An Expectation Maximization Algorithm for Inferring Offset-Normal Shape Distributions

Max Welling
School of Information and Computer Science
University of California Irvine
Irvine CA 92697-3425 USA
welling@ics.uci.edu

Abstract

The statistical theory of shape plays a prominent role in applications such as object recognition and medical imaging. An important parameterized family of probability densities defined on the locations of landmark-points is given by the offset-normal shape distributions introduced in [7]. In this paper we present an EM algorithm for learning the parameters of the offset-normal shape distribution from shape data. To improve model flexibility we also provide an EM algorithm to learn mixtures of offset-normal distributions. To deal with missing landmarks (e.g. due to occlusions), we extend the algorithm to train on incomplete data-sets. The algorithm is tested on a number of real-world data sets and on some artificially generated data. Experimentally, this seems to be the first algorithm for which estimation of the full covariance matrix causes no difficulties. In all experiments the estimated distribution provided an excellent approximation to the true offset-normal shape distribution.

1 INTRODUCTION

The statistical analysis of shape has important applications in fields as diverse as biology, anatomy, genetics, medicine, archeology, geology, geography, agriculture, image analysis, computer vision, pattern recognition and chemistry (see e.g. [9]). As an important example, we can represent an object (e.g. a face, skull, etc.) as a collection of landmarks at certain positions (in figure space). To compare objects it is then useful to discard differences in location, orientation and scale. (i.e. their pose). The remaining degrees of freedom are called the *shape* of an object. For a meaningful comparison of objects by their shape we need the tools of "statistical shape analysis". For instance, we may want to know whether two objects are *significantly* different (using a hypothesis test), or we may be interested in classifying or clustering objects by their shape. The statistical analysis of shape has a long history dating back to the late seventies [15, 10, 11, 12, 1, 2, 3, 4].

The work that we will present here is based on a more recent development in statistical shape analysis, namely the introduction of the *offset-normal* distribution [14, 7, 8, 9]. Offset-normal probability densities describe the distribution of shapes as represented by collections of landmark points in two dimensions. The assumption is that the landmarks in figure space are normally distributed. Pose is removed by mapping two landmarks to fixed positions (e.g. $(0, 0)$ and $(1, 0)$), while the remaining landmarks represent the shape information. Perhaps surprisingly, this distribution over the remaining landmarks can be expressed in analytic form [7]. However, a reliable method to infer the distribution parameters from shape data in the most general case (full covariance matrix), is not available. The fact that certain singular normal distributions map to the same offset-normal shape distribution has obstructed the formulation of estimation procedures for general covariance matrices.

In this paper we will derive EM update rules for unrestricted parameters of the offset-normal shape distribution, i.e. a mean vector and a full covariance matrix. As it turns out, both E- and M-step can be computed analytically, providing an efficient update scheme. In pattern recognition, it may happen that landmarks are occluded. To deal with this difficulty which is often encountered in practical problems we extend the EM procedure to learn from incomplete data. For cases where the data are not well approximated by an offset-normal shape distribution, we provide EM-learning rules for *mixtures* of offset-normal shape distributions. We conclude with experiments on some real world data-sets.

2 THE OFFSET-NORMAL SHAPE DISTRIBUTION

In order to be self contained, we explain and re-derive the offset-normal shape distribution in this section. Some results in later sections will follow a similar derivation.

Let an object in two dimensions be represented by the positions $\{x_i, y_i\}$ of p landmarks. Let $\mathbf{x}$ be distributed according to a $2p$ dimensional normal distribution, $\mathbf{x} \sim \mathcal{N}_{2p}[\boldsymbol{\nu}, \boldsymbol{\Omega}]$.

We will first remove translational content by applying the following transformation,

$$\mathbf{x} = [x_1, ..., x_p, y_1, ..., y_p]^T \rightarrow \mathbf{L}\mathbf{x} \quad (1)$$

with

$$\mathbf{L} = \begin{bmatrix} \mathbf{I} - \mathbf{1}\mathbf{e}_1^T + \mathbf{e}_1\mathbf{e}_1^T & \mathbf{0} \\ \mathbf{0} & \mathbf{I} - \mathbf{1}\mathbf{e}_1^T + \mathbf{e}_1\mathbf{e}_1^T \end{bmatrix} \quad (2)$$

and integrate out the first landmark. In this equation $\mathbf{I}$ and $\mathbf{0}$ are the $p \times p$ dimensional identity and zero matrices respectively, $\mathbf{1}$ is a $p \times 1$ dimensional vector of ones and $\mathbf{e}_1$ is the $p \times 1$ dimensional vector $[1, 0, ..., 0]^T$. This transformation shifts all landmarks, except the first one, by an amount x_1, y_1. Notice that if we had also shifted the first landmark, it would be fixed at the location $(0, 0)$, producing a singular probability distribution. Since the above transformation is linear, the coordinates $\mathbf{Lx}$ are also normally distributed with mean $\boldsymbol{\mu} = \mathbf{L}\boldsymbol{\nu}$ and covariance $\boldsymbol{\Sigma} = \mathbf{L}\boldsymbol{\Omega}\mathbf{L}^T$. Integrating out x_1, y_1 for a normal distribution is simply accomplished by deleting the corresponding entries in the mean and covariance. The remaining coordinates are denoted by $\mathbf{x}^* = [x_2^*, ..., x_p^*, y_2^*, ..., y_p^*]^T$ and have dimension $2p - 2$.

Next, we remove rotation and scale content by following a similar procedure. First, we transform $\mathbf{x}^*$ as follows,

$$\begin{aligned} u_2 &= x_2^*, \qquad u_i = \frac{(x_i^* x_2^* + y_i^* y_2)}{{x_2^*}^2 + {y_2^*}^2} \quad i = 3, ..., p \\ v_2 &= y_2^*, \qquad v_i = \frac{(y_i^* x_2^* - x_i^* y_2)}{{x_2^*}^2 + {y_2^*}^2} \quad i = 3, ..., p \end{aligned} \quad (3)$$

This transformation would have moved the second landmark to the location $(1, 0)$, not allowing any spread and generating a singular pdf. Therefore, we will leave the second landmark untouched, while treating all the other ones as if the second landmark were moved to the reference position $(1, 0)$. Finally, to remove information on orientation and scale we need to integrate out the second landmark, which we will do in the following.

