Tina Memo No. 2010-009
Internal Report

Shape Analysis using Measurement Covariances

H. Ragheb, N.A. Thacker

Last updated
22 / 12 / 2010

ISBE, Medical School,
University of Manchester,
Stopford Building, Oxford Road,
Manchester, M13 9PT, UK.

Shape Analysis using Measurement Covariances

H. Ragheb and N.A. Thacker
ISBE, Medical School,
University of Manchester,
Manchester, M13 9PT, UK.
hossein.ragheb@manchester.ac.uk

Abstract

Conventional methods for shape analysis, based upon Procrustes and PCA, seem incapable of dealing with 'non-landmark' features, meaning measured locations not associated with well defined locations. We argue here that this is due to an assumption of homogenous errors, associated with an attempt to extract linear models with biologically meaningful descriptions. This document contains the mathematical definition of a shape analysis system based upon the description of landmarks with measurement covariance which will extend the modelling process to 'pseudo-landmarks' such as boundaries and surfaces [1]. As this breaks with convention we discuss the properties of this approach and how these covariances can be considered characteristic of the local shape. The idea has been implemented in software and is now being tested on measurements from fly wings. We will use these data to explore possible advantages and disadvantages over the use of Procrustes/PCA.

1 Introduction

Over a decade during 70's, bio-mathematical and biometrical aspects of biological shape studies were treated separately. This early work was later criticised during 80's by Bookstein [1], Goodall [2] and Kendall [3]. Later Bookstein [5] worked towards converging of these notations, by Goodall, Kendall and himself, upon one single foundation for the biometric analysis of landmark data in a bio-mathematically interpretable framework of shape processes. As a consequence of these efforts, the standard method for analysis of variation in landmark position is generally regarded as "Procrustes". It comprises a least squares alignment of a set of landmark features to a mean shape, and this is often followed by eigen-vector analysis of the linear correlations in variation around that mean. While the technique is now very popular the approach has several limitations with regard to the types of variation with which it can deal. One of these limitations is due to the assumption associated with taking least-squares differences and eigen vector summaries of distributions. Though many regard these as simply definitional, and in particular associated with "shape", any statistical interpretation suggests that data are measures with homogenous noise. It is this statistical interpretation which we intend to investigate here.

Although landmarks are generally carefully chosen in order to allow accurate measurements of position within the image, problems will occur if 'pseudo landmarks', measured from smooth curves and only accurately localised in one dimension (1D), are input to the analysis. Landmarks with a high degree of variability can act as outliers in the alignment stage, generating correlated compensating shifts and rotations of the other points. As PCA aims to describe the main sources of variation, high levels of such correlated movement will then necessarily contaminate the extraction of eigen-vectors. This contamination cannot be considered a genuine biological variation, as it has occurred purely due to the uncertainty in the measurement. This in turn follows from the subjective definition of the landmark leading to the view that problems can be avoided via appropriate definition. The mathematical concept of homology (and mapping) underlies many of the considerations behind much theoretical work which is described with the mathematical formalisms of iso-morphism. Because of such restrictions on the definition of landmarks, semi-landmarks were introduced [6] in order to allow inclusion of other points which are not homologous among the specimens. By this we mean that a unique corresponding location can not be uniquely defined. Measurement at these locations must be augmented by a constraint, such as bending energy [8, 6], in order to recover the information missing due to the nature of local structure. However, it is easy to construct examples where such resolution of corresponding locations is not biologically meaningful.

From a statistical perspective an homology must be augmented by distributions indicative of the extent to which a correspondence can be established. The standard way to deal with inappropriate weighting of data in a least-squares fit is to introduce weighting, or more generally to generalise the least-squares cost to a Mahalanobis distance,

[1] This work was funded by the Max Planck Institute for Evolutionary Biology [14].

computed using measurement covariances. By avoiding the requirement of specifying a unique homologous location, this has the advantage of accommodating varying information in measured data without having to try to re-create missing data.

There are a considerable number of recommendations in the morphometrics literature targeting biologists. When techniques start becoming complex, there is a tendency to suggest simpler but inconsistent statistical solutions instead. As an example, we focus on the statements regarding semi-landmarks in a well know text book in this area [9] (page 399). Writing about the possibility of using weighted Procrustes in order to reduce the effects of semi-landmarks, the authors stated that such an approach does not lead to Kendall's shape space. Claiming that "statistical analysis cannot employ parametric models", they suggested that resampling-based methods must be used instead. Another reason for rejecting the idea of a weighted Procrustes was said to be a "lack of clear criteria for determining appropriate weighting of semi-landmarks". These criticisms can only really be interpreted once a method for weighting is specified. Goodall [10] (page 301) suggested a method which assumed a single covariance for all landmark perturbation. It has been noted that such a matrix is inestimable [11]. Goodall himself acknowledged that "as a model of measurement error this is a drawback, as the direction of greatest variation may vary considerably between landmarks". Despite this problem, later work [12] generalised this idea to a Bayesian framework. We believe that it makes sense instead to suggest an approach which can support the process of landmark location as measurement, with a covariance describing the localisation of each landmark separately. We will now investigate possible generalisations of Procrustes along these lines, and show one of several ways that such a measurement covariance can be estimated. As a key issue here is the computability of these covariances, the stability of the resulting analysis is investigated using repeat mark-up data for a range of sample sizes. In order to explain the relevance of these ideas to conventional notions of shape we introduce the concept of a ghost point, as a procedure for defining landmark locations such that they have isotropic measurement error.

2 Detailed Method

Suppose that there are K shapes (2D) in our data-set and each shape vector $\mathbf{w}_k$ contains N landmark points, i.e. $\mathbf{w}_k = [w_{1x}, w_{1y}, w_{2x}, w_{2y}, ..., w_{Nx}, w_{Ny}]$. We then apply a scale s_k, a rotation R_k and a translation $\mathbf{t}_k$ to the original data to get an aligned version of the data called $\mathbf{z}_k$.

$$\mathbf{z}_{kn} = s_k R_k(\mathbf{w}_{kn} - \mathbf{t}_k) \tag{1}$$

The mathematical description of the model so far can accommodate any value of scale or orientation for the definition of mean model. We therefore define the orientation of mean shape so that the line between a specified pair of points is horizontal. This also has the benefit that initial estimates of alignment for sample k can be set according to the relative positions of these points. We also use the average distance between these same landmarks to rescale the mean shape at each iteration so that scale remains numerically defined.

We assume a common, fixed, 2x2 covariance for each landmark derived from the measurement process. These are composed into the matrix C. This is a tri-diagonal matrix, the diagonal line of which contains data for individual landmarks. Outside of the 2x2 covariances, the off diagonal elements of C are zero, i.e. there are no correlations between landmarks. The use of a fixed data covariance cancels out when taking the weighted mean, to regenerate the conventional formula, i.e.

$$\mathbf{m} = \sum_{k=1}^{K} \mathbf{z}_k C^{-1} \left[\sum_{k=1}^{K} C^{-1} \right]^{-1} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{z}_k \tag{2}$$

where $\mathbf{z}_k = [\mathbf{z}_{k1}, \mathbf{z}_{k2}, ..., \mathbf{z}_{kN}]$. In order to apply the PCA as a standard statistical tool and obtain a model, we introduce ghost points (Figs. 1 and 2). The reason for this is that the points $\mathbf{z}_k$ do not have uniform independent noise distribution around them which is one of the assumptions for the PCA to be applicable.

