Multivariate Analysis of Data in Curved Shape Spaces

Christopher G. Small

University of Waterloo

June 9, 2011

Outline

1. Landmark analysis
2. Size and shape coordinates
   - The configuration of landmarks
   - Goodall-Mardia coordinates
   - Kendall shape coordinates
3. Statistical analysis of size and shape
   - A quick two-sample test for shape differences
4. References

Outline

Selecting landmarks from an image or object:

Selecting landmarks from an image or object:

- To reduce the high dimensionality of an image, it is often useful to summarize important features using landmarks.

Selecting landmarks from an image or object:

- To reduce the high dimensionality of an image, it is often useful to summarize important features using landmarks.
- Landmark analysis of images is a type of vectorization.

Selecting landmarks from an image or object:

- To reduce the high dimensionality of an image, it is often useful to summarize important features using landmarks.
- Landmark analysis of images is a type of vectorization.
- Salient features of d-dimensional images are encoded as vectors in $\mathbb{R}^{n \times d}$ where n is the number of landmarks.

Selecting landmarks from an image or object:

- To reduce the high dimensionality of an image, it is often useful to summarize important features using landmarks.
- Landmark analysis of images is a type of vectorization.
- Salient features of d-dimensional images are encoded as vectors in $\mathbb{R}^{n \times d}$ where n is the number of landmarks.
- Landmarks can be chosen by an expert or by a simple heuristic procedure.

By an expert ....

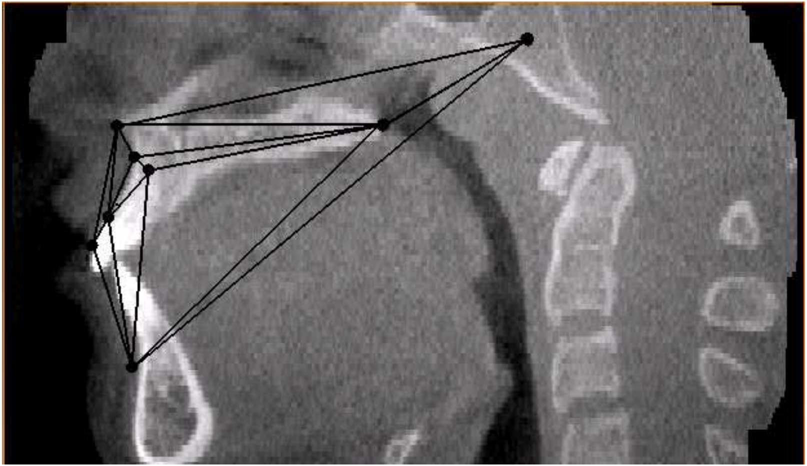


Figure: Seven landmarks from an orthodontics study

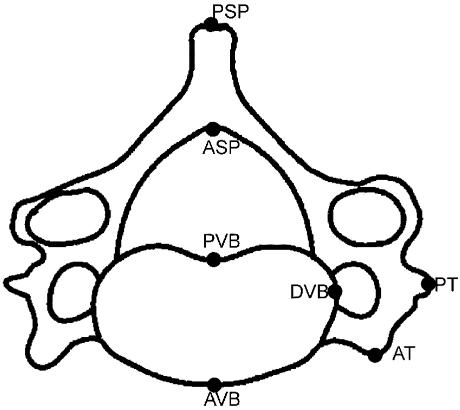


Figure: Landmarks for cervical gorilla vertebra (GGG & GGB)

By a simple heuristic procedure ....

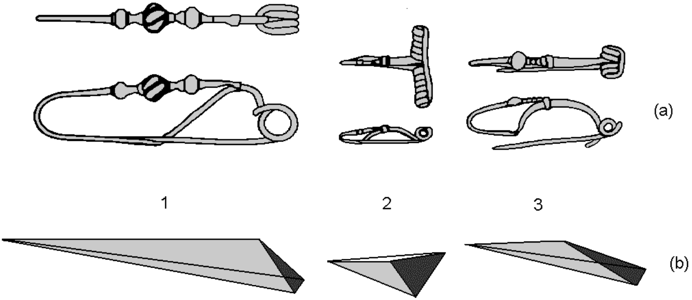


Figure: Iron Age brooch shapes encoded with 4 landmarks

The aim of such a landmark analysis is

The aim of such a landmark analysis is

- to cluster images into groups,

The aim of such a landmark analysis is

- to cluster images into groups,
- to perform a discriminant analysis when images are classified by some variable (training),

The aim of such a landmark analysis is

- to cluster images into groups,
- to perform a discriminant analysis when images are classified by some variable (training),
- to test an hypothesis concerning two or more groups of images.

