**Paper SA04_05**

# Statistical Shape Analysis with SAS

## Katherine Gerber, University of Virginia, Charlottesville, VA

## ABSTRACT

The field of statistical shape analysis involves methods for studying the geometrical properties of random objects invariant under translation, scaling and rotation. It is often extremely useful to measure, compare and categorize the shape of objects in a wide variety of disciplines, ranging from code recognition to medicine, archaeology, and geology. Shape techniques are most often applied to the area of biology known as morphometrics, the statistical study of biological shape and shape changes.

Utilities for data collection in pedagogical situations are widely available. With test data in hand, Base SAS®, SAS/STAT®, SAS/IML®, and SAS/GRAPH® are excellent tools for demonstrating the main ideas of shape analysis and performing statistical analysis on the shape data. An intermediate level of SAS programming is assumed; however, mathematically curious beginning level SAS programmers are likely to enjoy the material as well.

## INTRODUCTION

The familiar adage that a picture is worth a thousand words is a premise taken for granted when working with data and analyzing it statistically. Visual tools typically play a crucial role in deriving meaningful information from numerical data and communicating that information clearly and intuitively to others. Size and location, and somewhat less frequently, shape and color are used to communicate summary information about numerical data. Visual summaries in the form of charts and graphs are used to provide needed information with clarity and speed, often critical in the context of decision making. Even when the underlying data are as simple as four or five numbers and require very little intermediate interpretation, the much maligned pie chart applies an elementary geometric algorithm in order to convey their relative size to most audiences far more effectively than a listing of the data.

Suppose, however, that the geometric entities themselves serve as the starting point rather than the result. How, for example, can we arrive at legitimate statistical conclusions concerning polygonal samples that correspond to physical features of biological organisms? The field of **statistical shape analysis** involves methods for quantifying visual data and deriving information from it. Synthesizing techniques from statistics, geometry and more general mathematics, conclusions are arrived at through analysis and summaries of shape data objects.

Automatic object recognition is an obvious application of shape analysis. Perhaps not as well known to those working outside of the fields, traditional biological and medical applications study how shape

- Changes during growth
- Changes during evolution
- Is related to size
- Is affected by disease
- Is related to other covariates

Applications of shape analysis are indeed concentrated in the fields of biology, geology and medicine; however, the theory and techniques can be applied to appropriate configuration matrices regardless of the discipline from which they arise. Although shape analysis continues to be a developing field, existing mathematical underpinnings are rigorous and well developed, and a large body of relevant computational methods have been implemented, many of them in SAS®. The intention of this expository paper is to serve as a starting point for SAS® programmers who are interested in learning how to work with shape data or who may encounter a need for such analysis in the course of their future work.

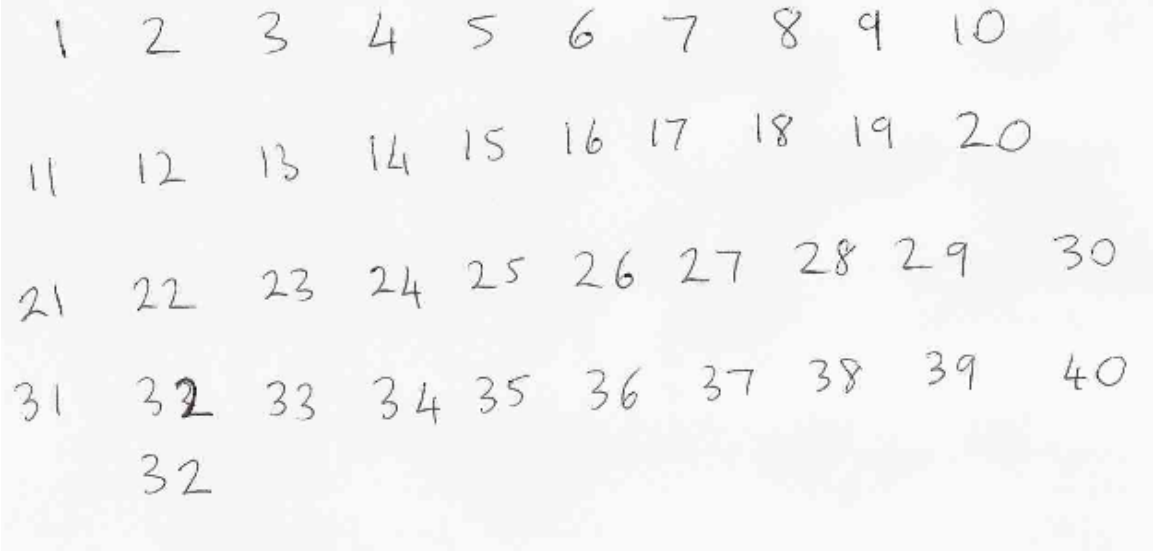
## SIMPLE DATA ACQUISITION OF SIMPLE SHAPE DATA

For someone just getting started with shape analysis, the morphometrics web site at SUNY Stony Brook (http://life.bio.sunysb.edu/morph/ ) is an indispensable resource for programs, test data and utilities in shape analysis. A glossary and comprehensive bibliography are also maintained at the site.

