz with 8 GB.

and $\mathbf{\Sigma}_S^\dagger$ having an isotropic structure with $\sigma^2 = 1$.

	$K = 3$	$K = 4$	$K = 5$	$K = 6$
		Kan's formulation		
$T = 3$	0.05	0.12	0.27	0.62
$T = 4$	0.09	0.45	1.74	5.22
$T = 5$	0.19	1.87	10.26	40.00
$T = 6$	0.45	7.55	59.32	288.31
		Laguerre polynomials		
$T = 3$	0.26	6.8	86.11	-
$T = 4$	3.56	389.0	11693.86	-
$T = 5$	54.0	21780.00	-	-
$T = 6$	825.0	-	-	-

Table 1: CPU time (in seconds) required to compute the expectations in equations (9) and (13). Results refer to a single iteration of the E-step of the EM algorithm, assuming $N = 1$, $K = 3, \ldots, 6$ and T varying from 3 to 6.

		N			N			N	
T	1	10	20	1	10	20	1	10	20
		$\phi_1 = 0.25$			$\phi_1 = 0.50$			$\phi_1 = 0.75$	
8	0.008 (0.087)	0.004 (0.029)	0.017 (0.027)	-0.266 (0.127)	0.010 (0.024)	0.011 (0.021)	0.002 (0.096)	0.005 (0.029)	0.009 (0.022)
15	0.011 (0.080)	0.003 (0.025)	0.002 (0.018)	-0.111 (0.067)	0.002 (0.024)	0.003 (0.014)	-0.002 (0.048)	0.005 (0.014)	0.004 (0.013)
30	-0.008 (0.065)	-0.001 (0.025)	0.000 (0.018)	-0.041 (0.050)	0.001 (0.015)	-0.001 (0.009)	0.004 (0.034)	0.006 (0.011)	-0.002 (0.008)
50	0.003 (0.037)	0.005 (0.011)	0.003 (0.009)	-0.027 (0.028)	0.002 (0.012)	0.005 (0.007)	0.014 (0.028)	-0.004 (0.010)	0.002 (0.007)

Table 2: The bias (and standard errors) of the autoregressive parameter estimates from 50 simulations for different N and T. The true parameters are $\phi_1 = 0.25,\ 0.50$ and 0.75 and the mean, $\tilde{\boldsymbol{\mu}}$, is constant.

Results suggest that the bias, the standard errors and the MSE for ϕ_1 become negligible as T and N increase. The bias is larger for $N = 1$, especially for $\phi_1 = 0.5$ for which it appears always negative. However, in general, as the size of the sample increases, the estimates become closer to the actual parameter and the standard errors are also reasonably small.

For the same set of simulations, Table 3 also shows the means and the standard errors of the Full Procrustes distance ρ, computed between the estimated mean, $\hat{\tilde{\boldsymbol{\mu}}}_i,\ i = 1, \ldots, 50$, and the true mean, $\tilde{\boldsymbol{\mu}}$. In general, we have that the estimated mean does not differ much from the mean shape parameter since ρ, averaged over the 50 simulations, is quite small. However, there is also the hint that the estimation of the mean shape appears slightly more difficult as the temporal correlation increases.

The concepts of a mean shape and shape correlation have an underpinning role in our applications. Regarding the estimation of these model parameters, it may be possible that different approaches can give rise to different results. The following simulation thus aims at motivating the use of the offset-normal

		N			N			N	
T	1	10	20	1	10	20	1	10	20
		ρ			ρ			ρ	
8	0.025 (0.010)	0.008 (0.003)	0.005 (0.002)	0.068 (0.036)	0.028 (0.014)	0.017 (0.006)	0.155 (0.084)	0.053 (0.023)	0.022 (0.015)
15	0.018 (0.006)	0.006 (0.003)	0.004 (0.002)	0.055 (0.030)	0.020 (0.008)	0.012 (0.006)	0.129 (0.061)	0.044 (0.021)	0.019 (0.011)
30	0.014 (0.008)	0.004 (0.002)	0.003 (0.001)	0.037 (0.014)	0.013 (0.005)	0.008 (0.003)	0.104 (0.051)	0.035 (0.018)	0.011 (0.007)
50	0.012 (0.005)	0.004 (0.002)	0.002 (0.001)	0.031 (0.012)	0.010 (0.005)	0.005 (0.001)	0.082 (0.033)	0.028 (0.014)	0.006 (0.005)