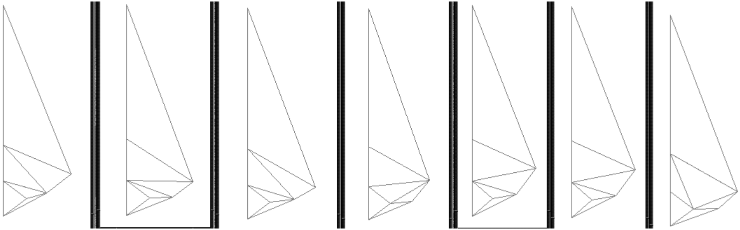


Figure: GGB cervical vertebrae

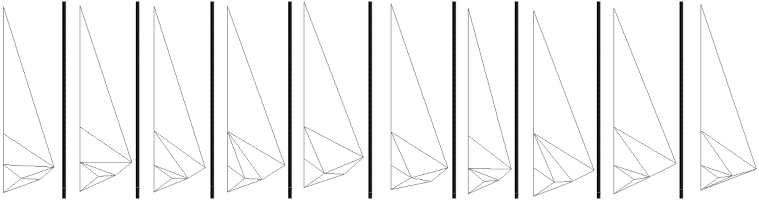


Figure: GGG cervical vertebrae

Outline

Outline

The configuration matrix

The configuration matrix

- Let

  $$X_i = (X_{i1}, \dots X_{id}) \quad i = 1, \dots, n\,.$$

  denote the n landmarks in $\mathbb{R}^d$.

The configuration matrix

- Let
$$X_i = (X_{i1}, \ldots X_{id}) \quad i = 1, \ldots, n\,.$$
denote the n landmarks in $\mathbb{R}^d$.

- We can arrange these vectors into a matrix of coordinates for all landmarks as follows.

The configuration matrix

- Let

$$X_i = (X_{i1}, \ldots X_{id}) \quad i = 1, \ldots, n \, .$$

denote the n landmarks in $\mathbb{R}^d$.

- We can arrange these vectors into a matrix of coordinates for all landmarks as follows.

Definition

The configuration matrix is defined as

$$X = (X_{ij}), \quad i = 1, \ldots, n; \, j = 1, \ldots, d \, .$$

Invariance to choice of coordinate system

Invariance to choice of coordinate system

- In many configuration matrices we are not concerned with information about the location and orientation of the configuration.

Invariance to choice of coordinate system

- In many configuration matrices we are not concerned with information about the location and orientation of the configuration.
- So the statistical methods should be invariant under any common transformation of the landmarks by an isometry of $\mathbb{R}^d$.

Invariance to choice of coordinate system

- In many configuration matrices we are not concerned with information about the location and orientation of the configuration.
- So the statistical methods should be invariant under any common transformation of the landmarks by an isometry of $\mathbb{R}^d$.

Definition

Two $n \times d$ configuration matrices

$$(X_{ij}) = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \quad \text{and } (Y_{ij}) = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

are said to have the same size and shape if there is some isometry $\tau : \mathbb{R}^d \to \mathbb{R}^d$ such that $Y_i = \tau(X_i)$ for all $i = 1, \ldots, n$.

Size and shape class of X

Size and shape class of X

- When $n \times d$ matrices X and Y have the same size and shape we write $X \sim Y$.

Size and shape class of X

- When $n \times d$ matrices X and Y have the same size and shape we write $X \sim Y$.
- Then $\sim$ defines an equivalence relation on the set of all configuration matrices.

Size and shape class of X

- When $n \times d$ matrices X and Y have the same size and shape we write $X \sim Y$.
- Then $\sim$ defines an equivalence relation on the set of all configuration matrices.
- We write the equivalence class of all configuration matrices Y with the same size and shape as X as $s(X)$.

Size and shape class of X

- When $n \times d$ matrices X and Y have the same size and shape we write $X \sim Y$.
- Then $\sim$ defines an equivalence relation on the set of all configuration matrices.
- We write the equivalence class of all configuration matrices Y with the same size and shape as X as $s(X)$.
- Equivalently,

$$s(X) = \{Y \ : \ X \sim Y\}.$$

Outline

Goodall and Mardia (1992), Goodall and Mardia (1993) proposed a representation of the size and shape of landmark cofigurations using lower triangular matrices and the QR-factorization.........

Removing location

Removing location

- Assume that the landmarks are in general position. That is, no three landmarks lie on a straight line, no four in some common plane, ..., subset of m lies in some common $(m-2)$-flat.