We will simplify notation for the second landmark by writing $\mathbf{x}_2^* = \mathbf{h}$, while $\mathbf{u} = [u_3, ...u_p, v_3, ..., v_p]^T$. In the coordinates $\{\mathbf{h}, \mathbf{u}\}$ the pdf is given by,

$$P(\mathbf{h}, \mathbf{u}) = \frac{1}{(2\pi)^{p-1}\sqrt{\det \boldsymbol{\Sigma}}} \exp[-\frac{1}{2}G] |\det \mathbf{J}|, \quad (4)$$

with,

$$\begin{aligned} G &= (\mathbf{W}\mathbf{h} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{W}\mathbf{h} - \boldsymbol{\mu}), \quad (5) \\ \det \mathbf{J} &= (h_x^2 + h_y^2)^{p-2}, \quad (6) \end{aligned}$$

where $\mathbf{J}$ is the Jacobian of the transformation (3) and

$$\mathbf{W}^T = \begin{bmatrix} 1 & u_3 & \cdots & u_p & 0 & v_3 & \cdots & v_p \\ 0 & -v_3 & \cdots & -v_p & 1 & u_3 & \cdots & u_p \end{bmatrix}. \quad (7)$$

The integration over $\mathbf{h}$ is facilitated by rewriting G as,

$$G = (\mathbf{h} - \boldsymbol{\xi})^T \boldsymbol{\Gamma}^{-1} (\mathbf{h} - \boldsymbol{\xi}) + g \quad (8)$$

with

$$\begin{aligned} \boldsymbol{\Gamma}^{-1} &= \mathbf{W}^T \boldsymbol{\Sigma}^{-1} \mathbf{W} \quad (9) \\ \boldsymbol{\xi} &= \boldsymbol{\Gamma} \mathbf{W}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \quad (10) \\ g &= \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} - \boldsymbol{\xi}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{\xi} \quad (11) \end{aligned}$$

We can simplify (4) further by transforming to the eigenbasis of $\boldsymbol{\Gamma}$,

$$\begin{aligned} &\boldsymbol{\Gamma} = \mathbf{R}\mathbf{D}\mathbf{R}^T, \\ &\boldsymbol{\zeta} = \mathbf{R}^T \boldsymbol{\xi} \qquad \mathbf{z} = \mathbf{R}^T \mathbf{h}. \end{aligned} \quad (12)$$

Noticing that the determinant of the Jacobian is invariant with respect to rotations, this gives,

$$\begin{aligned} P(\mathbf{z}, \mathbf{u}) = &\frac{1}{(2\pi)^{p-2}} \sqrt{\frac{\det \boldsymbol{\Gamma}}{\det \boldsymbol{\Sigma}}} e^{-\frac{1}{2}g} \times \\ &\times \mathcal{N}_{z_x}[\zeta_x, \sigma_x]\, \mathcal{N}_{z_y}[\zeta_y, \sigma_y]\, (z_x^2 + z_y^2)^{p-2} \end{aligned} \quad (13)$$

where

$$\sigma_x = \sqrt{D_{xx}} \qquad \sigma_y = \sqrt{D_{yy}} \quad (14)$$

Finally, we use the binomial expansion to rewrite the Jacobian as,

$$(z_x^2 + z_y^2)^{p-2} = \sum_{i=0}^{p-2} \binom{p-2}{i} z_x^{2i}\, z_y^{2p-4-2i}. \quad (15)$$

We are now ready to perform the integrations over $\mathbf{h}$, required for the definition of the offset-normal shape distribution,

$$\begin{aligned} P_{\mathcal{S}}(\mathbf{u}) = &\int d\mathbf{h}\, p(\mathbf{h}, \mathbf{u}) = \int d\mathbf{z}\, p(\mathbf{z}, \mathbf{u}) = \\ &\frac{1}{(2\pi)^{p-2}} \sqrt{\frac{\det \boldsymbol{\Gamma}}{\det \boldsymbol{\Sigma}}} e^{-\frac{1}{2}g} \times \\ &\times \sum_{i=0}^{p-2} \binom{p-2}{i} \mathbf{E}[z_x^{2i} | \zeta_x, \sigma_x]\, \mathbf{E}[z_y^{2p-4-2i} | \zeta_y, \sigma_y] \end{aligned} \quad (16)$$

where,

$$\mathbf{E}[z^k | \mu, \sigma] = \left(\frac{\sqrt{2}\sigma}{2i}\right)^k H_k\left(\frac{i\mu}{\sqrt{2}\sigma}\right), \quad (17)$$

denotes a Gaussian expectation and H_k denotes the Hermite polynomial of order k. Equation (17) is the offset-normal shape distribution [7], which is invariant with respect to translations, rotations and scalings of the data. It is expressed in terms of the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ which are not invariant with respect to orientation and scale changes (the translations were taken out in going from $\boldsymbol{\nu} \rightarrow \boldsymbol{\mu}, \boldsymbol{\Omega} \rightarrow \boldsymbol{\Sigma}$). It follows that the parameter set must be redundant, i.e. orientation and scale transformations of the parameters map to the same offset-normal shape distribution. Technically, this implies that the offset-normal shape distribution is described by an equivalence class of parameters. Therefore, when we mention in the rest of this paper that some random variable is distributed according to an offset-normal shape distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, we refer to the equivalence class of all $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ that map to the same offset-normal shape distribution. Sometimes it will be useful to remove this ambiguity by defining a canonical parameter set,

$$\begin{aligned} \boldsymbol{\mu}_c &= \mathbf{K}\boldsymbol{\mu} = [1, \mu_{3x}, ..., \mu_{px}, 0, \mu_{3y}, ..., \mu_{py}]^T, \\ \boldsymbol{\Sigma}_c &= \mathbf{K}\boldsymbol{\Sigma}\mathbf{K}^T, \end{aligned} \tag{18}$$

where the mean of the second landmark has been mapped to $(1, 0)$. More study is required to see for which offset-normal shape distributions the above transformation removes all redundancies and which have a still larger set of invariant transformations. It is important to notice the difference with the non-linear mapping (3). In contrast, (18) is a linear transform, depending on $\boldsymbol{\mu}_2$. In [7] it is observed that also some singular normal pdfs or even non-normal pdfs may map to the same offset-normal shape distribution, enlarging further the redundancy. In this paper we will not concern us with those.

Transformation Properties: We will now state two important properties of the offset-normal shape distribution, which will help us derive the learning algorithm in the subsequent sections.

