

# Investigating patterns in shape analysis

John T Kent<sup>1</sup>

Department of Statistics, University of Leeds, e-mail: j.t.kent@leeds.ac.uk

**Abstract:** Shape analysis is concerned with features of geometric objects that are invariant under changes in location, rotation and scale of the object. The methodology to identify patterns includes methods from multivariate analysis such as PCA and cluster analysis. But there are also specific types of structure that are unique to shape analysis, such as symmetry and singularity points.

**Keywords:** Clustering; Independent component analysis; Principal component analysis; Procrustes analysis; Singularity; Symmetry

## 1. Overview of shape space

A “configuration” is a  $k \times m$  matrix containing the coordinates for  $k$  landmarks in  $m$ -dimensional space. Configurations in the plane ( $m = 2$ ) are the most straightforward to visualize in practice, though higher values of  $m$  can be easily treated mathematically. A “shape” (or, more specifically “similarity shape”) is an equivalence class of configurations under the set of similarity transformations. In  $m = 2$  dimensions,  $2k - 4$  degrees of freedom are needed to specify a shape ( $2k$  df for the configuration, minus 2 df for translation in the horizontal and vertical directions, minus 1 df for scale and minus 1 df for rotation). In order to treat shape statistically, it is necessary to represent shape in some coordinate system. Possibilities include Bookstein coordinates (for each configuration choose a similarity transformation to move two specified landmarks to fixed positions, say (0,0) and (1,0); then the coordinates of the remaining landmarks can be treated as shape coordinates) and Procrustes tangent coordinates (essentially least squares residuals after matching a set of data configurations as closely as possible to an “average” configuration). For this paper we shall focus on the use of Procrustes tangent coordinates.