Removing location

- Assume that the landmarks are in general position. That is, no three landmarks lie on a straight line, no four in some common plane, ..., subset of m lies in some common $(m-2)$-flat.

- To remove location we choose the coordinate system so that one of the landmarks (say X_1) is at the origin. Equivalently we can subtract the first row of X from all other $n-1$ rows.

Removing location

- Assume that the landmarks are in general position. That is, no three landmarks lie on a straight line, no four in some common plane, ..., subset of m lies in some common $(m-2)$-flat.
- To remove location we choose the coordinate system so that one of the landmarks (say X_1) is at the origin. Equivalently we can subtract the first row of X from all other $n-1$ rows.
- We then delete the first row, yielding the $(n-1) \times d$ matrix

$$\widetilde{X} = (X_{ij} - X_{1j}); \quad i = 2, \cdots n; \quad j = 1, \ldots, d\,.$$

called the pre-size-and-shape matrix.

Removing orientation:

Removing orientation:

- Next, we must remove orientation information from $\widetilde{X}$.

Removing orientation:

- Next, we must remove orientation information from $\widetilde{X}$.
- We perform a QR-factorization on $\widetilde{X}$ so that

$$\widetilde{X} = s(X)\,\Gamma(X)\,.$$

where $\Gamma(X)$ is a $d \times d$ special orthogonal matrix and $s(X)$ is an $(n-1) \times d$ lower triangular matrix.

Removing orientation:

- Next, we must remove orientation information from $\widetilde{X}$.
- We perform a QR-factorization on $\widetilde{X}$ so that

$$\widetilde{X} = s(X)\,\Gamma(X)\,.$$

where $\Gamma(X)$ is a $d \times d$ special orthogonal matrix and $s(X)$ is an $(n-1) \times d$ lower triangular matrix.

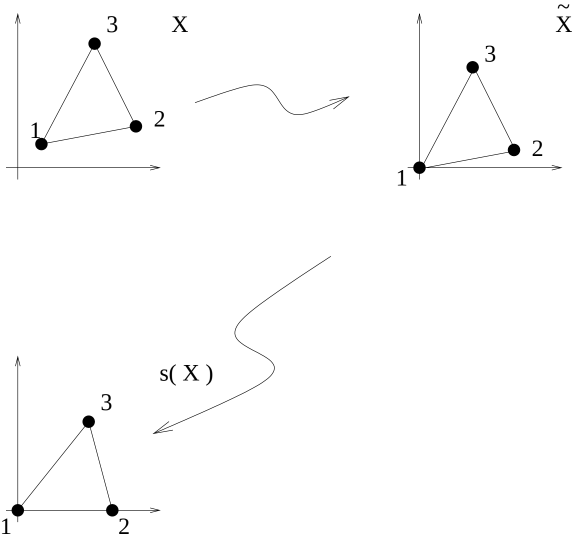
- The matrix $s(X)$ is a coordinate representation of the size and shape of X.

Goodall-Mardia coordinates for size and shape

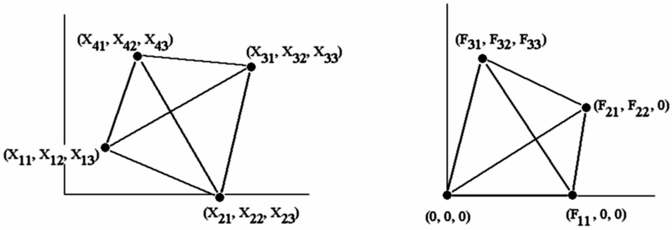
$$s(X) = \begin{pmatrix} F_{11} & 0 & 0 & \cdots & 0 \\ F_{21} & F_{22} & 0 & \cdots & 0 \\ F_{31} & F_{32} & F_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ F_{d1} & F_{d2} & F_{d3} & \cdots & F_{dd} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ F_{(n-1)1} & F_{(n-1)2} & F_{(n-1)3} & \cdots & F_{(n-1)d} \end{pmatrix}$$

where $F_{11}, \ldots, F_{(d-1)(d-1)} > 0$.

For three landmarks in $\mathbb{R}^2$



For four landmarks in $\mathbb{R}^3$



Goodall-Mardia coordinates for shape

To eliminate size, so that only shape information remains, we scale the elements of the size and shape matrix so that $\sigma_{ij} = F_{ij}/F_{11}$. This gives us the shape matrix

$$\sigma(X) = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \sigma_{21} & \sigma_{22} & 0 & \cdots & 0 \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \sigma_{d3} & \cdots & \sigma_{dd} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sigma_{(n-1)1} & \sigma_{(n-1)2} & \sigma_{(n-1)3} & \cdots & \sigma_{(n-1)d} \end{pmatrix}$$

where again $\sigma_{22}, \ldots, \sigma_{(d-1)(d-1)} > 0$.