Table 3: The means (and standard errors) of the Full Procrustes distance ρ computed for each estimated mean, $\hat{\tilde{\boldsymbol{\mu}}}_i,\ i = 1, \ldots, 50$, and the true mean shape, $\tilde{\boldsymbol{\mu}}$.

distribution by comparing EM parameter estimates with those obtained by using Procrustes tangent coordinates (Dryden and Mardia, 2016). Since this is a commonly used approach in practical shape analysis, one of the first things to investigate is whether the distribution of points in the tangent space may be used as satisfactory approximation to their distribution in shape space. To this purpose, Table 4 shows estimation results for simulated AR(1) with $\phi_1 = 0.5$. The numbers of landmarks, time points and independent configurations are fixed as, $K = 8$, $T = 8$ and $N = 20$, respectively. An isotropic structure, with two different levels of variability ($\sigma = 1, 3$), is also considered for the error term. The mean of the process, $\tilde{\boldsymbol{\mu}}$, assumed as constant in time, is represented by the mean shape of the data used in example 5.3 (i.e. the happiness configuration) and its coordinates range from 279.42 to 356.25.

Results suggest that the differences in estimating the mean are not so large as the Full Procrustes distances between and $\hat{\tilde{\boldsymbol{\mu}}}$ and $\tilde{\boldsymbol{\mu}}$ are similar. However, using the Procrustes tangent coordinates clearly affects the estimate of the autoregressive parameter ϕ since, working in the tangent space, both the bias and MSE appear larger than those observed for the ML approach. Also, we note that, because of the linear approximation, working in the tangent space may limit the interpretation of ϕ, especially for larger error variances.

	$\sigma = 1$				$\sigma = 3$			
T	ρ_{pr}	ρ_{ml}	$\hat{\phi}_{pr}$	$\hat{\phi}_{ml}$	ρ_{pr}	ρ_{ml}	$\hat{\phi}_{pr}$	$\hat{\phi}_{ml}$
8	0.001	0.000	0.476 (0.033)	0.499 (0.028)	0.002	0.002	0.474 (0.031)	0.500 (0.028)

Table 4: Comparison of the estimates obtained by using Procrustes tangent coordinates and ML. $\hat{\phi}_{pr}$ and $\hat{\phi}_{ml}$ represent the the means (and standard errors), taken over 100 simulations of the AR(1) process, of the autoregressive parameter, $\phi = 0.5$. The Full Procrustes distances, ρ_{pr} and ρ_{ml}, represents the distance between the estimated mean, $\hat{\tilde{\boldsymbol{\mu}}}$, and the true mean shape, $\tilde{\boldsymbol{\mu}}$, using Procrustes tangent coordinates and the EM approach, respectively.

Finally, to asses the performance of the EM estimator when the data exhibit departures from the normal distribution, we have carried out further simulations with non-Gaussian noise. In Table 5, we thus show results from 50 simulations for an AR(1) process with error terms following a Student's t-distribution. The mean is assumed constant, $K = 8$, $T = 15$, $N = 20$ and $\phi = 0.5$ as used in Table 2 above. Results show that the estimator seems to be robust to plausible deviations from normality. However, when the number of degrees of freedom result smaller than 5, then both the bias and the standard errors for ϕ appear to increase significantly. The Full Procrustes distance, ρ_{ml}, between the estimated mean, $\hat{\tilde{\boldsymbol{\mu}}}$, and the true mean shape, $\tilde{\boldsymbol{\mu}}$, also appears to increase considerably.

	df			
	5	3	2.5	2.2
$\hat{\phi}_{ml}$	0.010 (0.016)	0.022 (0.019)	0.046 (0.061)	0.057 (0.070)
ρ_{ml}	0.023	0.038	0.146	0.184

Table 5: For different degrees of freedom (df) of the Student's t-distribution, the first row shows the bias (and standard errors) of the autoregressive parameter estimates from 50 simulations with $N = 20$, $T = 15$, $\phi_1 = 0.50$ and constant mean. The second row, gives the Full Procrustes distance, ρ_{ml}, between the estimated mean, $\hat{\tilde{\boldsymbol{\mu}}}$, and the true mean shape, $\tilde{\boldsymbol{\mu}}$.