Although transformation of data can be considered as a new space, it can also be interpreted as an affine reprojection. Ghost points are accordingly defined in the original coordinate system and, being scaled projections relative to the shape centroid, are an alternative way to summarise the original measurement. The process amplifies the spatial variation in directions which are well measured relative to those which are not so that the resulting locations have isotropic errors. In turn, this allows accurately measured structure to be encoded in the most significant eigen vectors of the linear model.

We hence transform $\mathbf{z}_k$ to ghost points $\mathbf{g}_k$ using the matrix W in the same coordinate system.

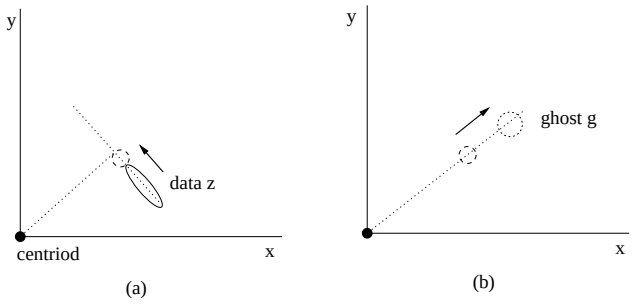

Figure 1: Graphically, we can describe the generation of ghost points in two steps: a) transforming the elliptical distribution (along the axis of elongation using the mean as origin and the elliptical aspect ratio as scale) to produce a circular distribution (left); b) scaling the circular distribution by a global scale (along the axis connecting the centre of circle to the centroid) for all landmarks in the shape (right).

$$\mathbf{g}_k^T = W(\mathbf{z}_k - \mathbf{m})^T \tag{3}$$

The whitening matrix W is computed by applying singular value decomposition to C^{-1}, i.e.

$$C^{-1} = U^T V U \tag{4}$$

By making $W^T I W$ equivalent to C^{-1} we find that

$$W = V^{1/2} U \tag{5}$$

Application of PCA to $\mathbf{g}_k$ follows for construction of the shape covariance

$$F \;=\; \sum_k^K \mathbf{g}_k^T \mathbf{g}_k \;\approx\; \sum_j^J \mu_j \mathbf{e}_j^T \mathbf{e}_j$$

giving the eigen vectors $\mathbf{e}_j$ and values μ_j for the whitened space of ghost points as those which minimise the unexplained variance for fixed $J < N$.

For any specific shape example k, linear factors λ_{jk} can be computed to best approximate $\mathbf{z}_k$ with the model

$$\mathbf{z}'_k = \mathbf{m} + W^{-1} \sum_{j=1}^{J} \lambda_{jk} \mathbf{e}_j \tag{6}$$

where J parameter is the number of eigen vectors used in the model and $\lambda_{jk} \;=\; \mathbf{e}_j.\mathbf{g}_k$.

In order to find the best $R_k, \mathbf{t}_k, s_k$ parameters for each k, the cost function we would expect to be minimised (e.g. using simplex) is a Mahalanobis distance given by

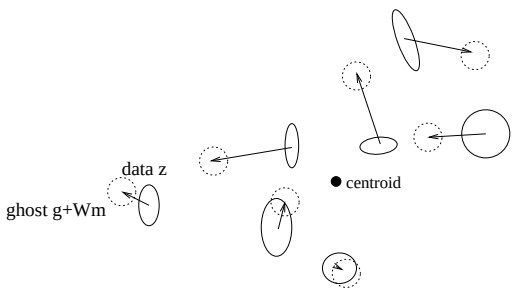

Figure 2: Ghost points for multiple landmarks. Each point is the affine projection of a landmark performed in such a way as to achieve homogenous spatial errors on the resulting location. Ghost points are therefore merely a redefinition of the measured location.

$$\log(P_{kz'}) = (\mathbf{z}'_k - \mathbf{z}_k)^T C^{-1} (\mathbf{z}'_k - \mathbf{z}_k) \tag{7}$$

This is simply the Likelihood estimate for the location of the shape given the linear model and the assumed measurement covariances. However, if we do this while leaving the linear shape parameters free to vary we will trade off changes in orientation and translation against eigen vector components and not achieve a tight grouping of shape (see below). Clearly this degeneracy of representation would not happen if the parameters associated with alignment were incorporated as orthogonal components of the linear model. The standard Procrustes technique avoids this problem by minimising the least squares difference to the mean, in advance of building the PCA model. The model then describes the minimum variation possible following alignment (i.e. orthogonalising the location and linear model parameter groups). A generalisation of this approach which takes account of the measurement of each point is needed if we wish to reduce the sensitivity of the alignment to outliers and support use of pseudo-landmarks. In order to minimise the information which the linear model needs to describe $\mathbf{z}'_k$ is replaced by $\mathbf{m}$. This is equivalent to minimising

$$F(P_{kz'}) \;=\; \log(P_{kz'}) \;+\; \lambda_k^2 \tag{8}$$

By replacing C with I this reduces to the least squares function for standard Procrustes. We can therefore interpret this as a generalisation of the standard approach. However, we cannot interpret $F(P_k)$ as a conventional Likelihood, rather it is a modification to deal with non-independence of the parameters, associated with the information remaining within the aligned sample. The use of C here accords with a direct analogy with Fisher information (the volume of the data space scaled by measurement accuracy).

We can also approach the construction of a Likelihood in an alternative manner. A genuine Likelihood should be based upon the variation of the data around the assumed model. Failure to do this results in only partial registration and residual distributions which cannot be meaningfully interpreted [2]. Using this argument, if we wish to align to the mean shape we should use a covariance B consistent with the distribution around this mean. It therefore quantifies the entire variation of the sample as opposed to the noise in the measurement. The corresponding Likelihood function is then;

$$\log(P_{km}) = (\mathbf{m} - \mathbf{z}_k)^T B^{-1} (\mathbf{m} - \mathbf{z}_k) \tag{9}$$

with

$$B = \frac{1}{K} \sum_{k=1}^{K} (\mathbf{m} - \mathbf{z}_k)(\mathbf{m} - \mathbf{z}_k)^T \tag{10}$$

This also stabilises the alignment process with respect to poorly localised landmarks, but unlike C, B is independent of eigen vector estimation. We can therefore iteratively compute the best alignment and B prior to construction of the linear shape model. We therefore have two alternative methods to consider, and as use of either Eq. (8) or Eq. (9) require an initial estimate of the model and transformed data $\mathbf{z}_k$, it makes sense to start from the Procrustes result.

To reach the best possible alignment we should iteratively estimate R_k, $\mathbf{t}_k$ and s_k using the assumed $\mathbf{e}_j$, $\mathbf{m}$, C^{-1} (or B^{-1}) and W^{-1}. This gives us a new $\mathbf{z}_k$, and so, a new $\mathbf{m}$ and F for construction of $\mathbf{e}_j$.

For fixed covariances, convergence can be monitored via construction of the total Likelihood

$$\log(P) \;=\; \sum_{k}^{K} \log(P_k)$$

One may use the final estimates of $\mathbf{z}_k$ and $\mathbf{z}'_k$ after convergence to find the measurement covariance

$$C = \frac{1}{K} \sum_{k=1}^{K} (\mathbf{z}'_k - \mathbf{z}_k)(\mathbf{z}'_k - \mathbf{z}_k)^T \tag{11}$$

either as a new estimate of the assumed C, or as a consistency check.