Advantages of Goodall-Mardia coordinates:

Advantages of Goodall-Mardia coordinates:

- Goodall-Mardia coordinates provide a simple recipe for representing landmark shapes as multivariate data.

Advantages of Goodall-Mardia coordinates:

- Goodall-Mardia coordinates provide a simple recipe for representing landmark shapes as multivariate data.
- While the coordinates are represented in lower triangular matrix form, they may be encoded as vectors in standard statistical packages such as R, etc.

Advantages of Goodall-Mardia coordinates:

- Goodall-Mardia coordinates provide a simple recipe for representing landmark shapes as multivariate data.
- While the coordinates are represented in lower triangular matrix form, they may be encoded as vectors in standard statistical packages such as R, etc.
- The statistician who wishes to analyse shapes can calculate these coordinates and apply standard multivariate procedures. (While standard distribution assumptions will not hold for the coordinates, they are never realised in practice anyway.)

Problems with Goodall-Mardia coordinates:

Problems with Goodall-Mardia coordinates:

- The QR-factorization is not symmetric in the landmark labels. That is, if the landmarks are labelled in different order the shape matrix undergoes a transformation.

Problems with Goodall-Mardia coordinates:

- The QR-factorization is not symmetric in the landmark labels. That is, if the landmarks are labelled in different order the shape matrix undergoes a transformation.
- So statistical inference will depend upon arbitrary labelling of vertices.

Problems with Goodall-Mardia coordinates:

- The QR-factorization is not symmetric in the landmark labels. That is, if the landmarks are labelled in different order the shape matrix undergoes a transformation.
- So statistical inference will depend upon arbitrary labelling of vertices.
- For example, suppose $X^{(1)}, \ldots, X^{(k)}$ are k configurations each of n landmarks in $\mathbb{R}^d$, and let $\pi\left(X^{(1)}\right), \ldots, \pi\left(X^{(k)}\right)$ be row-permuted (i.e., relabelled) versions of the original configurations.

Problems with Goodall-Mardia coordinates:

- The QR-factorization is not symmetric in the landmark labels. That is, if the landmarks are labelled in different order the shape matrix undergoes a transformation.
- So statistical inference will depend upon arbitrary labelling of vertices.
- For example, suppose $X^{(1)}, \ldots, X^{(k)}$ are k configurations each of n landmarks in $\mathbb{R}^d$, and let $\pi\left(X^{(1)}\right), \ldots, \pi\left(X^{(k)}\right)$ be row-permuted (i.e., relabelled) versions of the original configurations.
- Then there is no simple affine relationship between the two sets of shapes

$$\sigma\left(X^{(1)}\right), \ldots, \sigma\left(X^{(k)}\right) \qquad \text{and} \qquad \sigma\left(\pi(X^{(1)})\right), \ldots, \sigma\left(\pi(X^{(k)})\right).$$

Problems with Goodall-Mardia coordinates:

- The QR-factorization is not symmetric in the landmark labels. That is, if the landmarks are labelled in different order the shape matrix undergoes a transformation.
- So statistical inference will depend upon arbitrary labelling of vertices.
- For example, suppose $X^{(1)}, \ldots, X^{(k)}$ are k configurations each of n landmarks in $\mathbb{R}^d$, and let $\pi\left(X^{(1)}\right), \ldots, \pi\left(X^{(k)}\right)$ be row-permuted (i.e., relabelled) versions of the original configurations.
- Then there is no simple affine relationship between the two sets of shapes

$$\sigma\left(X^{(1)}\right), \ldots, \sigma\left(X^{(k)}\right) \qquad \text{and} \qquad \sigma\left(\pi(X^{(1)})\right), \ldots, \sigma\left(\pi(X^{(k)})\right).$$

- A statistician using the former may reach different conclusions from a statistician using the latter.

Outline

In 1986, David G. Kendall proposed a coordinatisation of size and shape based upon procrustes fitting. This is the approach to shape analysis taken in Small (1996, 2011), Dryden and Mardia (1998), Kendall et al. (1999) .....

- By a scale variable of order p we mean

- By a scale variable of order p we mean
  - a nonnegative function of size-and-shape: $\lambda(X) = \lambda\,[\,s(X)\,]$,

- By a scale variable of order p we mean
  - a nonnegative function of size-and-shape: $\lambda(X) = \lambda\,[\,s(X)\,]$,
  - which is homogeneous of order p: $\lambda\,[\,s(c\,X)\,] = c^p\,\lambda\,[\,s(X)\,]$ for all $c \in \mathbb{R}^+$.