Appendix 7: ML estimates of the landmarks and temporal covariance matrices for the example 5.1

	(x_1, y_1)	(x_2, y_2)	(x_3, y_3)	(x_4, y_4)	(x_5, y_5)	(x_6, y_6)	(x_7, y_7)	(x_8, y_8)
(x_1, y_1)	1.48+0.00i	0.19 + 0.02i	0.36 - 0.11i	0.72 + 0.33i	1.27 + 0.20i	1.36 - 0.21i	1.31 - 0.00i	1.09 + 0.15i
(x_2, y_2)	0.19 - 0.02i	0.02+0.00i	0.04- 0.02i	0.10 + 0.03i	0.16 + 0.01i	0.17 - 0.04i	0.17 - 0.02i	0.14 + 0.00i
(x_3, y_3)	0.36 + 0.11i	0.04 + 0.02i	0.10+0.00i	0.15 + 0.13i	0.30 + 0.14i	0.35 + 0.05i	0.32 + 0.09i	0.26 + 0.12i
(x_4, y_4)	0.72 - 0.33i	0.10 - 0.03i	0.15 - 0.13i	0.43+0.00i	0.66 - 0.19i	0.62 - 0.41i	0.64 - 0.30i	0.57 - 0.17i
(x_5, y_5)	1.27 - 0.20i	0.16 - 0.01i	0.30 - 0.14i	0.66 + 0.19i	1.12+0.00i	1.14 - 0.36i	1.12 - 0.19i	0.96 - 0.02i
(x_6, y_6)	1.36 + 0.21i	0.17 + 0.04i	0.35 - 0.05i	0.62 + 0.41i	1.14 + 0.36i	1.28+0.00i	1.20 + 0.17i	0.99 + 0.29i
(x_7, y_7)	1.31 + 0.01i	0.17 + 0.02i	0.32 - 0.09i	0.64 + 0.30i	1.12 + 0.19i	1.20 - 0.17i	1.15+0.00i	0.97 + 0.14i
(x_8, y_8)	1.09 - 0.15i	0.14 - 0.00i	0.26 - 0.12i	0.57 + 0.17i	0.96 + 0.02i	0.99 - 0.29i	0.97 - 0.14i	0.83+0.00i

Table 6: ML estimates of the complex covariance matrix $\mathbf{\Sigma}_S$ for the 8 landmarks of the CC in the preform space. The estimates refer to the pooled sample and are obtaied using equation (10).

	t_1	t_2	t_3
t_1	1.00	0.41	0.36
t_2	0.41	1.00	0.44
t_3	0.36	0.44	1.00

Table 7: ML estimates of the temporal correlation matrix $\mathbf{\Sigma}_T$ for the three visits. The estimates refer to the pooled sample and are obtaied using equation (11).