When attempting to estimate C, the use of free parameters during model fitting reduces the sample covariance obtained from residuals. A possible outcome of this is the over weighting of landmarks leading to a runaway convergence on one landmark, during iterative estimation. For a single scale parameter associated with an approximate linear vector $\mathbf{f}$ we can estimate the expected reduction in the covariance for each 2x2 landmark component of the matrix as

$$\Delta_n C \;=\; \frac{\mathbf{f_n}^T \otimes \mathbf{f_n}}{\mathbf{f}^T C^{-1} \mathbf{f}}$$

[2] Bookstein: "Wherever there is partial registration the true value of a (vector deformation) is inaccessible."

Here the denominator is the change in χ^2 expected due to a unit change in $\mathbf{f}$, and $\mathbf{f}_n = D_n\mathbf{f}$ is the 2D component of $\mathbf{f}$ corresponding to landmark n. So that

$$\Delta C = \sum_n^N \Delta_n C$$

The parameters of the linear model, including scale, rotation, translation and linear model weightings can be written in this way. Though in this work we currently only correct for the linear model, as this provides an estimate of the variation for each landmark following alignment. For an eigen vector $\mathbf{e}$ defined in the whitened ghost space this would suggest a total correction of

$$\Delta C = \frac{W^{-1}\mathbf{e}^T \otimes W^{-1}\mathbf{e}}{W^{-1}\mathbf{e}^T C^{-1} W^{-1}\mathbf{e}} = W^{-1}\mathbf{e}^T \otimes W^{-1}\mathbf{e}$$

which is correct for the required tri-diagonal components, and must be added to the sample covariance. The known structure of the covariance can be enforced by zeroing relevant off-diagonal terms.

A statistical test is required to give an indication of the appropriateness of the assumed linear model for fixed J. For this work, we have used a simple χ^2 test based upon the construction of these corrected covariances on one data set and then used for the calculation of χ^2 for a second set. If any variance component was found to have been limited (by the methods described above), the local 2x2 diagonal section of C was rescaled to give a $\chi^2/DoF = 1$ (degrees of freedom) before testing on the second data set. Here we had access to repeated mark-ups for this process, though it could just as easily be done by splitting the data in two. For large numbers of samples ($K > 30$) the resulting statistic when applied to each 2D landmark is expected to be approximately Gaussian with mean $2K$ and variance $4K$. So that, for example, for 100 samples we expect 95 % ($\pm$ 3 S.D.) of the data to lie in the range $140 < \chi^2 < 260$. As in all cases with such tests it was necessary to be wary of the possible effects of outliers.

Although this model has been defined for independent landmarks, it should be noted that it would be possible to define C so as to include (off diagonal) measurement correlations along extended strings, so supporting general shape analysis. The theory presented here can thus be considered as an extension to both Procrustes based shape analysis and active shape models.

3 Experiments

We have experimented with a dataset of manual mark-ups for fly wing images (left wings of female flies) containing 200 samples [13]. A sample picture from the fly wing data is shown in Fig. 3. There are 15 landmark points on each fly wing as shown in Fig. 5.

First, we study how Likelihood would work if only the log-Likelihood term was used for optimisation (Eq. 7). In Fig. 4, we show the spread of the aligned shapes (z_k) in the top, while the spread of the models (z'_k) are shown in the bottom of the figure. Here while each shape is well aligned to its corresponding model (based on the error histograms we have monitored), the aligned shapes are not necessarily aligned to a mean shape which is perhaps the common definition of alignment. These results are suggestive of the information available for the definition of a shape co-ordinate system without introducing the additional requirement of minimising the variance described by the model. Hence, in what follows we focus on Procrustes and the two alternative methods explained earlier which are based upon Likelihood.

For the Likelihood approaches, 20 iterations were used involving building a three component model using PCA, optimisation process using simplex, and estimation of the covariance matrices as described earlier. The same number of iterations were used for Procrustes, involving computing the mean shape, and optimisation using simplex. Although the eigen vectors are not used directly in Procrustes, these can easily be computed for behaviour analysis of the method. The results obtained from these experiments are shown in Figs. 5-8 and Tables 1-5.

In the top row of Fig. 5, we show the mean shape to which the shapes are aligned when Procrustes is used. Here, we have labelled the landmarks from 1 to 15. In the middle and bottom rows of the figure, we show the spread of the models, to which the corresponding shapes are aligned when Likelihood is used in conjunction with additional term λ_k^2 (Eq. 8), and when rather than the covariance matrix C the biological covariance matrix B is used in the optimisation (Eq. 9) and the mean shape $\mathbf{m}$ is used in place of the model z'_k. In what follows, we refer to the former method as LC and the latter as LB.

The spread of the aligned points are shown in Fig. 6. Here from the top to the bottom the subfigures correspond to Procrustes, LC and LB methods. Apart from being comparable, these results cannot tell us much about these three methods.

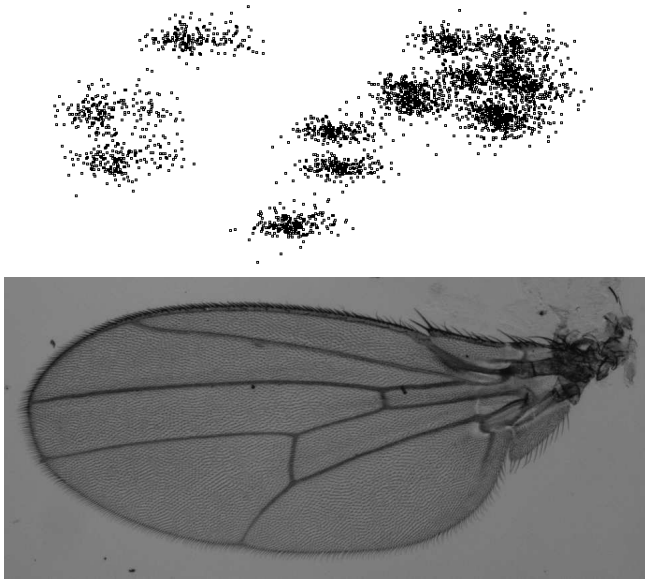

Figure 3: The original data for the 15 landmark points on 200 fly wings superimposed (FemaleLeft:L1) together with a sample fly wing image.

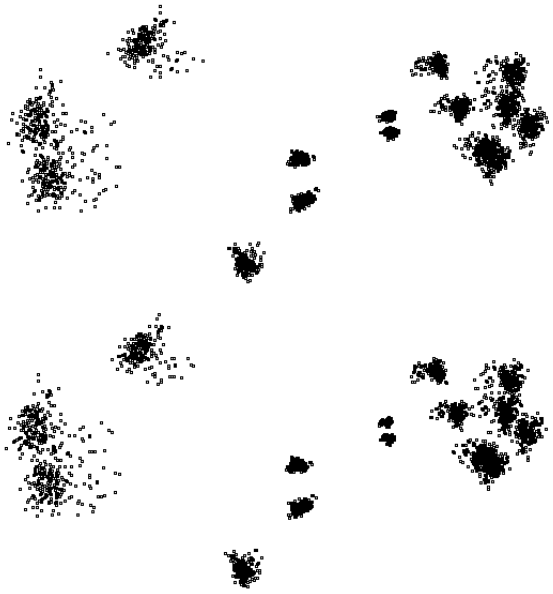

Figure 4: The aligned shapes (top) and the models to which the shapes are aligned (bottom) using Likelihood when optimising the log-Likelihood only (Eq. 7).