- By a scale variable of order p we mean
  - a nonnegative function of size-and-shape: $\lambda(X) = \lambda\,[\,s(X)\,]$,
  - which is homogeneous of order p: $\lambda\,[\,s(c\,X)\,] = c^p\,\lambda\,[\,s(X)\,]$ for all $c \in \mathbb{R}^+$.
- So when $\lambda > 0$,

$$s(X) = \{\,\sigma(X),\,\lambda(X)\,\}\ ,$$

- By a scale variable of order p we mean
  - a nonnegative function of size-and-shape: $\lambda(X) = \lambda[\,s(X)\,]$,
  - which is homogeneous of order p: $\lambda[\,s(c\,X)\,] = c^p\,\lambda[\,s(X)\,]$ for all $c \in \mathbb{R}^+$.
- So when $\lambda > 0$,

$$s(X) = \{\,\sigma(X),\, \lambda(X)\,\}\ ,$$

- and when $\lambda = 0$

$$s(X) = \star$$

- By a scale variable of order p we mean
  - a nonnegative function of size-and-shape: $\lambda(X) = \lambda\,[\,s(X)\,]$,
  - which is homogeneous of order p: $\lambda\,[\,s(c\,X)\,] = c^p\,\lambda\,[\,s(X)\,]$ for all $c \in \mathbb{R}^+$.
- So when $\lambda > 0$,

$$s(X) = \{\,\sigma(X),\,\lambda(X)\,\}\ ,$$

- and when $\lambda = 0$

$$s(X) = \star$$

the shape is degenerate.

- By a scale variable of order p we mean
  - a nonnegative function of size-and-shape: $\lambda(X) = \lambda\,[\,s(X)\,]$,
  - which is homogeneous of order p: $\lambda\,[\,s(c\,X)\,] = c^p\,\lambda\,[\,s(X)\,]$ for all $c \in \mathbb{R}^+$.
- So when $\lambda > 0$,

$$s(X) = \{\,\sigma(X),\,\lambda(X)\,\}\ ,$$

- and when $\lambda = 0$

$$s(X) = \star$$

  the shape is degenerate.
- We write size and shape space as

$$S\Sigma_d^n = \left(\Sigma_d^n \times \mathbb{R}^+\right) \cup \{\star\}.$$

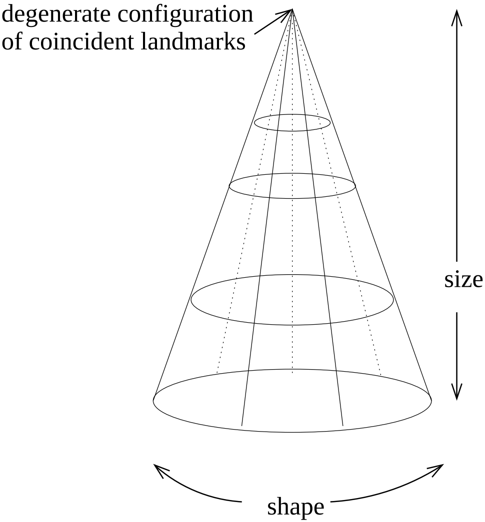


Figure: Schematic diagram of size-and-shape space

Kendall geometry of size-and-shape space

Kendall geometry of size-and-shape space

- To make the Kendall geometry explicit we need to put metrics on shape space Σ_d^n and size-and-shape space $S\Sigma_d^n$.

Kendall geometry of size-and-shape space

- To make the Kendall geometry explicit we need to put metrics on shape space Σ_d^n and size-and-shape space $S\Sigma_d^n$.
- We use a matching metric on $S\Sigma_d^n$ by setting

$$\rho\left(s(X),\, s(Y)\right) = \min\left\{||X - Z|| \;:\; s(Z) = s(Y)\right\}.$$

Warning: this is not geodesic distance!

Kendall geometry of size-and-shape space

- To make the Kendall geometry explicit we need to put metrics on shape space Σ_d^n and size-and-shape space $S\Sigma_d^n$.
- We use a matching metric on $S\Sigma_d^n$ by setting

$$\rho\,(s(X),\, s(Y)) = \min\left\{||X - Z|| \;:\; s(Z) = s(Y)\right\}.$$

Warning: this is not geodesic distance!

- As a scale variable, we choose

$$\lambda(X) = \min_{\tau}\left\{||\tau(X)||^2 \;:\; \tau \text{ isometry}\right\}$$

using the Frobenius norm $||\ ||$. Note: this is simply the trace of the landmark covariance matrix.