Lemma 1 *Let* $\mathbf{x} = [x_1, ..., x_p, y_1, ..., y_p]^T$ *be a random variable distributed according to a normal distribution with parameters* $\boldsymbol{\nu}$ *and* $\boldsymbol{\Omega}$*, and let* $\mathbf{u} = [u_3, ..., u_p, v_3, ..., v_p]^T$ *be the corresponding shape random variable, distributed according to the offset-normal shape distribution with parameters*

$$\boldsymbol{\mu} = \mathbf{L}_{p-1}\boldsymbol{\nu}, \qquad \boldsymbol{\Sigma} = \mathbf{L}_{p-1}\boldsymbol{\Omega}\mathbf{L}_{p-1}^T, \tag{19}$$

where $\mathbf{L}_{p-1}$ *is the matrix defined in (2) with the* $1st$ *and* $(p+1)st$ *row deleted. The random variable* $\mathbf{x}' = [x'_1, ..., x'_p, y'_1, ..., y'_p]^T = \mathbf{G}\,\mathbf{x}$*, where* $\mathbf{G}$ *is a matrix of dimension* $2p \times 2p$*, will be distributed according to a* $2p$ *dimensional normal distribution with parameters* $\boldsymbol{\nu}' = \mathbf{G}\boldsymbol{\nu}$ *and* $\boldsymbol{\Omega}' = \mathbf{G}\boldsymbol{\Omega}\mathbf{G}^T$*. The corresponding shape random variables* $\mathbf{u}' = [u'_3, ..., u'_p, v'_3, ..., v'_p]^T$ *will be distributed according to an offset-normal shape distribution with parameters,*

$$\boldsymbol{\mu}' = \mathbf{L}_{p-1}\mathbf{G}\boldsymbol{\nu}, \qquad \boldsymbol{\Sigma}' = \mathbf{L}_{p-1}\mathbf{G}\boldsymbol{\Omega}\mathbf{G}^T\mathbf{L}_{p-1}^T. \tag{20}$$

The proof of this lemma is straightforward and relies on some well known properties of normal pdfs [5]. It is actually not necessary to assume that $\mathbf{G}$ is a square matrix. In general $\mathbf{G}$ can be of size $2g \times 2p$, where $2 < g \leq p$. This is useful if we want to integrate out variables from the offset-normal shape distribution [5].

Define the pair of baseline variables to be the ones which are mapped to $(0, 0)$ and $(1, 0)$. By choosing $\mathbf{G}$ to be a permutation matrix we can transform the offset-normal shape distribution between any pair of baseline variables in terms of the figure space parameters $\boldsymbol{\nu}$ and $\boldsymbol{\Omega}$. But does this still hold if we only have access to the parameters of the offset-normal shape distribution (i.e. the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$)? The following lemma answers this in the affirmative:

Lemma 2 *Let* $\mathbf{x} = [x_1, ..., x_p, y_1, ..., y_p]^T$ *be a random variable distributed according to a normal distribution with parameters* $\boldsymbol{\nu}$ *and* $\boldsymbol{\Omega}$*, and let* $\mathbf{u} = [u_3, ..., u_p, v_3, ..., v_p]^T$ *be the corresponding shape random variable, distributed according to the offset-normal shape distribution with parameters* $\boldsymbol{\mu}$ *and* $\boldsymbol{\Sigma}$*. Furthermore, let* $\mathbf{x}' = [x'_{\pi(1)}, ..., x'_{\pi(p)}, y'_{\pi(1)}, ..., y'_{\pi(p)}]^T = \mathbf{P}\mathbf{x}$ *be a permutation of* $\mathbf{x}$*, which is distributed according to a normal distribution with parameters* $\boldsymbol{\nu}' = \mathbf{P}\;\boldsymbol{\nu}$ *and* $\boldsymbol{\Omega}' = \mathbf{P}\;\boldsymbol{\Omega}\;\mathbf{P}^T$*. Then, the shape random variables* $\mathbf{u}' = [u'_{\pi(3)}, ..., u'_{\pi(p)}, v'_{\pi(3)}, ..., v'_{\pi(p)}]^T$ *are distributed according to an offset-normal shape distribution with parameters,*

$$\boldsymbol{\mu}' = \mathbf{B}\boldsymbol{\mu}, \qquad \boldsymbol{\Sigma}' = \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^T \qquad \mathbf{B} = \mathbf{L}_{p-1}\mathbf{P}\mathbf{E}, \tag{21}$$

Here $\mathbf{E}$ is the $2p \times 2p-2$ dimensional matrix, $\mathbf{E} = \begin{bmatrix} 0 & \cdots \\ \mathbf{I} & \mathbf{0} \\ 0 & \cdots \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$. This matrix has the effect of inserting zeros at the position of the first landmark, i.e. $\boldsymbol{\mu} \rightarrow [0, \boldsymbol{\mu}_x^T, 0, \boldsymbol{\mu}_y^T]^T$ and, $\boldsymbol{\Sigma} \rightarrow \begin{bmatrix} 0 & \cdots & 0 & \cdots \\ \vdots & \boldsymbol{\Sigma}_{xx} & \vdots & \boldsymbol{\Sigma}_{xy} \\ 0 & \cdots & 0 & \cdots \\ \vdots & \boldsymbol{\Sigma}_{yx} & \vdots & \boldsymbol{\Sigma}_{yy} \end{bmatrix}$

Proof of Lemma 2 *To prove this it we need to show that the following two transformations are equivalent:*

$$\mathbf{L}_{p-1}\mathbf{P} = \mathbf{B}\mathbf{L}_{p-1} \doteq \mathbf{L}_{p-1}\mathbf{P}\mathbf{E}\mathbf{L}_{p-1} \tag{22}$$

We will multiply left and right with the identity as follows,

$$\mathbf{E}^T\mathbf{E}\mathbf{L}_{p-1}\mathbf{P} = \mathbf{E}^T\mathbf{E}\mathbf{L}_{p-1}\mathbf{P}\mathbf{E}\mathbf{L}_{p-1} \qquad \mathbf{E}^T\mathbf{E} = \mathbf{I} \tag{23}$$

Next, we notice that we can rewrite the combination $\mathbf{E}\mathbf{L}_{p-1}$ *as,*

$$\mathbf{E}\mathbf{L}_{p-1} = \mathbf{I} - \mathbf{1}\mathbf{e}_1^T \tag{24}$$

i.e. it is the $2p \times 2p$ *dimensional matrix which translates the first landmark to the origin. Using this in eqn. 23 we find,*

$$\mathbf{E}^T\left(\mathbf{P} - \mathbf{1}\mathbf{e}_1^T\mathbf{P}\right) = \mathbf{E}^T\left(\mathbf{P} - \mathbf{1}\mathbf{e}_1^T\mathbf{P}\right)\left(\mathbf{I} - \mathbf{1}\mathbf{e}_1^T\right) \tag{25}$$

Writing this out and noting that $\mathbf{P}\mathbf{1}\mathbf{e}_1^T = \mathbf{1}\mathbf{e}_1^T$ *and* $\mathbf{1}\mathbf{e}_1^T\mathbf{1}\mathbf{e}_1^T = \mathbf{1}\mathbf{e}_1^T$*, we verify that the left hand side is indeed identical to the right hand side, which then proves the lemma.* □

The relevance of this lemma is that we can compute the offset-normal shape distribution for an arbitrary pair of baseline landmarks from the offset-normal shape-parameters of a given pair of baseline landmarks. This will allow us to estimate the parameters of the shape distribution, even if the data are presented in different reference frames; a situation which may occur if one of the baseline landmarks is occluded.