As mentioned earlier, the main advantage of Likelihood over Procrustes is that the former provides us with measurement error covariances. To highlight this point, we have applied the two Likelihood methods to the data and listed the estimated variances for each landmark in Table 1. Unconstrained estimates of landmark covariance can lead to an object location being driven by one feature. To avoid this we have enforced the lower limit 1.0 on variance values, which should be considered as the minimum credible localisation accuracy for any landmark. Obviously, to maintain the orientation of each 2x2 covariance matrix, the remaining 3 elements of the matrix need to be rescaled by the scale used to bring either of the two diagonal elements to 1.0. We believe that if the covariance estimate could be corrected to obtain a true unbiased estimate such a process would be unnecessary. Once the alignment is achieved, to obtain the final covariances, a further adjustment may be needed. This is to ensure that for each landmark a χ^2 test based on the corresponding error residuals and the 2x2 covariance matrix results in a unity value for the ratio χ^2/DoF. We choose the minimum number of linear components (eigen vectors) with

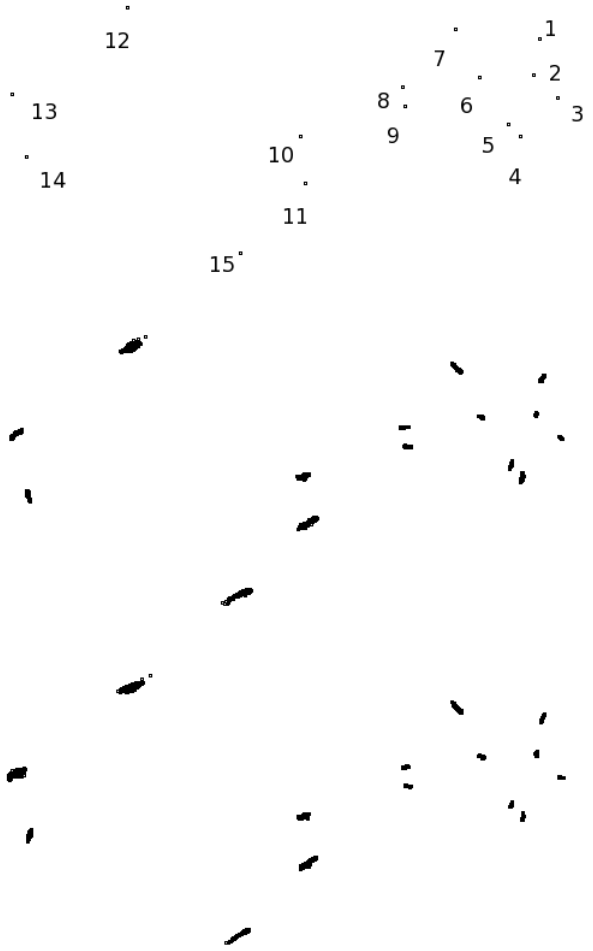

Figure 5: The mean shape to which all shapes are aligned in Procrustes with labels on the landmark points (top); the model to which each shape is aligned using Likelihood LC (middle) and LB (bottom); note that all models are shown superimposed.

which each method results in reasonable values of χ^2/DoF. For instance, this number is set here to 3 for both the LC method and LB alignment method. In a later experiment, we show that how this test can be used to perform similarity tests on the left-right wings. By looking at the variance values in Table 1 for LC and LB methods, it is clear that most points result in similar covariances, except for points 3, 13 and 14. Here the LB approach seems to have done a better job of accommodating the variation seen for point 14, while LC has done a better job for points 13 and 3.

To make it easier to compare the variances with the local structures in the fly wings, we have 2D oriented error bars on each landmark point computed from the corrected covariance matrix ($C + \Delta C$). These are shown in Fig. 7. It is also worth comparing these error bars with the spread of the alignments in Fig. 6 and corresponding models in Fig. 5. This unmodellable component of variation (error) does not simply reflect the initial distribution for the landmark (see for example landmark 14). There is however, good agreement with the example taxonomy discussed below (Fig. 9).

Here, we turn our attention to comparing the eigen vectors corresponding to the methods. We do not attempt to compare individual eigen vectors visually, as we believe that they work together in a manner that is not always straightforward to analyse. For instance, a process that is strong in the first eigen vectors of one method may be absent for those of another method, while it may appear more strongly in the second vectors, etc. However, as a measure of similarity, we compute the dot product of the vectors from each of the two Likelihood methods (LC and LB) against the ones from Procrustes. These results, which are listed in Table 2, provide quantitative measures to study the different behaviours of the two Likelihood methods in terms of eigen vector components. The differences are sufficient to say that the linear models constructed in each case are not well approximated by the Procrustes

Figure 6: The aligned shapes using Procrustes, Likelihood LC and LB methods (from-top-to-bottom).

landmark	C_{xx}(LC)	C_{yy}(LC)	C_{xx}(LB)	C_{yy}(LB)
1	1.050828	0.654479	0.932378	1.084497
2	1.154094	0.464422	1.070157	0.876925
3	2.735413	0.758391	2.954354	0.168542
4	0.587761	1.086068	0.700778	1.179136
5	0.604578	0.815126	0.425962	0.979304
6	5.537220	0.491463	5.309956	0.682133
7	3.654634	0.914281	3.665892	1.237434
8	10.541393	0.828501	9.851564	0.849361
9	11.046012	0.747111	10.767957	0.653714
10	11.880961	0.655518	11.239818	0.741850
11	5.075245	2.642622	5.181160	2.915924
12	3.581522	0.884512	4.593532	1.866610
13	0.569974	5.887439	3.189554	4.943126
14	1.354149	6.871092	0.941556	7.286369
15	11.009977	1.013384	14.271631	1.826566

Table 1: The diagonal elements of the covariance matrices (variances) for alignment of data using LC (Eq. 8) and LB (Eq. 9).

method.

To compare the methods quantitatively, we have computed the value of each cost function using the results given

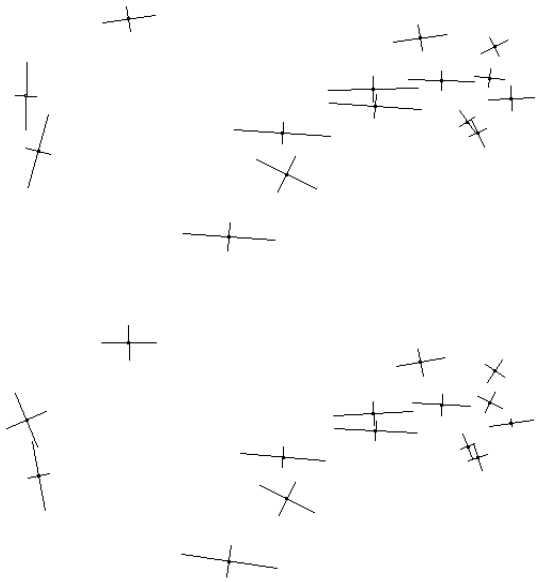

Figure 7: The unexplained statistical variation extracted following the Likelihood LC (top) and LB (bottom) methods; error bars were scaled about 10 times larger for visualisation purposes; see Table 1 for original values of variances.