Kendall geometry of size-and-shape space

- To make the Kendall geometry explicit we need to put metrics on shape space Σ_d^n and size-and-shape space $S\Sigma_d^n$.
- We use a matching metric on $S\Sigma_d^n$ by setting

$$\rho\,(s(X),\, s(Y)) = \min\left\{||X - Z|| \;:\; s(Z) = s(Y)\right\}.$$

Warning: this is not geodesic distance!

- As a scale variable, we choose

$$\lambda(X) = \min_{\tau}\left\{||\tau(X)||^2 \;:\; \tau \text{ isometry}\right\}$$

using the Frobenius norm $||\ ||$. Note: this is simply the trace of the landmark covariance matrix.

-

$$\Sigma_d^n = S\Sigma_d^n \;\cap\; \{\lambda = 1\}\,.$$

Kendall geometry of size-and-shape space

- To make the Kendall geometry explicit we need to put metrics on shape space Σ_d^n and size-and-shape space $S\Sigma_d^n$.
- We use a matching metric on $S\Sigma_d^n$ by setting

$$\rho\left(s(X), s(Y)\right) = \min\left\{||X - Z|| \ : \ s(Z) = s(Y)\right\}.$$

Warning: this is not geodesic distance!

- As a scale variable, we choose

$$\lambda(X) = \min_{\tau}\left\{||\tau(X)||^2 \ : \ \tau \text{ isometry}\right\}$$

using the Frobenius norm $||\ ||$. Note: this is simply the trace of the landmark covariance matrix.

- $$\Sigma_d^n = S\Sigma_d^n \ \cap \ \{\lambda = 1\}.$$

- The shape space Σ_d^n inherits its metric as a subspace of $S\Sigma_d^n$.

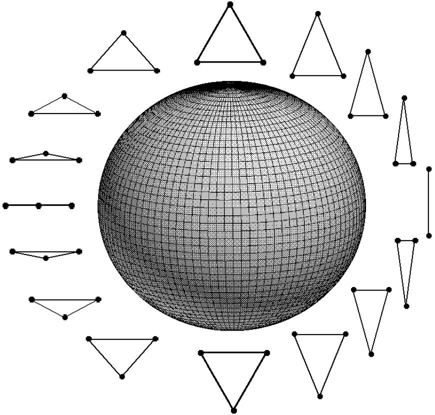
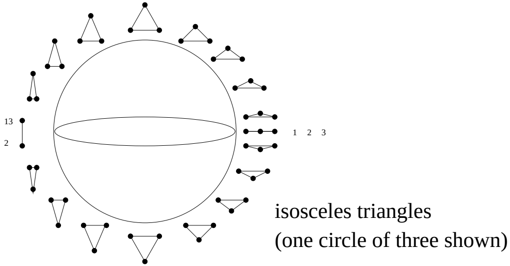


Figure: Shape space Σ_2^3 of triangle shapes in 2D



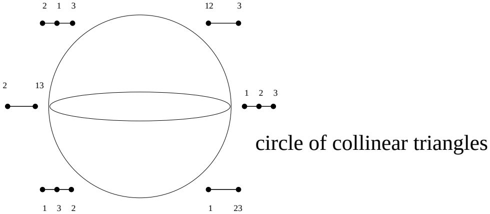


Figure: Shape space Σ_2^3 again. A line of "longitude" corresponding to isosceles triangle shapes (top). The "equator" of collinear triangle shapes (bottom).

Outline

- How do we conduct a statistical analysis of size and shape?

- How do we conduct a statistical analysis of size and shape?
- We will consider two ways ....

- In Goodall-Mardia coordinates, we encode size and shape as vectors (vectorized matrices) and apply standard multivariate methods.

- In Goodall-Mardia coordinates, we encode size and shape as vectors (vectorized matrices) and apply standard multivariate methods.
  - However, as discussed earlier, this distorts the nonlinear geometry of shape.

- In Goodall-Mardia coordinates, we encode size and shape as vectors (vectorized matrices) and apply standard multivariate methods.
  - However, as discussed earlier, this distorts the nonlinear geometry of shape.

- In Kendall geometry, we look for analogs of various multivariate statistical tools appropriate for the geometry of the manifold, E.g.,

$$\begin{aligned} \text{Euclidean distance} &\rightarrow \text{Matching distance/Geodesic distance} \\ \text{Centroid} &\rightarrow \text{Frechet mean} \\ \text{MANOVA} &\rightarrow \text{Analysis of geodesic dispersion} \end{aligned}$$

- In Goodall-Mardia coordinates, we encode size and shape as vectors (vectorized matrices) and apply standard multivariate methods.
  - However, as discussed earlier, this distorts the nonlinear geometry of shape.