In the case where we only interchange the second landmark with higher labelled landmarks, leaving the first landmark in place, the lemma slightly simplifies. In that case, $\mathbf{PE} = \mathbf{EP}_{p-1}$, where $\mathbf{P}_{p-1}$ is $2p-2 \times 2p-2$ dimensional permutation matrix. Therefore, using, $\mathbf{L}_{p-1}\mathbf{E} = \mathbf{I}$, we may write instead of (21),

$$\boldsymbol{\mu}' = \mathbf{P}_{p-1}\boldsymbol{\mu}, \qquad \boldsymbol{\Sigma}' = \mathbf{P}_{p-1}\boldsymbol{\Sigma}\mathbf{P}_{p-1}^T, \tag{26}$$

3 EM LEARNING ALGORITHM

Our main objective is to find parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ (or $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$) that maximize the log-likelihood of the offset-normal shape distribution given a data-set $\{\mathbf{u}_n\} \quad n = 1...N$. The log-likelihood is given by,

$$\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{N}\sum_{n=1}^{N} \log P_{\mathcal{S}}(\mathbf{u}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{27}$$

Although the analytic form of the offset-normal shape distribution is quite complicated, the joint distribution $P(\mathbf{h}, \mathbf{u})$ is much simpler. Unfortunately, $\mathbf{h}$ is not observed and may be considered a hidden variable for that reason. This makes this estimation problem a school example of the expectation maximization (EM) algorithm. In the EM framework one iteratively optimizes the following family of objective functions (depending on the iteration k),

$$Q(k|k-1) = \frac{1}{N}\sum_{n=1}^{N} \int d\mathbf{h}\ P_{k-1}(\mathbf{h}|\mathbf{u}_n)\ \log P_k(\mathbf{h}, \mathbf{u}_n), \tag{28}$$

where $Q(k|k-1)$ depends on the parameters $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ at iteration k, *given* the parameters $\boldsymbol{\mu}_{k-1}$ and $\boldsymbol{\Sigma}_{k-1}$ at iteration $k-1$. Maximization of the log-likelihood is obtained by alternating an M-step where Q is maximized with respect to the parameters $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, and an E-step where the posterior distribution $p(\mathbf{h}|\mathbf{u}_n)$ is determined, given the new parameters calculated in the previous M-step.

M-step: In the M-step we need to maximize $\langle \log P(\mathbf{h}, \mathbf{u}_n)\rangle_n$ with respect to $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$. Here $\langle . \rangle_n$ denotes a posterior average, $\langle f(\mathbf{h})\rangle_n = \int d\mathbf{h}\ P(\mathbf{h}|\mathbf{u}_n) f(\mathbf{h})$. The derivatives are given by,

$$\frac{\partial}{\partial \boldsymbol{\mu}_k} Q(k|k-1) = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{\Sigma}_k^{-1}(\mathbf{W}_n\langle\mathbf{h}\rangle_n - \boldsymbol{\mu}_k) \tag{29}$$

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_k^{-1}} Q(k|k-1) = \tag{30}$$

$$\frac{1}{2}\frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{\Sigma}_k - \mathbf{W}_n\langle\mathbf{h}\mathbf{h}^T\rangle_n\mathbf{W}_n^T + 2\mathbf{W}_n\langle\mathbf{h}\rangle_n\boldsymbol{\mu}_k^T - \boldsymbol{\mu}_k\boldsymbol{\mu}_k^T)$$

resulting in the following simple update rules:

$$\boldsymbol{\mu}_k = \frac{1}{N}\sum_{n=1}^{N} \mathbf{W}_n\langle\mathbf{h}\rangle_n \tag{31}$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N}\sum_{n=1}^{N} \mathbf{W}_n\langle\mathbf{h}\mathbf{h}^T\rangle_n\mathbf{W}_n^T - \boldsymbol{\mu}_k\boldsymbol{\mu}_k^T \tag{32}$$

After every M-step we also map the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ to the canonical parameters $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$, defined in (18), to avoid drifting. Because the offset-normal shape distribution is invariant with respect to this transformation, the log-likelihood will not change either.

E-step: In the E-step we need to calculate the mean $\langle\mathbf{h}\rangle_n$ and covariance $\langle\mathbf{h}\mathbf{h}^T\rangle_n$ of the posterior distribution $P(\mathbf{h}|\mathbf{u}_n)$. Using Bayes rule it is easily found that,

$$P(\mathbf{h}|\mathbf{u}_n) = \frac{P(\mathbf{h}, \mathbf{u}_n)}{P_{\mathcal{S}}(\mathbf{u}_n)}, \tag{33}$$

where $P_{\mathcal{S}}(\mathbf{u}_n)$ is simply the offset-normal shape distribution evaluated at $\mathbf{u}_n$. Calculation of the sufficient statistics thus involves the following integrals,

$$\langle\mathbf{h}\rangle_n = \frac{1}{P_{\mathcal{S}}(\mathbf{u}_n)}\int d\mathbf{h}\ \mathbf{h}\ P(\mathbf{h}, \mathbf{u}_n) \tag{34}$$

$$\langle\mathbf{h}\mathbf{h}^T\rangle_n = \frac{1}{P_{\mathcal{S}}(\mathbf{u}_n)}\int d\mathbf{h}\ \mathbf{h}\mathbf{h}^T\ P(\mathbf{h}, \mathbf{u}_n) \tag{35}$$

These integrals can be solved following the same strategy as the one used to calculate the offset-normal shape distribution in section 2. Again, we will transform to the $\mathbf{z}$ coordinates defined in (12) and notice that,

$$\langle\mathbf{h}\rangle_n = \mathbf{R}_n\langle\mathbf{z}\rangle_n, \tag{36}$$

$$\langle\mathbf{h}\mathbf{h}^T\rangle_n = \mathbf{R}_n\langle\mathbf{z}\mathbf{z}^T\rangle_n\mathbf{R}_n^T. \tag{37}$$

Using the binomial expansion (15) and the result (17) we can calculate the following posterior averages,

$$\langle z_x^a z_y^b\rangle_n = \frac{\sum_{i=0}^{p-2}\begin{pmatrix} p-2 \\ i \end{pmatrix} \mathbf{E}_n[z_x^{2i+a}]\ \mathbf{E}_n[z_y^{2p-4-2i+b}]}{\sum_{j=0}^{p-2}\begin{pmatrix} p-2 \\ j \end{pmatrix} \mathbf{E}_n[z_x^{2j}]\ \mathbf{E}_n[z_y^{2p-4-2j}]} \tag{38}$$

Using (38) for the pairs $\{(a,b) = (1,0),(0,1),(2,0),(1,1),(0,2)\}$ allows us to perform the E-step.