Appendix 8: Set of landmarks for the FG-NET data

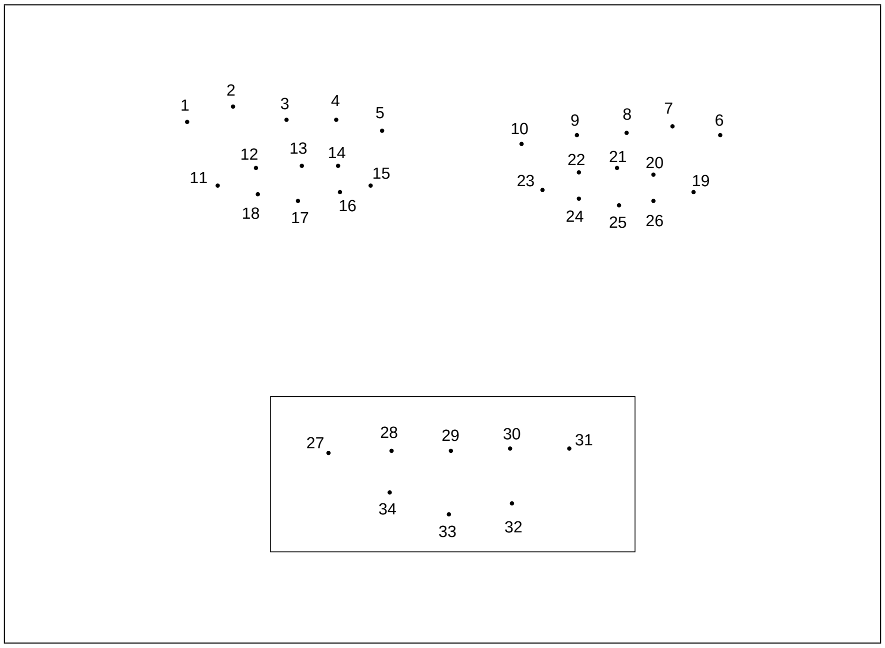


Figure 8: Example of facial landmark configuration with 34 landmarks. The box around the mouth highlights the landmarks of interest.

Appendix 9: Rat skulls data

A statistical analysis similar to that described in section 5.2, is described here for modelling the dynamics of eight biological landmarks measured on the skulls of 18 rats at ages $7, 14, 21, 30, 40, 60, 90$, and 150 days. Results from the analysis suggest that a first order autoregressive model with constant mean does not look reasonable since the estimated autoregressive parameter ϕ_1 is very close to 1. Kume and Welling (2010) have shown that the non-stationary features of the process can be much better described by using a polynomial function for the mean. The best AIC value among shape regression models assuming temporally independent errors and isotropic landmark covariance matrix, is found for a second order polynomial trend model, with $\mathcal{P} = 42$ parameters, which has AIC$= -4563.4$. The AIC value for the first order polynomial trend, with $\mathcal{P} = 28$, is equal to -4321.4. The use of shape regression models for the mean with an AR(1) structure for the errors suggests a slightly better model fit to the data. In fact, the AIC for the second order polynomial model, which has $\hat{\phi}_1 = 0.19$, is equal to -4700.3.

By making use of Bookstein shape coordinates, results of the estimation procedure is shown in Figure 9. For comparison purposes, the fit of our regression model is compared with the fit of the functional growth model (using the "special" parametrization) described in Kent et al. (2001). The figure shows that both models are able to provide a good description of the dynamics of the data and the fits represent an improvement compared with those provided by Kume and Welling (2010) and Le and Kume (2000).

Forecasting results are also shown in Figure 10 where the one-step ahead forecast is compared with the Bookstein shape coordinates of the rats at the last age (originally excluded from the analysis). Specifically, the upper panels show the best (left) and worse (right) predictions corresponding to the minimum (0.077) and maximum (0.340) values of the distribution of the root mean squared prediction errors obtained for the 18 rats. The forecast corresponding to the $75°$ percentile (0.210) is also represented in the middle panel. The bottom panels emphasize the magnitude of the prediction error at each landmark, with the length of each arrow (blown up for visibility) representing the distance between predicted and true coordinate shapes. In general, results suggest the quite good shape predictions at the eighth age can be obtained for most of the analyzed rats. For some of the rats, there seems to be difficulties in predicting the growth at the *Labda*, *Bregma*, *Spheno-ethmoidal synchondrosis*, *Intersphenoidal suture* and *Spheno-occipital synchondrosis* landmarks.