- In Kendall geometry, we look for analogs of various multivariate statistical tools appropriate for the geometry of the manifold, E.g.,

$$\begin{aligned} \text{Euclidean distance} &\rightarrow \text{Matching distance/Geodesic distance} \\ \text{Centroid} &\rightarrow \text{Frechet mean} \\ \text{MANOVA} &\rightarrow \text{Analysis of geodesic dispersion} \end{aligned}$$

  - The advantages of this are that the conclusions of such and analysis are not influenced by artificial coordinate systems designed to "make" the data multivariate.

- In Goodall-Mardia coordinates, we encode size and shape as vectors (vectorized matrices) and apply standard multivariate methods.
  - However, as discussed earlier, this distorts the nonlinear geometry of shape.

- In Kendall geometry, we look for analogs of various multivariate statistical tools appropriate for the geometry of the manifold, E.g.,

$$\begin{aligned} \text{Euclidean distance} &\rightarrow \text{Matching distance/Geodesic distance} \\ \text{Centroid} &\rightarrow \text{Frechet mean} \\ \text{MANOVA} &\rightarrow \text{Analysis of geodesic dispersion} \end{aligned}$$

  - The advantages of this are that the conclusions of such and analysis are not influenced by artificial coordinate systems designed to "make" the data multivariate.
  - The disadvantage of this is that an analog of a multivariate method may not be obviously available, or there may be many different analogs of one multivariate method.

Outline

1. Landmark analysis

2. Size and shape coordinates
   - The configuration of landmarks
   - Goodall-Mardia coordinates
   - Kendall shape coordinates

3. Statistical analysis of size and shape
   - A quick two-sample test for shape differences

4. References

Gorilla gorilla gorilla

Gorilla gorilla beringei

Testing for shape differences in cervical vertebrae of two gorilla subspecies

- We consider the shapes of the fifth cervical vertebrae of two subspecies of gorilla: G. g. gorilla and G. g. beringei.

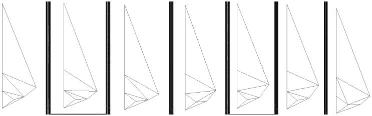
Figure: GGB cervical vertebrae (left), GGG cervical vertebrae (right)

- We consider the shapes of the fifth cervical vertebrae of two subspecies of gorilla: G. g. gorilla and G. g. beringei.

Figure: GGB cervical vertebrae (left), GGG cervical vertebrae (right)

- The most prominent shape differences between the two subspecies are to be seen in the vertex angle at the "top" of the configuration of landmarks: this angle is smaller for GGG than for GGB.

- We consider the shapes of the fifth cervical vertebrae of two subspecies of gorilla: G. g. gorilla and G. g. beringei.



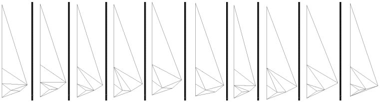


Figure: GGB cervical vertebrae (left), GGG cervical vertebrae (right)

- The most prominent shape differences between the two subspecies are to be seen in the vertex angle at the "top" of the configuration of landmarks: this angle is smaller for GGG than for GGB.
- But is this apparent difference significant? Secondly, which landmarks contribute most to the observed shape differences? Here, we shall only address the first question. See Albert, Le & Small (2003) for more on the second question.

- We test

$$H_0 \; : \; \mathcal{L}\left(\sigma_{GGG}\right) = \mathcal{L}\left(\sigma_{GGB}\right)$$
$$\text{versus}$$
$$H_1 \; : \; \mathcal{L}\left(\sigma_{GGG}\right) \neq \mathcal{L}\left(\sigma_{GGB}\right)$$

where $\sigma \in \Sigma_2^7$ is a random shape from the respective population.

- We test

$$H_0 \; : \; \mathcal{L}(\sigma_{GGG}) = \mathcal{L}(\sigma_{GGB})$$
$$\text{versus}$$
$$H_1 \; : \; \mathcal{L}(\sigma_{GGG}) \neq \mathcal{L}(\sigma_{GGB})$$

  where $\sigma \in \Sigma_2^7$ is a random shape from the respective population.

- We shall assume that any differences in shape between GGG and GGB are due to a shift in Frechet mean, and that the geodesic dispersions of the two populations are roughly equal.