method	first	second	third vectors
LC	0.494903	0.477170	0.373610
LB	0.518845	0.378013	0.246271

Table 2: Modulus of the dot product between the eigen vectors from Procrustes and the Likelihood methods LC and LB.

by these methods (Tables 3 and 4). This is a reasonable way of comparing the alignments with and without taking the covariance estimates into account, as the main difference between the least squares and Mahalanobis-based distances. From Table 3 and 4, it is clear that the two cost functions are smallest for the method using the corresponding cost function definition during alignment.

cost function	Procrustes	Likelihood LC
least squares	67294.480	82709.606
Equation (8)	65258.738	25087.276

Table 3: Cost function corresponding to each method is used to find the values corresponding to both methods; the covariance estimates from Likelihood LC are used when computing the Mahalanobis-based distance (Eq. 8) for the results given by Procrustes.

cost function	Procrustes	Likelihood LB
least squares	67294.480	77233.695
Equation (9)	11616.728	6000.000

Table 4: Cost functions corresponding to each method is used to find the values corresponding to both methods; the covariance estimates from Likelihood LB are used when computing the Mahalanobis distance (Eq. 9) for the results given by Procrustes.

Finally, we experiment with data from left and right wings (L and R) of up to 200 female flies. The data from left wings (L1) are part of the data used so far in our experiments. Moreover, two images of each wing has been taken from slightly different viewing positions (1 and 2). These images have been used for marking-up in order to also

test reproducibility. As a result, there are four datasets available: L1, L2, R1 and R2.

As mentioned earlier, to perform a symmetry test or a reproducibility test we use the quantity χ^2/DoF. In the first phase of the experiment, we align the L1 landmark data using LC and LB methods and estimate covariances (as described above), while in the second phase we use these covariance estimates as input to the alignment process of L2, R1 and R2 in turn. These input covariances will be fixed during the whole alignment process. We perform this experiment for three different numbers of samples from each dataset: $K = 200$, $K = 100$ and $K = 50$. For the case of $K = 200$, in Table 5, we list the χ^2/DoF ratios as a measure of similarity between LC-L1 on one side and LC-L2, LC-R1 and LC-R2 on the other side. The same comparison applies to LB-L1 on one side and LB-L2, LB-R1 and LB-R2 on the other side.

landmark	LC-L2	LC-R1	LC-R2	LB-L2	LB-R1	LB-R2
1	0.9174	0.9349	0.9303	0.7846	1.0016	1.0374
2	0.9163	0.8512	0.7328	0.9006	0.9261	0.8062
3	0.9682	1.0217	0.9499	0.9883	0.7869	0.7267
4	0.9203	0.9429	0.8922	0.9983	0.9925	0.9197
5	1.0292	0.8020	0.8905	1.2864	0.8637	0.9063
6	0.9724	0.8399	0.7933	0.9809	0.6596	0.5711
7	1.1393	1.0633	1.0357	0.9081	1.1864	1.1481
8	1.0358	0.9175	0.8633	0.9849	0.7751	0.7512
9	1.0411	0.8418	0.9119	0.9621	0.6780	0.8228
10	1.0550	1.0560	0.9408	1.0596	1.0043	0.9670
11	1.1637	1.0726	0.9490	1.0313	1.3109	1.3126
12	0.7741	1.1039	1.2720	1.3889	1.7464	1.7426
13	0.8271	0.7337	0.6881	0.9323	1.0858	0.9597
14	1.1878	0.9391	0.9336	1.0083	2.0093	2.1920
15	1.0415	1.0337	1.0939	1.0422	1.4421	1.3723

Table 5: The ratio χ^2/DoF as a similarity measure between: a) LC-L1 on one side and LC-L2, LC-R1 and LC-R2 on the other side; b) LB-L1 on one side and LB-L2, LB-R1 and LB-R2 on the other side; expected range for 200 samples in each dataset is [0.72, 1.28].

The χ^2 test results listed in Table 5 for $K = 200$ and also those corresponding to $K = 100$ and $K = 50$ are summarised in Fig. 8. Here, plots from the top to the bottom correspond to the experiments using 200, 100 and 50 samples in each dataset. We set the statistical test for significant difference on the basis of an allowable range of χ^2/DoF corresponding to ± 4 $S.D.$, i.e. [0.72, 1.28] for 200 samples, [0.6, 1.4] for 100 samples, and [0.43, 1.57] for 50 samples. The corresponding allowable range is shown using two straight lines in each plot. Also, ratio results for all 15 landmarks obtained by applying each of the two methods (LC and LB) to each of the three datasets (L2, R1 and R2) are marked using six different markers. The results are more consistent for the LC method. We believe that this is due to the slight differences in initial data alignment resulting from estimation of the B matrix for the second test data set.

4 Discussion

The aim of this work was to investigate ways of incorporating poorly localised landmarks (such as pseudo-landmarks) into Procrustes. The resulting approach deals with linear modelling of shape by treating the data generation process as originating from both linear correlating factors and random fluctuations. This latter process being inherent to the way that corresponding points have been defined and measured (for *any* definition of feature correspondence). The overall analysis estimates both the linear correlations and the distribution of the statistical fluctuations. By performing the analysis in this way we ensure that the assumptions regarding data behaviour are applied in a self-consistent manner during estimation of the linear model.

Although the definition of a ghost point might be seen as a 'non-shape' property, which could vary with experiment, as the measurement accuracy is a property of the structure of the region around a landmark, this is just a way of encoding the relative local shape. For simple line based structures we can construct an example taxonomy which illustrates how specific local structures lead to corresponding measurement covariances (Fig. 9). Although more sophisticated landmarks (such as defined by image patches) would be more difficult to make predictions for, a given region of information with fixed structure will non-the-less give rise to a fixed structural covariance. Pseudo landmarks, have poor localisation in one direction, invalidating their use in conventional Procrustes/PCA analysis.