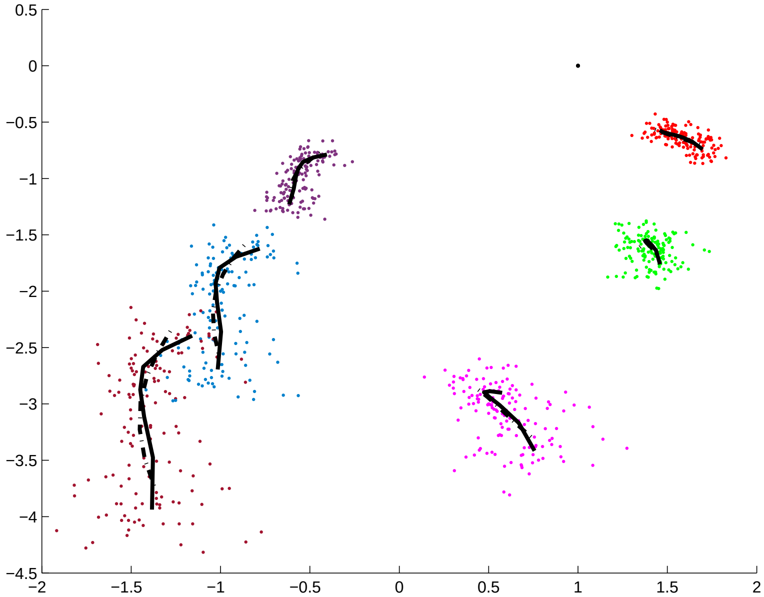


Figure 9: Rat skulls in Bookstein shape coordinates with baseline defined by points (0,0) and (1,0). The dots represent the observed configuration landmarks while the fitted paths are from the second order shape regression model with AR(1) errors (solid line) and the functional growth model (dashed line)

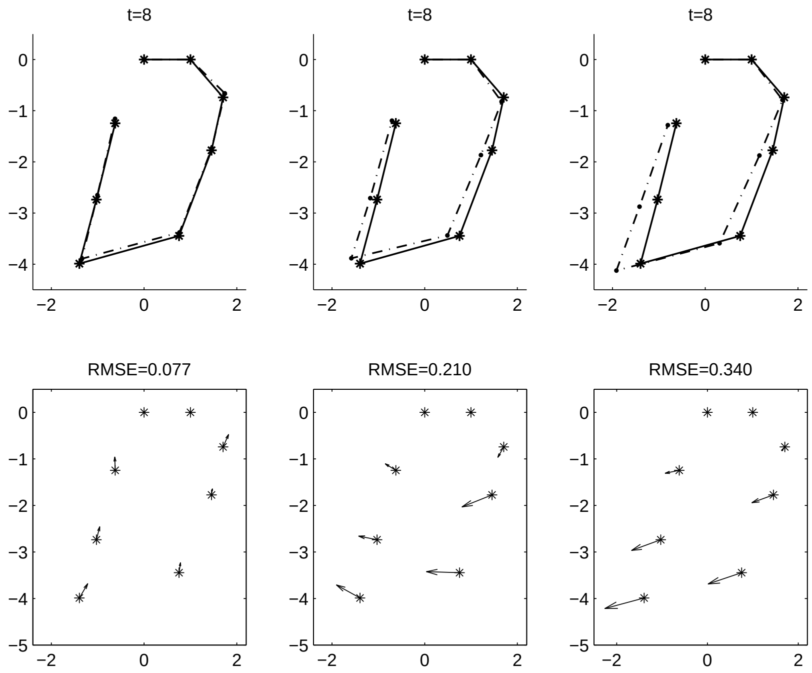


Figure 10: Forecasts (continuous line) of the rat shapes at the eighth age. The configurations are represented using Bookstein shape coordinates with baseline defined by points (0,0) and (1,0). The upper panels show the best (left) and worse (right) predictions corresponding to the minimum (0.077) and maximum (0.340) root mean squared prediction errors computed among the 18 rats. The forecast corresponding to the $75°$ percentile (0.210) is also represented in the middle panel. The bottom panels emphasize the magnitude of the prediction error at each landmark, with the length of each arrow (blown up for visibility) representing the distance between predicted and true coordinate shapes.