- Samples:

- Samples:
  - GGG: $\sigma_{1,1},\ \sigma_{1,2},\ \ldots,\ \sigma_{1,10}$
  - GGB: $\sigma_{2,1},\ \sigma_{2,2},\ \ldots,\ \sigma_{2,7}$

- Samples:
  - GGG: $\sigma_{1,1}, \sigma_{1,2}, \ldots, \sigma_{1,10}$
  - GGB: $\sigma_{2,1}, \sigma_{2,2}, \ldots, \sigma_{2,7}$

- Proposed test statistic

$$T = \frac{\sum_{j=1}^{10} \sum_{k=1}^{7} \rho^2(\sigma_{1j}\,, \sigma_{2k})}{\sum_{j=1}^{9} \sum_{k=j+1}^{10} \rho^2(\sigma_{1j}\,, \sigma_{1k}) + \sum_{j=1}^{6} \sum_{k=j+1}^{7} \rho^2(\sigma_{2j}\,, \sigma_{2k})}$$

where $\rho(\sigma\,, \tau)$ is the geodesic distance in Σ_2^7 which is the Fubini-Study metric on the complex projective space $\Sigma_2^7 \cong \mathbb{C}P^5(4)$.

- To approximate the distribution of the statistic T under H_0 we can use a permutation test.

- To approximate the distribution of the statistic T under H_0 we can use a permutation test.
- Shuffle the 10+7=17 shapes into random order:

$$\Pi : \{1, 2, \ldots, 17\} \to \{1, 2, \ldots, 17\},$$

where Π is a random permutation of 1, 2, ... , 17. then partition the shuffled 17 shapes into two new groups:

$$\sigma^*_{1,1}, \ldots, \sigma^*_{1,10}\,; \quad \sigma^*_{2,1}, \ldots, \sigma^*_{2,7}\,.$$

- To approximate the distribution of the statistic T under H_0 we can use a permutation test.
- Shuffle the 10+7=17 shapes into random order:

$$\Pi : \{1, 2, \ldots, 17\} \to \{1, 2, \ldots, 17\},$$

where Π is a random permutation of 1, 2, ... , 17. then partition the shuffled 17 shapes into two new groups:

$$\sigma^*_{1,1}, \ldots, \sigma^*_{1,10}; \quad \sigma^*_{2,1}, \ldots, \sigma^*_{2,7}.$$

- We computed the test statistic T^*, iterated 10000 times, and computed the number of times out of 10000 that $T^* > T$.

- To approximate the distribution of the statistic T under H_0 we can use a permutation test.
- Shuffle the 10+7=17 shapes into random order:

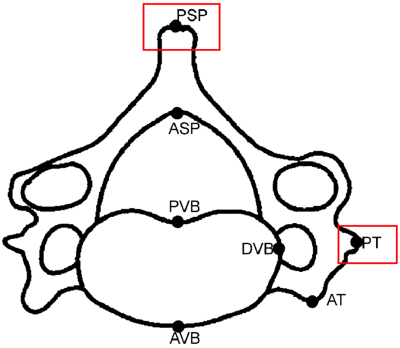
$$\Pi : \{1, 2, \ldots, 17\} \to \{1, 2, \ldots, 17\} ,$$

where Π is a random permutation of 1, 2, ... , 17. then partition the shuffled 17 shapes into two new groups:

$$\sigma^*_{1,1}, \ldots, \sigma^*_{1,10} ; \quad \sigma^*_{2,1}, \ldots, \sigma^*_{2,7} .$$

- We computed the test statistic T^*, iterated 10000 times, and computed the number of times out of 10000 that $T^* > T$.
- For the given data, we found only this to be true in only 0.07 % of cases!!!

A more detailed analysis – see Albert, Le & Small (2003) – shows that two landmarks are particularly responsible for most of the between sample shape variation.



Outline

- Albert MH, Le H, Small CG. Assessing landmark influence on shape variation. *Biometrika* 2003, **90**: 669–678.
- Dryden IL, Mardia KV. *Statistical Shape Analysis*. Chichester: Wiley; 1998.
- Goodall CR, Mardia KV. Multivariate aspects of shape theory. *Ann. Statist.* 1993, **21**: 848–866.
- Huckemann S. Intrinsic inference on the mean geodesic of planar shapes and tree discrimination by leaf growth. Accepted for *Ann. Statist.*, arXiv, 1009.3203 [stat.ME]
- Kendall DG. Shape manifolds, procrustean metrics, and complex projective spaces. *Bull. Lond. Math. Soc.* 1984, **16**: 81–121.
- Kendall DG, Barden D, Carne TK, Le H. *Shape and Shape Theory*. Chichester: Wiley; 1999.
- Small CG. *The Statistical Theory of Shape*. New York: Springer; 1996.
- Small CG. Statistics of shape. WIREs Comp. Stat. 2011, **3**: to appear.

Thank you