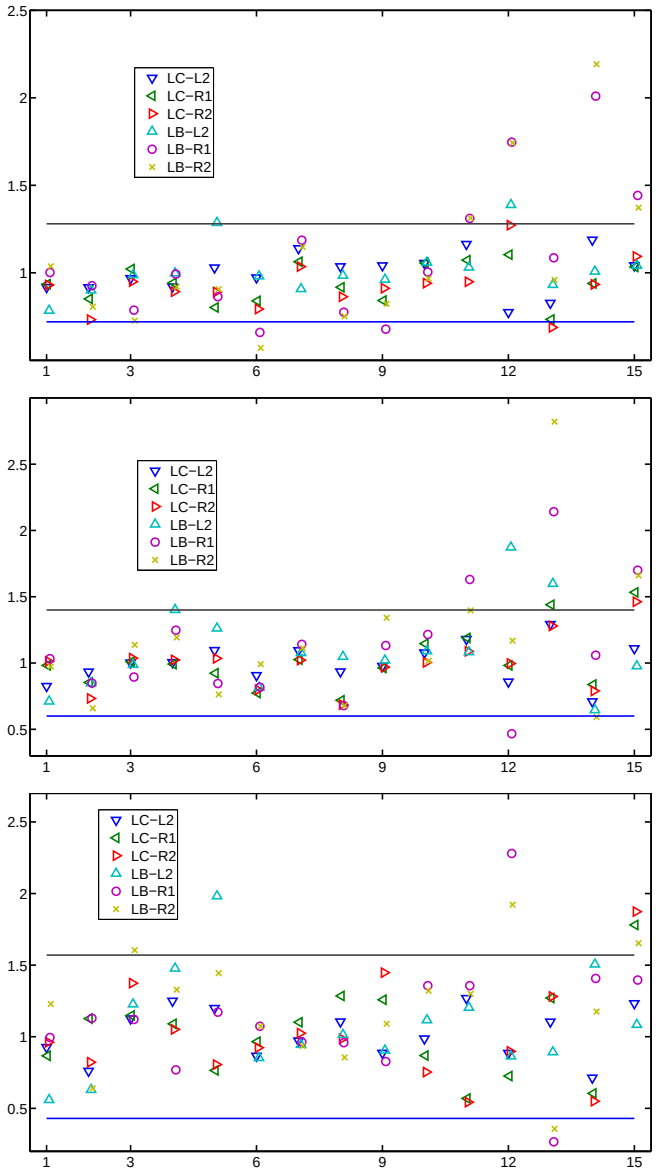

Figure 8: Plots for the ratios χ^2/DoF (as given in Table 5) when experimenting with 200 (top), 100 (middle) and 50 (bottom) samples.

It is exactly this kind of behaviour which we wish to address via the use of ghost points. The expected advantages of this approach are several;

- It will lead to a consistent way of incorporating all forms of landmark into our analysis.
- It removes the instabilities inherent in the analysis due to poorly determined points.
- It affords the application of an eigen vector analysis statistical rigour.
- It offers the possibility of interpreting the linear modelling process as a statistical approximation, with consequent interpretations of the requirement for the number of linear model components.
- Finally, generalisation of the approach would seem to be possible which would support the analysis of curves.

We will now consider the objections to using a weighting process mentioned in the introduction. Now that we have a specific definition for how to weight landmark data, we can see that using ghost points does not invalidate use of Kendall's statistics as suggested in [9]. The use of these approaches follows due to scale normalisation of the shape data, it is not an intrinsic property of the use of the original landmarks co-ordinates per-se. Ghost points are just a new way to define landmark location relative to local shape structure. As a consequence, normalisation

of scale in the ghost space $\mathbf{g} + W\mathbf{m}$ will result in an equivalent behaviour suitable for Kendall's approach. We can reproject scaled shapes onto the tangent space defined in the transformed ghost space if we wish, in order to remove local curvature arising from scale normalisation [3]. The remaining objection to weighting, objective definition, is discussed below.

Procrustes chooses a space consistent with Euclidean distance, but this has problems of poorly defined landmarks. Our suggestion here is to use the Mahalanobis distance defined by measurement. This method can be equivalently interpreted as a redefinition of the landmark location as ghost points.

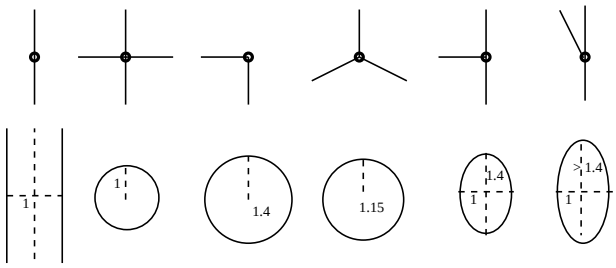

Figure 9: A simple taxonomy of structures showing the expected link between linear features and localisation covariance. Covariances are defined assuming that the local (Fisher) information for spatial localisation is proportional to boundary length.

Far from there being no objective way to define these covariances [9], [11], there are at least three; we can estimate them directly from repeatability of measurements; they can be directly estimated via conventional statistical means when using Likelihood based landmark location (Cramer-Rao Bound); or they can be estimated as the unexplained stochastic variation (residuals) in fitted data (see Fig. 7). For the latter, when estimated using residuals of the fitted shape model, we will see contributions additional to the measurement process, this is the stochastic (therefore unmodellable) behaviour of the biology itself. The CRB estimated by the second approach, (suitable for automated measurement systems) is defined only by measurement, and is therefore also a bound on the minimum observable variation possible for each landmark. Agreement between Fig. 7 and the predictions from Fig. 9 for landmarks 6-15 are an indication that the observed variances are being driven by the spatial information encoded in local structure. However, because of structural complexity, we can make no simple predictions for points 1 to 5.

The question arises of how we should now interpret ghost points and the linear model which describes them. The conventional interpretation of Procrustes is that the resulting linear model is a pure shape description which can be directly associated with biological processes. Some may argue that extending the approach to weight data, even to accommodate pseudo-landmarks, breaks with this tradition.

However, it is our belief that any distinction between the original landmark and a ghost point, as locations which are somehow true measurements of biology in one case but not the other is arbitrary (Appendix). Both need to be calibrated using known samples with identifiable biological cause in order to make any interpretation. Also, reweighting of data using a covariance is statistically equivalent to modifying the information available by changing the specified set of landmarks. If we object to using measurement covariances then we should also object to the arbitrary initial choice of landmarks. When viewed like this there should be no objection to defining measured locations so that their statistical properties at least accord with the way we intend to analyse the data (i.e. PCA). We need to be careful in both cases that we do not over-interpret the results of any analysis and any intrinsic relevance to biology.

Use of a least squares measure (which assumes isotropic errors) does not introduce some pure measurement of biology. We should not interpret a fixed difference in the millimetre position of a landmark as a quantitative biological measure, invariant to its location in the organism. For example, well measured structures at small scales can be correlated with variations in near-by structures at larger scales. A Procrustes based approach is certain to mask small scale effects with any large scale variation. Whereas covariance weighted analysis would identify these correlations if the measurement is appropriate.

[3] We note this while at the same time pointing out that few researchers seem to use Kendall's methods in practice. We should not over value the ability to approximate a distribution using a particular computational form. It would be unlikely for biological variation to "co-operate" with what are solely algebraic and computational restrictions. Monte-Carlo resampling methods have no down-side in the light of modern computing capabilities.

5 Conclusions

In this work we have associated the problems of working with pseudo-landmarks in biological shape analysis as being a consequence of the statistical assumptions implicit to analysis techniques such as Procrustes/PCA. We have implemented a new method which takes appropriate account of measurement and landmark localisation stability in order to obtain a new form of analysis which is consistent with a Likelihood based definition of the alignment and model building tasks. By introduction of the concept of ghost points, we have tried to explain how this new analysis differs from conventional methods only by the choice of landmark co-ordinate system (Fig 2). This, in effect, was always an arbitrary (implicit) choice, similar in statistical behaviour to the initial choice of landmarks.

Our interpretation of the resulting linear model should simply be this; the locations we intend to define as landmark positions are those points in the image plane which can be projected from observed structure with homogenous measurement noise (ghost points). The model constructed is then a summary of correlations which can be inferred from the set of measurements we have chosen to take. As such the results should be no less meaningful as a starting point for biological interpretation than those of Procrustes. Though, by adhering more closely to a Likelihood interpretation, they should be a better summary of the measured data. There will however, be consequent limits on the quantitative value of results. Researchers may need to agree not only on the set of landmarks (and pseudo-landmarks) they intend to use, but also on the associated covariances.