Initialization: To initialize the parameters we use the approximation described in [7]. If the variances of the landmarks are small compared to the mean length of the baseline, then the offset-normal shape distribution becomes similar to a normal distribution with mean $\boldsymbol{\lambda} = [\mu_{c3x}, ..., \mu_{cpx}, \mu_{c3y}, ..., \mu_{cpy}]^T$ and covariance $\boldsymbol{\Lambda} = \mathbf{F}\boldsymbol{\Sigma}_c\mathbf{F}^T$, where $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ are are the canonical parameters and $\mathbf{F}$ is the $2p-4 \times 2p-2$ dimensional matrix,

$$\mathbf{F} = \begin{bmatrix} \lambda_1 & \mathbf{I} & \gamma_1 & \mathbf{0} \\ \vdots & & \vdots & \\ \lambda_{2p-4} & \mathbf{0} & \gamma_{2p-4} & \mathbf{I} \end{bmatrix}$$

$$\boldsymbol{\gamma} = [\mu_{3y}, ..., \mu_{py}, -\mu_{3x}, ..., -\mu_{px}]^T \tag{39}$$

To initialize our algorithm we therefore calculate the sample mean and covariance of the shape data, $\boldsymbol{\lambda} = \frac{1}{N}\sum_{n=1}^N \mathbf{u}_n$ and $\boldsymbol{\Lambda} = \frac{1}{N-1}\sum_{n=1}^N(\mathbf{u}_n - \boldsymbol{\lambda})(\mathbf{u}_n - \boldsymbol{\lambda})^T$. The initial values of the mean $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are then given by,

$$\boldsymbol{\mu} = [1, \boldsymbol{\lambda}_x, 0, \boldsymbol{\lambda}_y] \tag{40}$$

$$\boldsymbol{\Sigma} = \mathbf{F}_+\boldsymbol{\Lambda}\mathbf{F}_+^T, \qquad \mathbf{F}_+ = \mathbf{F}^T(\mathbf{F}\mathbf{F}^T)^{-1} \tag{41}$$

where $\mathbf{F}_+$ is the pseudo-inverse of $\mathbf{F}$.

4 MIXTURE DISTRIBUTIONS

In practice it might happen that the data in figure-space are not well described by a normal distribution. In that case, we may approximate it by a mixture of Gaussians. The corresponding distribution in shape-space turns out to be a mixture of offset-normal shape distributions according to the following lemma [5]

Lemma 3 *Under a multivariate normal mixture model for the figure-space coordinates,*

$$P_{\mathbf{MoG}}(\mathbf{x}) = \sum_{a=1}^M \mathcal{N}_{2p}[\mathbf{x}|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a]\ \pi_a, \tag{42}$$

the joint probability distribution function of the shape vector $\mathbf{u}$ *is a mixture of offset-normal shape distributions,*

$$P_{\mathbf{Mo}\mathcal{S}}(\mathbf{u}) = \sum_{a=1}^M P_{\mathcal{S}}[\mathbf{u}|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a]\ \pi_a \tag{43}$$

The proof is simple if one realizes that every mixture component is mapped to an offset-normal shape distribution, which are then combined using the a priori probabilities π_a. To find update rules for the parameters π_a, $\boldsymbol{\mu}_a$ and $\boldsymbol{\Sigma}_a$ we start with the log-likelihood,

$$\mathcal{L} = \frac{1}{N}\sum_{n=1}^N \log \sum_{a=1}^M P_{\mathcal{S}}(\mathbf{u}_n|a; \boldsymbol{\mu}^a, \boldsymbol{\Sigma}^a)\ \pi_a. \tag{44}$$

We will consider the labels a and the variables $\mathbf{h}$ *hidden*. The function to be iteratively maximized is therefore given by,

$$Q(k|k-1) = \tag{45}$$

$$\frac{1}{N}\sum_{n=1}^N\sum_{a=1}^M \int d\mathbf{h}\ P_{k-1}(a, \mathbf{h}|\mathbf{u}_n)\ \log\{P_k(\mathbf{h}, \mathbf{u}_n|a)\ \pi_k^a\} =$$

$$\frac{1}{N}\sum_{n=1}^N\sum_{a=1}^M P_{k-1}(a|\mathbf{u}_n) \int d\mathbf{h}\ P_{k-1}(\mathbf{h}|\mathbf{u}_n, a) \times$$

$$\times \{\log P_k(\mathbf{u}_n, \mathbf{h}|a) + \log \pi_k^a\},$$

where we used, $P(\mathbf{h}, a|\mathbf{u}) = P(a|\mathbf{u})\ P(\mathbf{h}|\mathbf{u}, a)$. The M-step involves again maximizing this expression at every iteration with respect to π_k^a, $\boldsymbol{\mu}_k^a$ and $\boldsymbol{\Sigma}_k^a$. Taking derivatives with respect to these variables and equating them to zero we find,

$$\pi_k^a = \frac{1}{N}\sum_{n=1}^N P_{k-1}(a|\mathbf{u}_n), \tag{46}$$

$$\boldsymbol{\mu}_k^a = \frac{\sum_{n=1}^N P_{k-1}(a|\mathbf{u}_n)\ \mathbf{W}_n\langle\mathbf{h}\rangle_n^a}{\sum_{m=1}^N P_{k-1}(a|\mathbf{u}_m)}, \tag{47}$$

$$\boldsymbol{\Sigma}_k^a = \frac{\sum_{n=1}^N P_{k-1}(a|\mathbf{u}_n)\ \mathbf{W}_n\langle\mathbf{h}\mathbf{h}^T\rangle_n^a\mathbf{W}_n^T}{\sum_{m=1}^N P_{k-1}(a|\mathbf{u}_m)} - \boldsymbol{\mu}_k^a\boldsymbol{\mu}_k^{aT}, \tag{48}$$

where we have defined,

$$\langle f(\mathbf{h})\rangle_n^a = \int d\mathbf{h}\ P(\mathbf{h}|\mathbf{u}_n, a) f(\mathbf{h}). \tag{49}$$

These update rules are very similar to (31) and (32). In the mixture case however, the influence of every data point on $\boldsymbol{\mu}_k^a$ and $\boldsymbol{\Sigma}_k^a$ is weighted by a factor $\frac{P_{k-1}(a|\mathbf{u}_n)}{\sum_{m=1}^N P_{k-1}(a|\mathbf{u}_m)}$ which expresses the probability that mixture component $P(\mathbf{u}_n|a)$ is responsible for the generation of datum $\mathbf{u}_n$.