Jean-Pierre Dujardin's Collection of Coordinates (COO) program is available at the site and allows one to retrieve *x-y* coordinates of landmark data quickly and inexpensively from scanned images. The novice can begin with scanned images of leaves or handwriting samples, then the coordinate data and harmonic coefficients can be saved to a text file and the appropriate lines of data can be retrieved with SAS®.

The handwriting sample below provides a sample of twos and threes. By first establishing guidelines for landmark location and then collecting coordinates with a program such as COO, we have sufficient data for making use of the principles and algorithms of shape analysis. One can derive mean shapes for the twos and threes, test for shape difference, and determine whether a new shape object is more likely to be a two or three.

Such an exercise in obtaining data manually helps to make one aware of some of the pitfalls inherent in shape data collection. For example, the mean shape of a collection of twos will not at all resemble a two if an initial crook or loop is included in the outline of some of the shape samples and not included in others. It may be necessary to establish a rule that ignores such a feature.




## A LANDMARK APPROACH

More formally, **statistical shape analysis** involves methods for studying the shapes of objects where location, rotation and scale information can be removed. We will begin by focusing on situations where the objects under study are summarized by key points called landmarks, and a few preliminary definitions are needed.

**Shape** is all the geometrical information that remains when location, scale and rotational effects are filtered out from an object.

A **landmark** is a point of correspondence on each object that matches between and within populations. Each landmark is associated with Cartesian coordinates, that is, either with an ordered pair in the plane or with a triple in 3-space.

An **anatomical landmark** is a point assigned by an expert that corresponds between objects of study in a way meaningful in the context of the disciplinary context. In addition to the Cartesian coordinates, each landmark has a name denoting correspondence from shape object to shape object, for example, the point of the right elbow.

**Mathematical landmarks** are points located on an object according to some mathematical or geometrical property of the figure.

**Pseudo-landmarks** are constructed points on an object, either around the outline or in between anatomical or mathematical landmarks.

The **configuration** is the set of landmarks on a particular object.

The **configuration matrix** $X$ is the $k$ x $m$ matrix of Cartesian coordinates of the $k$ landmarks in $m$ dimensions.

The **configuration space** is the space of all possible landmark coordinates. In applications we have $k>=3$ landmarks in $m = 2$ or $m = 3$ dimensions.