Our results indicate that measurement covariances can be reliably estimated in our data for sample sizes at least as small as 50, contradicting the common view that this should be impossible. This may be the result of using a corrected estimate rather than the conventional sample covariance. Although this is an encouraging start, more work is needed to assess the stability of the new analysis technique, and to determine the best way to use independently estimated parameter covariances associated with measurements. The covariances estimated from sample data (as here) have the possibility of being reduced by accommodating the observed variation within the linear model. While this can be mitigated by setting a lower limit on the diagonal elements of C, it would probably be more appropriate to fix the covariance to the measured values. In particular, as measurement noise is spherical in the ghost space, the PCA model should be estimated for

$$F \;=\; \sum_{k}^{K} \mathbf{g}_k^T \mathbf{g}_k \;-\; I$$

so that pure noise does not appear in the eigen vector approximation. Unambiguous interpretation of data (i.e. as genuine biology as distinct to measurement error) may be impossible without this step.

We might also consider first a partial improvement over Procrustes, whereby we set covariances in one of two fixed forms homogenous 2D (landmarks), or heterogeneous 1D (pseudo-landmarks). This might deal directly with the main problem of constructing a model of smooth shape while introducing less variation between analyses.

Appendix : The Limits of Statistical Shape Analysis

From a more philosophical standpoint we can consider what we are doing when we identify landmark locations and attempt to compare them between sets (shapes). We do not expect that biology manipulates the locations of our chosen landmarks directly, they simply appear to move around as the net effect of distributed developmental influences. Recent considerations of biology have introduced the phrase "palimpsest" [7], as an analogy with repeatedly erasing and rewriting text in ancient parchment, to describe the way that structures develop. Notice that the initial choice of landmarks is subjective, not only in terms of the features selected but also how we chose to define their locations. To start with there is no clearly identifiable biological process which we can associate to unique locations in shape geometry. This latter point becomes more obvious once we attempt to automatically estimate their position (subjective definitions defy logical analysis). Any quantitative estimate of a landmark position, such as using template based correlation, must always define a structure and associated location for it. However, once we have a structure the location itself depends upon thechoice of origin for template co-ordinates, do we use the centre of a patch or the bottom left corner? If we decide that the intersection of linear structures is a good way to define location, then we could just as easily have chosen another point consistently defined relative to this location. The information pertaining to shape is not provided by the absolute positions of features themselves but by the correlations seen between them. These correlations remain fixed for any fixed shift of the individual landmarks. Our concept of a landmark being at a specific location is purely based upon a pre-conception of co-ordinates, both for origin and specification of axes. This is driven by anatomical naming conventions, which encourage researcher to define approximate regions as localisable objects at ever finer scales in order to support correspondance matching. Issues arising due to semi-landmarks are then the inconvenient aspects of this problem

which cannot be re-defined away. In summary, a landmark is the result of a localisation procedure (partly influenced by multiple biological considerations) which has an associated positional uncertainty. Acceptance of this situation allows us to extend the process of co-ordinate selection to the idea of ghost points.

We will now try to understand the shape analysis task starting from a well defined concept of statistical similarity. In a fixed co-ordinate frame, we can define the difference between two measured sets of (approximately) corresponding data points z_1 and z_2 with associated fixed covariances as defined in the main text[4] C_1 and C_2 as

$$log(P) \propto -(z_1 - z_2)^T (C_1 + C_2)^{-1} (z_1 - z_2)$$

Where P here is the probability that z_1 could be perturbed to z_2 on the basis of the allowed random (noise) fluctuations. The size of these fluctuations, determine our ability to identify separable locations in this data space. The proportionality is used to remind us that we are working with probability densities on the RHS. In this way we can define a biological variation s.t. $z_1 \rightarrow z_2$. Unfortunately, in biological applications we have no intrinsic co-ordinate frame in which to appropriately define such a metric. In the absence of a fixed co-ordinate frame we can only compute the minimum supportable statistical difference, once the over-all effect of a transformation (T) has been removed $z_1' = T(z_1, \theta_{z_1,z_2})$.

$$log(P) \propto -(z_1' - z_2)^T (C_1' + C_2)^{-1} (z_1' - z_2) \qquad (1)$$

We can at this point approximate the net measurement covariance and it is reasonable to assume that it will be fixed ($C_1' + C_2 = C$) for localised regions of the data space, i.e. similar shapes. As the true transformation parameters are unknown, optimising (1) in order to determine θ results in a LOWER limit on the true (generating) difference. The process of transforming the data erases any aspect of biological variablity which is on average similar to an allowable transformation, for example due to scale or rotation. Importantly, this means that if the data were actually generated by a model with known parameters (or specific $z_1 \rightarrow z_2$), our analysis of the data will not regenerate them. This mechanism is in operation at the core of Kendall's definition of shape. As a consequence, it can also be argued that the linear parameters describing biological variation should strictly be made orthogonal to those used for alignment so that Kendall's definition is enforced.

In essence any specific method of aligning data in order to define shape is necessarily an arbitrary choice, selected to reduce the shape variation prior to model building. Non-the-less, such behaviour is suitable for inference. This includes construction of hypothesis tests which aim only to test equivalence, or pattern recognition systems which associate locations in the space with other relevant information (see below).

The interpretation of the similarity function as a limit can also be concluded by considering the subjective nature of the selection of landmarks for a model. Shapes can only be distinguished according to the measured information available. This information is determined by the defined set of landmarks, and the associated measurement covariance. If you believe it is possible to separate two groups of data in shape space, simply delete the landmark which contributes most strongly to the discriminating direction and look again. Two nearby landmarks can (in the limit) be treated as a single landmark with suitably modified covariance. Similarly, a landmark well located in 2D can be interpreted as two superimposed (orthogonal) semi-landmarks. In this respect the idea of covariance and selection of the number of landmarks are statistically equivalent, in that they both modify the positions, shape and separability of embedded distributions, and can be considered interchangeable.

If we now decide to construct a standard co-ordinate frame (perhaps defined by a mean shape m) in which to represent each shape with fixed location $\theta_{z,m}$, we will only be able to construct meaningful differences between shapes if the transformation applied between shapes is consistent. i.e.

$$\theta_{z_1,z_2} = \theta_{z_1,m} - \theta_{z_2,m}$$

As θ was determined above in order to obtain the best $log(P)$, any variation from this equality will imply that statistical shape differences in the new co-ordinate frame can only be taken as an UPPER limit on computed similarity, implying a contradiction with the limit discussed above. Consequently, as the behaviour needed for construction of hypothesis tests is lost, this interpretation needs to be avoided.