The E-step involves the calculation of $P(a|\mathbf{u}_n)$, $\langle\mathbf{h}\rangle_n^a$ and $\langle\mathbf{h}\mathbf{h}^T\rangle_n^a$. $P(a|\mathbf{u}_n)$ is simply given by,

$$P(a|\mathbf{u}_n) = \frac{P_{\mathcal{S}}(\mathbf{u}_n|a)\ \pi_a}{\sum_{b=1}^M P_{\mathcal{S}}(\mathbf{u}_n|b)\ \pi_b}, \tag{50}$$

where $P_{\mathcal{S}}(\mathbf{u}_n|a)$ is an offset-normal shape distribution with parameters $\boldsymbol{\mu}^a$ and $\boldsymbol{\Sigma}^a$. According to (49), the calculation of $\langle\mathbf{h}\rangle_n^a$ and $\langle\mathbf{h}\mathbf{h}^T\rangle_n^a$ is identical to those described in section 3, where we use parameters $\boldsymbol{\mu}^a$ and $\boldsymbol{\Sigma}^a$ for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. We thus see that the learning rules for a mixture of offset-normal shape distributions are straightforward generalizations of the one component learning rules.

5 INCOMPLETE DATA

In practice, it may happen that landmarks are occluded and only incomplete data are provided. First, we will assume that the missing information does not concern the baseline points (i.e. landmarks 1 and 2). This will be generalized to arbitrary missing landmarks later in this section.

Assume that we have N, possibly incomplete samples, $\{\mathbf{u}_n\},\ n = 1...N$. For every sample we define an index m denoting the missing dimensions, and an index o denoting the observed dimensions. We will always assume that both the x and the y component of a landmark are missing, implying that m and o are necessarily even dimensional. We thus have $\mathbf{u}_n = [\mathbf{u}_n^m, \mathbf{u}_n^o]^T$ (the dependence of m and o on n is omitted for notational convenience). The question we want to answer is; *Can we use the information of incomplete data-vectors in the estimation of the parameters of the offset-normal shape distribution ?* To answer this, we first write the log-likelihood,

$$\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{N} \sum_{n=1}^{N} \log P_{\mathcal{S}}(\mathbf{u}_n^o | \boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{51}$$

which now only depends on the observed data. This implies that we may treat the missing dimensions as hidden variables, alongside the variable $\mathbf{x}_2^*$. Thus, for every n, we have a different set of hidden variables, denoted by $\mathbf{h}_n = [\mathbf{x}_{2,n}^*, \mathbf{u}_n^m]$. In fact, it turns out to be more convenient to represent the unobserved landmarks in figure space, so that the set of missing variables becomes $\mathbf{h}_n = [\mathbf{x}_{2,n}^*, \mathbf{x}_n^{*m}]$. The auxiliary functions $Q(k|k-1)$ in terms of the above variables are given by,

$$Q(k|k-1) = \frac{1}{N} \sum_{n=1}^{N} \int d\mathbf{h}\ P_{k-1}(\mathbf{h}|\mathbf{u}_n^o)\ \log P_k(\mathbf{h}, \mathbf{u}_n^o). \tag{52}$$

The formula for $P(\mathbf{h}, \mathbf{u}^o)$ is very similar to (4) with 2 important differences. Firstly, since more variables are defined in figure space, the Jacobian of the transformation is slightly different,

$$|\det \mathbf{J}| = (x_2^{*2} + y_2^{*2})^{p-2-q}, \tag{53}$$

where q denotes the number of missing landmarks (which may be different for each data case n). Assuming for a moment that the missing dimensions have the lowest indices (i.e. $m = 3, 4, \ldots$), we define,

$$\mathbf{W}^T = \begin{bmatrix} 1 & 0 & 0 & \cdots & u_{q+1} & \cdots & u_p & 0 & 0 & 0 & \cdots & \cdots & v_{q+1} & \cdots & v_p \\ 0 & 1 & 0 & \cdots & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & \cdots & 0 & \cdots & 0 \\ \vdots & & & & & & \vdots & & & & & & & & \vdots \\ 0 & 0 & 0 & \cdots & -v_{q+1} & \cdots & -v_p & 1 & 0 & 0 & \cdots & \cdots & u_{q+1} & \cdots & u_p \\ 0 & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & 1 & 0 & \cdots & \cdots & 0 & \cdots & 0 \\ \vdots & & & & & & \vdots & & & & & & & & \vdots \end{bmatrix} \tag{54}$$

To generalize this to arbitrary missing dimensions we simply need to permute the columns of $\mathbf{W}^T$.

The M-step of the EM algorithm proceeds exactly as explained in section (3), where averages are now taken w.r.t. the posterior distribution $P(\mathbf{h}|\mathbf{u}_n^o)$. Evaluating these averages, which is part of the E-step, proceeds analogously as in section (3). Using equations (34) and (35) we note that the difficult part of that calculation is computing the following expectations,

$$\int d\mathbf{h}\, f(\mathbf{h}) P(\mathbf{h}, \mathbf{u}^o) = C\, \mathbf{E}[f(\mathbf{h})\ (h_1^2 + h_{q+1}^2)^{2-p+q} |\ \boldsymbol{\xi}, \boldsymbol{\Gamma}], \tag{55}$$

where $\mathbf{E}[.\ |\ \boldsymbol{\xi}, \boldsymbol{\Gamma}]$ denotes taking the average over a multivariate normal pdf with mean $\boldsymbol{\xi}$ and covariance $\boldsymbol{\Gamma}$ and $f(\mathbf{h}) = \mathbf{h}$ or $f(\mathbf{h}) = \mathbf{h}\mathbf{h}^T$. Unfortunately, the transformation in eqn. (12) will not leave the Jacobian invariant, since

$$h_1^2 + h_{q+1}^2 = \mathbf{h}^T \boldsymbol{\Omega} \mathbf{h} \qquad \boldsymbol{\Omega} = \mathbf{e}_1 \mathbf{e}_1^T + \mathbf{e}_{q+1} \mathbf{e}_{q+1}^T \tag{56}$$