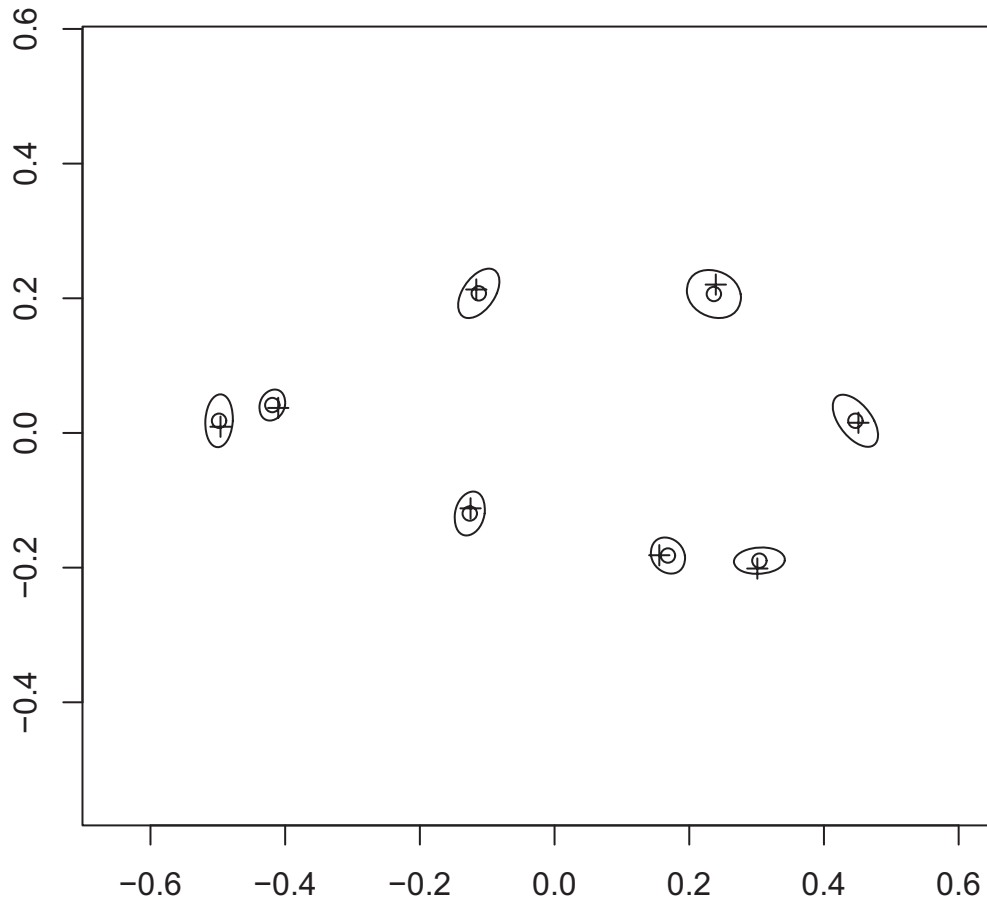
## 2. Multivariate methods

For concentrated shapes, there are methods of analysis analogous to multivariate analysis. Given a dataset of  $k \times m$  configurations  $X_i$ ,  $i = 1, \dots, n$ , the first step is to find a “mean shape”, represented by a centered and scaled configuration  $\hat{\mu}$ , say, by generalized Procrustes analysis; then compute the Procrustes residual matrices  $T_i$  with respect to  $\hat{\mu}$ . Details are given in Kent and Mardia (2001). These residuals satisfy  $km - m - 1 - m(m - 1)/2$  structural linear constraints (depending on  $\hat{\mu}$ ).

To carry out conventional multivariate analysis, it is convenient to stack the columns of each tangent matrix  $T_i$  into a  $(km \times 1)$  vector. Methods such as PCA and ICA can be applied to these tangent vectors (making due allowance for the structural singularities).

---

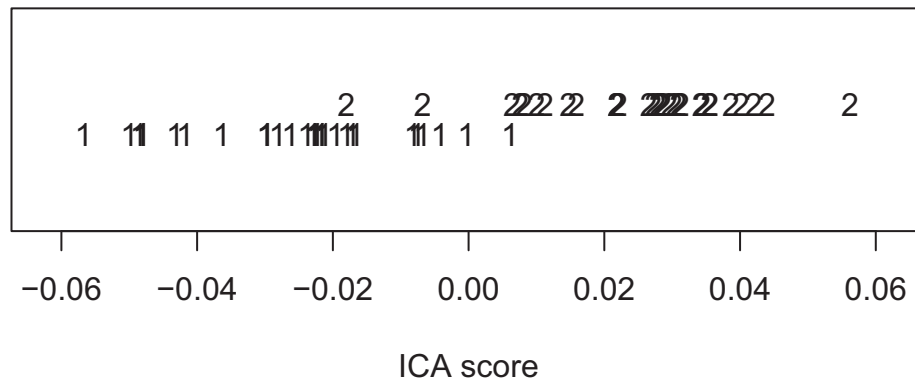
<sup>1</sup>Address of correspondence: Dept. of Statistics, University of Leeds, Leeds, LS2 9JT, UK



**Figure 1:** *Procrustes tangent coordinates for male and female gorillas. Mean male landmarks (o) and mean female landmarks (+) are plotted with a 95% probability ellipse at each landmark, centered at the mean male landmarks.*

One of the classic problems in multivariate analysis is deciding whether and how to cluster a set of data into two or more groups. Tyler *et al.* (2009) have recently presented a unified a set of methods for clustering, which involve constructing two estimates of a scatter matrix with different amounts of robustness, and then looking at a relative eigenanalysis of one matrix with respect to the other to ‘identify possible clustering directions. A variant of their methods is given by looking for the direction with minimal (more specifically, large in absolute value and negative) kurtosis; see e.g. Bugrien and Kent (2005). Finding the projection direction to minimize the kurtosis can be a delicate numerical problem, as the objective function can have multiple local minima. One possible approach

is to start the optimization algorithm from the direction of the first principal component, though the resulting algorithm is no longer invariant under affine transformations of the data.