This is an alternative way to think about the issue of consistency to that presented in [4]. Conventional statistical definitions always consider (average) expectations which presuppose that the average value of a parameter is the best way to summarise multiple estimates. Here, we are interested in knowing that the estimated statistical difference between the pair of shapes will be the same in the shape space as it was for the original data. The simple process of computing orientation based upon any pair of lines will be completely consistent, as will a parameterisation which

[4]If the observed covariances are a non-linear function of the measurement space this would generate a statistical process with heteroscedastic properties, necessitating more complex similarity functions with associated metrics and non-linear minimum cost paths.

makes the transformation completely orthognal to the linear model (as mentioned above). Orthogonalisation is required to prevent linear shape descriptors playing the same role as alignment parameters, so generating degenerate alignments. Procrustes shape alignment achieves only approximate consistency, so that if we compute the rotation between two samples the angular difference will not be the difference of the aligned orientations to the mean. Without some form of orthogonalisation, biological variation which could have been interpretted as a transformation will only be partly erased from the linear model (contradicting Kendall's definition of shape). This will also be the case for Likelihood based methods, with assumed residuals matching the observed data distribution, such as equations (7) and (8). But there is no need to expect that use of anisotropic measurement covariances will result in a less consistent co-ordinate frame than Procrustes. Indeed, measurement covariance based weighting of data is accepted as best practice in science, in that it achieves the minimum variance on the transformation parameter estimates. If we want better consistency we should consider tranformation and model orthogonalisation and align each shape using the full model description, rather than using the partial alignment of Procrustes (see earlier Bookstein footnote).

Following the argument above, distances measured in the shape space resulting from Procrustes alignment can at best be only interpreted as approximating a lower limit on measurable biological variation. Any principle components of variation used to summarise the distribution of a set of data should be correspondingly interpreted, not as the "maximum correlations in biological shape variation", but as the "maximum OBSERVABLE correlations in the set of chosen landmarks subject to measurement and following alignment". Although it is tempting to adopt Kendall's definition of shape and accept the first description, our models are then easy to over interpret. As any eigen-vector can be augmented with unobservable alignment transformations, neither the orientation, lengths or relations between them, can be interpreted absolutely in terms of biology. In any case true (causal) biological correlations would not be expected to be either orthogonal or linear.

What we can say is the set of principle axes, taken together with an unspecified transformation, provide a compact density model of the data. Any inference as to the relative separation of multiple groups (e.g. A is further from B than C) must be made with extreme care. The subjective definition of the set of landmarks has the consequence that any conclusion is only valid if it is the same for any linear re-weighting of the data (including for example affine re-projection). This represents a severe restriction on the interpretation of shape models. We can use the resulting shape spaces only to construct densities, and it is the properties of these densities (such as overlap) which we might quantify, not their position or distribution. This favours any technique which is invariant to linear transformation of co-ordinates, such a Fishers Linear Discriminants and Canonical Analysis. However, attempts to interpret the separation of distributions directly presupposes an equivalence of the effects of alignment and uniqueness of the chosen shape space. Unfortunately, such suppositions are untestable by the very nature of the shape analysis problem.

Use of a least squares measure (which assumes isotropic errors) does not change these facts and introduce some pure measurement of biology. Though it eliminates alternative representations of data it constitutes only an arbitrary choice which (as has been noted) has no relevance to underlying biological variation. We should not interpret a fixed difference in the millimetre position of a landmark as a quantitative biological measure, invariant to its location in the organism. For example, well measured structures at small scales can be correlated with variations in near-by structures at larger scales if the measurement covariance is appropriate for this. Whereas, a Procrustes based approach is certain to mask small scale effects with any large scale variation. Any who believe that the physical size of a structural variation can be used as a surrogate for a measure of biological (eg. genetic) change should consider relevant examples, such as; variations in skull shape (1cm) and details of the inner ear (0.01cm). It is also well known that the interesting biological correlations may show up in lower order eigen-vectors, as well as in the higher ones. We must expect processes to correlate due to genetic variation all the way from the molecular to the macro-structural (even behavioural) level, in the process of generating viable biological systems.

Our argument regarding the interchangeability of covariance and landmark selection tells us that the shape space generated by Procrustes is not special. It further constitutes an additional approximation regarding our initial definition of measured statistical difference. Rather than being a true measure of biological shape, it can only be regarded as suboptimal use of the available data for both alignment and model construction. Such observations already seems to have been made before in the literature (see Bookstein's comments in [10]).

We can only safely justify aligned shape models for discrimination. And if we want better discrimination, in order to improve the lower bound on the separability of any two shapes, we must either measure more points, or measure those we have already more accurately. Or if we we using Procrustes, we might just choose to make better use of the data we already have by using measurement covariances. Whether using measurement covariances or not, our ability to make use of the density model is restricted by the need to define the meanings of specific locations in the shape space. With regard to the separation of multiple groups described above, we can only state that A is further from B than C if we have already calibrated the corresponding locations in shape space with regard to

something biologically meaningful. Evidence that such a calibration is possible and leads to simple interpretations in some datasets should not lead us to conclude that this should be expected in general. This would constitute an over interpretation.

In summary, a linear model constructed for describing shape variation in a normalised space is a low dimensional summary of the measurements supporting a lower limit approximation of statistical difference. It only has biological meaning when locations in the resulting space are calibrated with other biological knowledge. This is no less true for "Procrustes/PCA" analysis as for the more sophisticated approach suggested in this document. We should therefore be able to use this new approach to resolve problems with analysis of psuedo-landmarks, without destroying any innate biological validity of a shape space derived using Procrustes.

References

[1] Bookstein, F.L., Tensor Biometrics for Changes in Cranial Shape, Ann. Human Biol. 11, 413-437, 1984.

[2] Goodall C.R., The Statistical Analysis of Growth in Two Dimensions, Doctoral dissertation, Department of Statistics, Harvard University, 1983.

[3] Kendall, D.G., Shape-Manifolds, Procrustean Metrics, and Complex Projective Spaces, Bull. London Math. Sot. 16, 81-121, 1984.

[4] Kent, J.T., Mardia K.V., Consistency of Procrustes Estimators. J.R. Statist, Soc (B), 59, 1, 281-190, 1997.

[5] Bookstein, F.L., Size and Shape Spaces for Landmark Data in Two Dimensions, Statis. Sci. 1, 181-242, 1986.

[6] Gunz, P., Mitteroecker, P. and Bookstein, F.L., Semilandmarks in Three Dimensions, in Modern Morphometrics in Physical Anthropology, Edited by D.E. Slice, New York: Kluwer Academic/Plenum Publishers, 73-98, 2005.

[7] Halgrimsson B., Lieberman D.E, Young N.M, Parsons T., Wat S. Evolution of Covariance in Mammalian Skull, Novartis Found Symp, 284, 164-185, 2007.

[8] Bookstein, F.L., Landmark Methods for Forms Without Landmarks: Morphometrics of Group Differences in Outline Shape, Med. Image. Anal. 1(3): 225-243, 1997.

[9] Zelditch, M.L., Swiderski, D.L., Sheets H.D. and Fink W.L., Geometric Morphometrics for Biologists A Primer, Elsevier Academic Press, New York, 2004.

[10] Goodall, C., Procrustes Methds in the Statistical Analysis of Shape, J. Roy. Stat. Soc. Series B (Methodological), 53(2): 285-339, 1991.

[11] Lele, S. and Richtsmeier, J.T., Statistical Models in Morphometric: Are They Realistic? Syst. Zool., 39(1): 60-69, 1990.

[12] Theobald, D.L. and Wuttke, D.S., Empirical Bayes Hierarchical Models for Regularizing Maximum Likelihood Estimation in the Matrix Gaussian Procrustes Problem, Proc. Natl. Acad. Sci. USA, 103(49): 18521-18527, 2006.

[13] Klingenberg Lab, *http://www.flywings.org.uk.*

[14] The Max Planck Institute for Evolutionary Biology, Plon, Germany, *http://www.mpg.de/english/portal/index.html.*