is not invariant with respect to $\mathbf{h} \to \mathbf{z} = \mathbf{R}^T \mathbf{h}$. However, if we transform,

$$\begin{aligned} \boldsymbol{\Gamma} &= \mathbf{R}\mathbf{D}\mathbf{R}^T \doteq \mathbf{F}\mathbf{F}^T \\ \boldsymbol{\zeta} &= = \mathbf{F}^{-1}\boldsymbol{\xi} = \mathbf{U}\mathbf{D}^{-\frac{1}{2}}\mathbf{R}^T\boldsymbol{\xi} \\ \mathbf{z} &= \mathbf{F}^{-1}\mathbf{h} \end{aligned} \tag{57}$$

then the normal distribution transforms to, $\mathcal{N}_{\mathbf{h}}[\boldsymbol{\xi}, \boldsymbol{\Gamma}] \to \mathcal{N}_{\mathbf{z}}[\boldsymbol{\zeta}, \mathbf{I}]$ while we can still choose the orthonormal matrix $\mathbf{U}$ such that the Jacobian remains as simple as possible,

$$\mathbf{h}^T \boldsymbol{\Omega} \mathbf{h} = \mathbf{z}^T \mathbf{F}^T \boldsymbol{\Omega} \mathbf{F} \mathbf{z} = \mathbf{z}^T \boldsymbol{\Lambda} \mathbf{z} \tag{58}$$

The matrix $\boldsymbol{\Lambda}$ can be chosen diagonal by using the following eigenvalue decomposition, $\mathbf{F}^T \boldsymbol{\Omega} \mathbf{F} = \mathbf{V}\mathbf{H}\mathbf{V}^T$ which is always possible because $\mathbf{F}^T \boldsymbol{\Omega} \mathbf{F}$ is a symmetric rank-2 matrix. Thus, by choosing $\boldsymbol{\Lambda} = \mathbf{H}$ and $\mathbf{U} = \mathbf{V}^T$ we obtain the desired result. We now need to expand the Jacobian in a binomial series expansion and use eqns. (34) and (35) to arrive at an expression for the desired averages similar to eqn. (38).

Alternatively, a good approximation can be obtained by sampling from the normal distribution $\mathcal{N}[\mathbf{h}\mid\boldsymbol{\xi}, \boldsymbol{\Gamma}]$ and subsequent calculation of the sample average.

Missing Baseline Landmarks: Next, we treat the case where one or both of the baseline landmarks is missing from the data. For such a data case, the locations of the other landmarks should be represented in a different reference frame, i.e. using a different (observed) baseline pair. In that frame, the situation reduces to the case treated above. It remains to be understood how to incorporate data in different reference frames in the estimation process. We will first choose one, arbitrary, baseline pair and invoke lemma 2 (section 2) to write the distribution in any other frame as,

$$P_{\mathcal{S}}(\mathbf{u}'|\,\boldsymbol{\mu}',\boldsymbol{\Sigma}') = P_{\mathcal{S}}(\mathbf{u}'|\mathbf{B}\boldsymbol{\mu},\mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^T), \qquad (59)$$

where $\mathbf{B} = \mathbf{L}_{p-1}\mathbf{P}\mathbf{E}$ and $\mathbf{L}_{p-1}$, $\mathbf{P}$ and $\mathbf{E}$ are defined in section 2. Since every data point may be defined in a different reference frame, $\mathbf{B}$ depends on n. Taking derivatives with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in the M-step then generates the following update rules,

$$\boldsymbol{\mu}_k = \frac{1}{N}\sum_{n=1}^{N}\mathbf{B}_n^{-1}\mathbf{W}_n\langle\mathbf{h}\rangle_n \qquad (60)$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N}\sum_{n=1}^{N}\mathbf{B}_n^{-1}\mathbf{W}_n\langle\mathbf{h}\mathbf{h}^T\rangle_n\mathbf{W}_n^T\mathbf{B}_n^{-T} - \boldsymbol{\mu}_k\boldsymbol{\mu}_k^T, \qquad (61)$$

where $\mathbf{W}_n$ and $\langle . \rangle_n$ are defined in their own reference frame.

In the E-step we compute the posterior mean and covariance as usual, using parameters $\boldsymbol{\mu}_n' = \mathbf{B}_n\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_n' = \mathbf{B}_n\boldsymbol{\Sigma}\mathbf{B}_n^T$ for data case n.

6 EXPERIMENTS

To test the algorithm on real world data, we downloaded 5 data-sets from the web [1]. Some data-sets contain data directly in shape space, while others have data in figure space, which we converted to shape space by mapping two landmarks to locations $(0,0)$ and $(1,0)$ respectively. Before transforming to shape space we extracted the sample mean and covariance to establish 'ground truth', since these are the parameters which describe the offset-normal shape distribution. Note however that many different normal distributions map to the same offset-normal shape distribution, so that comparing the parameters directly is not very meaningful.

Figure 1 shows the results when the sample mean and covariance were available in figure space. The data-sets used in Figure 1 are "Brizalina", 'Globorotalia' (described in [4]) and 'Mouse vertebrae' (Small group) (described in [9]). Figure 2 shows the results on the datasets 'Gorilla skulls' (female) (described in [9]) and 'Rat calvarial growth' (studied in [4]). These data-sets are defined in shape space, which implies that we have no access to ground truth. Finally, in figure 3, we present an example where we artificially generated 100 samples from a 'challenging' offset-normal shape distribution.

The algorithm usually converges within 20 iterations. Notice however, that for every data-point a SVD needs to

[1]The data-sets can be found at:
http://www.amsta.leeds.ac.uk/iand/Shape-S/datasets.html
http://life.bio.sunysb.edu/morph/index.html