**Figure 2:** *ICA scores for male (1) and female (2) gorillas, separated vertically for clarity.*

*Example.* Dryden and Mardia (1998) give a set of configuration data for gorilla skulls. There are  $k = 12$  landmarks in  $m = 2$  dimensions on  $n_1 = 29$  males and  $n_2 = 30$  females. For the purpose of this example, we ask if it is possible to separate the sexes without making explicit use of the sex information. A plot of the data in Procrustes tangent coordinates is given in Figure 1. Although the two sexes do not seem very strongly separated, a Hotelling's  $T^2$  is highly significant. A plot of the kurtosis-based projection (with the search starting from the first principal component direction) is given in Figure 2. Note that the separation is good, but not perfect. In contrast Fisher's linear discriminant function gives perfect separation of the sexes in this example, though of course it uses knowledge of the sex information in the construction of the projection direction.

### 3. Symmetry

Consider landmarks on the surface of e.g. the human face, either in a 2d frontal projection or in 3d. There is a line (in 2d) or plane (3d) of approximate left-right symmetry running through the midline (midplane) of the face. Some landmarks will lie on the midline (midplane), called “solos”, and others will arise as “pairs”, one on the left side and one on the right side.

There are at least two statistical questions that can be asked in this setting.

- (a) Given a single set of data, is there any evidence of asymmetry – i.e. is the sample mean shape compatible with symmetry. For humans, there is a suggestion of “left-side” dominance.

- (b) Given two sets of data, each set can be “symmetrized” by including a reflected configuration along with the original configuration in an augmented data set. The mean configuration for the augmented dataset is automatically exactly symmetric. We can then ask how the symmetric components of shape differ between the two groups.

For more information, see e.g. Kent and Mardia (2001); Kolamunnage and Kent (2003, 2005).

## 4. Singularities in shape space

Huckemann *et al.* (2009) drew attention to a problem which arises in  $m \geq 3$  dimensions when looking at the shape of a configuration whose rank is at most  $m - 2$  (e.g. a set of collinear landmarks in  $\mathbb{R}^3$ ). Such shapes are “singular” points of shape space, and Huckemann *et al.* (2009) proposed a “geodesic-based” method of principal component analysis instead of the more common “Euclidean-based” PCA in order to look for parsimonious reduced-dimension descriptions of the data.

A good analogy to this situation is given by data on the unit sphere in  $\mathbb{R}^3$  which are uniformly spread around the equator. Since the equator is an example of a geodesic on the sphere, the data are one-dimensional from a geodesic perspective. On the other hand two Euclidean coordinates are needed to represent the equator, so the data are two-dimensional from a Euclidean perspective. Similar issues arise near singularities in shape space, and we shall illustrate the problem with a set of tree crown data. Each configuration consists of several landmarks in the trunk near the base of the tree and one landmark in the crown of the tree.

## References

- Bugrien J.B. and Kent J.T. (2005) Independent component analysis: an approach to clustering, in: *Proceedings in Quantitative Biology, Shape Analysis and Wavelets*, Barber S., Baxter P.D., Mardia K.V. and Walls R.E., eds., Leeds University Press, 111–114.
- Dryden I.L. and Mardia K.V. (1998) *Statistical Shape Analysis*, Wiley, Chichester.
- Huckemann S., Hotz T. and Munk A. (2009) Intrinsic shape analysis: geodesic PCA for Riemannian manifolds modulo isometric lie group actions, *Statistica Sinica*, to appear.
- Kent J.T. and Mardia K.V. (2001) Shape, Procrustes tangent projections and bilateral symmetry, *Biometrika*, 88, 469–485.
- Kolamunnage R. and Kent J.T. (2003) Principal component analysis for shape variation about an underlying symmetric shape, in: *Proceedings in Stochastic Geometry, Biological Structure and Images*, Aykroyd R.G., Mardia K.V. and Langdon M.J., eds., Leeds University Press, 137–139.
- Kolamunnage R.R. and Kent J.T. (2005) Decomposing departures from bilateral symmetry, in: *Proceedings in Quantitative Biology, Shape Analysis and Wavelets*, Barber S., Baxter P.D., Mardia K.V. and Walls R.E., eds., Leeds University Press, 75–78.
- Tyler D.E., Critchley F., Dümbgen L. and Oja H. (2009) Invariant co-ordinate selection, *Journal of the Royal Statistical Society, Series B*, to appear.