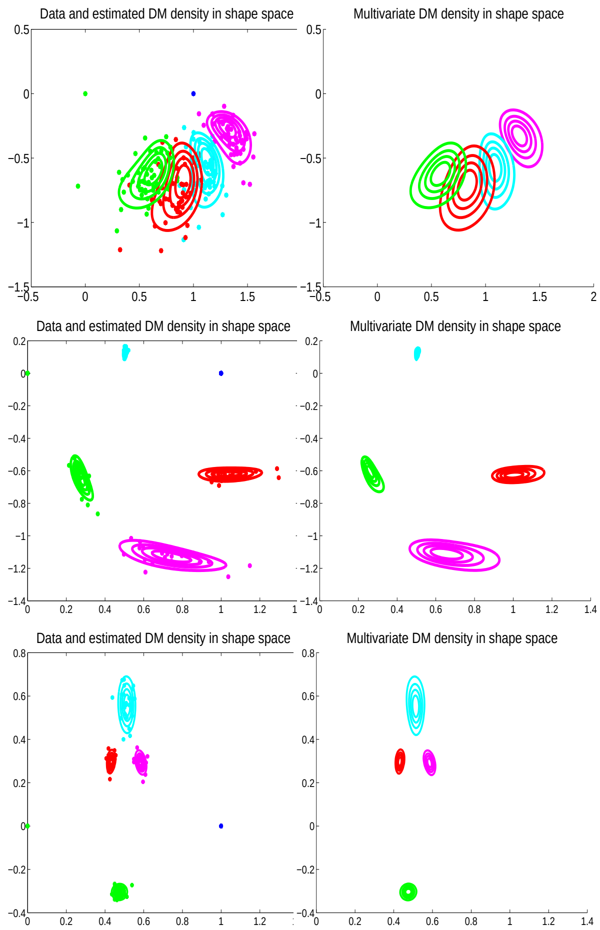

Figure 1: Estimation of offset-normal shape distributions for the following data-sets provided in figure space (from top to bottom): "Brizalina", "Globorotalia" and "Mouse vertebrae (small group)". The first column depicts the data overlaid with the offset-normal distributions estimated in shape space, while the second column shows the offset-normal distributions estimated in figure space.

be computed, resulting in unfavorable scaling behavior for large amounts of data.

We have encountered no problems in the estimation of the full covariance matrix, as described in [7], [9]. Also, few data are needed to find a reliable estimate of the distribution (around 20).

7 DISCUSSION

In this paper we have shown how to infer the parameters of a full covariance offset-normal shape distribution using the expectation maximization algorithm. In addition, we have addressed to important issues which open the door to practical applications. Firstly, the data may not be well described by an offset-normal shape distribution and secondly, the data may be incomplete, e.g. due to occlusion. The first problem was addressed by providing a learning

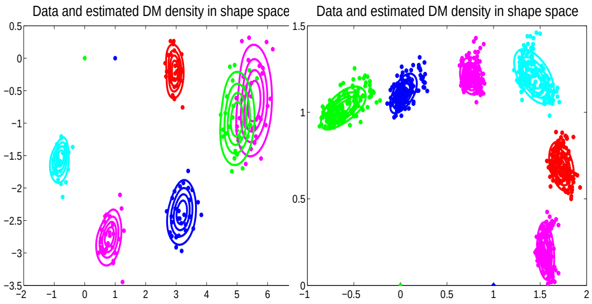

Figure 2: Estimation of offset-normal shape distributions for the following data-sets provided in shape space (left to right): "Gorilla Skulls (female)" and "Rat calvarial growth" (small group). These data-sets are only provided in shape-space, so no figure space estimates are available.

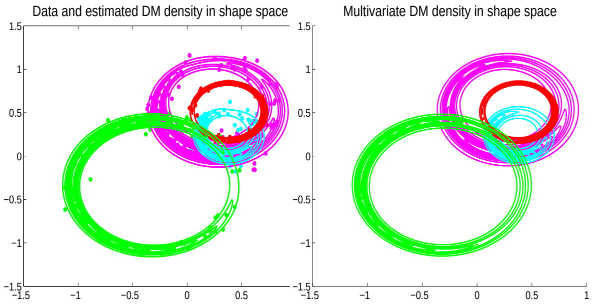

Figure 3: As in figure 1 with artificially generated data.

algorithm for mixtures of offset-normal shape distributions which improves model flexibility. The second issue was addressed by showing how to incorporate incomplete data into the estimation process.

We think the presented learning algorithms could find important applications in the field of object (class) and pattern recognition. In [6] a face recognition system was proposed where the geometry of certain feature detectors (e.g. eye-corner, nose) was described by the offset-normal shape distribution. This model also accounts for uncertainties in the labelling and the positions of the landmarks. The parameters of that model were determined in *figure* space. This was possible only because the data were acquired under carefully controlled circumstances. In more realistic situations, we want to learn the model using (possibly incomplete) shape data, which is precisely the topic of the present paper.

An important generalization of the offset-normal shape distribution is the affine invariant shape distribution proposed in [13]. There, a third landmark is mapped to a fixed position (e.g. $(x, y) = (0, 1)$), rendering the resulting distribution invariant with respect to affine transformations. The presented EM algorithm is easily extended to cover that case as well, which will be described in a future publication.

ACKNOWLEDGEMENTS

We thank CNSE and the sloan foundation for financial support. We are also grateful for discussions with Pietro Perona, Mike Burl and Markus Weber.

References

[1] F.L. Bookstein. *Lecture Notes on Biomathematics, Vol. 24*. Springer Verlag, 1978.

[2] F.L. Bookstein. A statistical method for biological shape comparison. *J. Theor. Biol.*, 107:475–520, 1984.

[3] F.L. Bookstein. Size and shape spaces for landmark data in two dimensions. *Statistical Science*, 1(2):181–242, 1986.

[4] F.L. Bookstein. *Morphometric tools for landmark data*. Cambridge University Press, 1991.

[5] M.C. Burl. *Recognition of visual object classes*. PhD thesis, Department of Electrical Engineering, California Institute of Technology, Pasadena, CA, 1997.

[6] M.C. Burl, T.K. Leung, and P. Perona. "Face Localization via Shape Statistics". In *Int Workshop on Automatic Face and Gesture Recognition*, 1995.

[7] I.L. Dryden and K.V. Mardia. General shape distributions in a plane. *Advanced Applied Probability*, 23:259–276, 1991.

[8] I.L. Dryden and K.V. Mardia. Size and shape analysis of landmark data. *Biometrika*, 79:57–68, 1992.

[9] I.L. Dryden and K.V. Mardia. *Statistical shape analysis*. Wiley, 1998.

[10] D.G. Kendall. The diffusion of shape. *Advances Applied Probability*, 9:428–430, 1977.

[11] D.G. Kendall. Shape manifolds, procrustean metrics, and complex projective spaces. *Bull. London Math Soc.*, 16:81–121, 1984.

[12] D.G. Kendall. A survey of the statistical theory of shape. *Statistical Science*, 4(2):87–120, 1989.

[13] T.K. leung, M.C. Burl, and P. Perona. Probabilistic affine invariants for recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 1998.

[14] K.V. Mardia and I.L. Dryden. Shape distributions for landmark data. *Adv. Appl. Prob.*, 21:742–755, 1989.

[15] H. Ziezold. On expected figures and a strong law of large numbers for random elements in quasi-metric spaces. In *Trans. 7th Praque Conf. Information Theory, Statistical Decision Functions, Random Processses and of the 1974 European Meeting of Statisticians*, volume A, pages 591–602, Prague, 2000.