## Examples of Mathematical Landmarks and Pseudo-landmarks

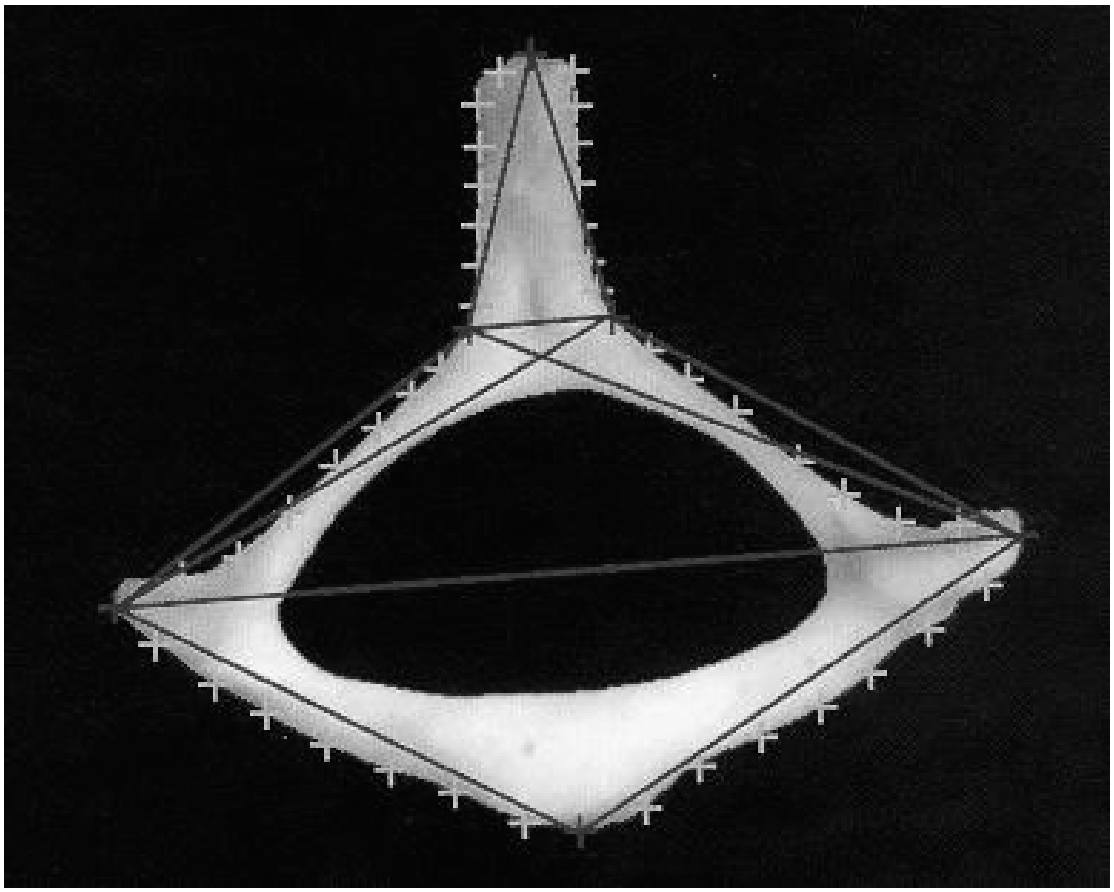


Image of a T2 mouse vertebra with six mathematical landmarks as well as 42 pseudo-landmarks (Book cover image from Dryden and Mardia)

A **size measure** $g(X)$ is any positive real valued function of the configuration matrix such that $g(aX) = ag(X)$ for any positive scalar $a$.

The **centroid size** is given by $S(X) = \|CX\| = \sqrt{\sum_{i=1}^{k}\sum_{j=1}^{m} (X_{ij} - \overline{X}_j)^2}$, $X \in \mathbb{R}^{km}$

where $X_{ij}$ is the $(i,j)th$ entry of $X$, the arithmetic mean of the $jth$ dimension is $\overline{X}_j = \frac{1}{k}\sum_{i=1}^{k} X_{ij}$, $C = I_k - \frac{1}{k}1_k 1_k^T$ is the centring matrix, $\|X\| = \sqrt{trace(X^T X)}$ is the Euclidean norm, $I_k$ is the $k \times k$ identity matrix and $1_k$ is the $k \times 1$ vector of ones.

$S(X)$ is the square root of the sum of squared Euclidean distances from each landmark to the centroid:

$$S(X) = \sqrt{\sum_{i=1}^{k} \|(X)_j - \overline{X}\|^2},$$

where $(X)_j$ is the $jth$ row of $X$ $(j = 1,\ldots,k)$ and $\overline{X} = (\overline{X}_1,\ldots\overline{X}_m)$ is the centroid.

The analysis of landmark data requires the choice of a suitable coordinate system so that the similarity transformations of translation, rotation and rescaling are removed. First a pair of baseline landmarks is selected. The first, landmark 1, will be sent to (-1/2,0), and landmark 2 will be sent to (1/2,0).

**Bookstein coordinates** - $(u_j^B, v_j^B)^T$, $j = 3,\ldots.k$ are the remaining coordinates of an object after translating, rotating, rescaling the baseline to $(-\frac{1}{2}, 0)$ and $(\frac{1}{2}, 0)$ so that

$u_j^B = \{(x_2 - x_1)(x_j - x_1) + (y_2 - y_1)(y_j - y_1)\}/D_{12}^2 - \frac{1}{2}$,

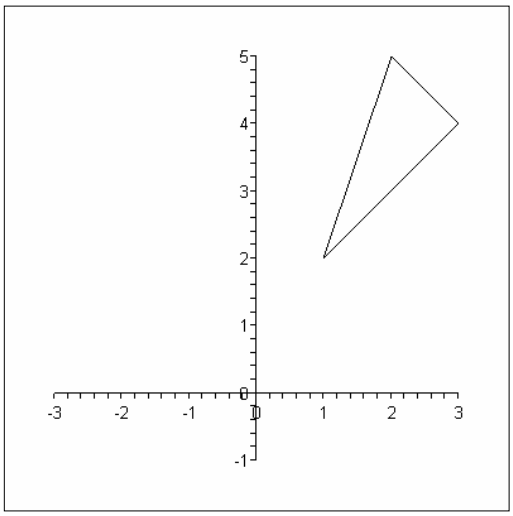
$v_j^B = \{(x_2 - x_1)(y_j - y_1) + (y_2 - y_1)(x_j - x_1)\}/D_{12}^2$,

where $j = 3,\ldots,k$, $D_{12}^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2 > 0$ and $-\infty < u_j^B, v_j^B < \infty$

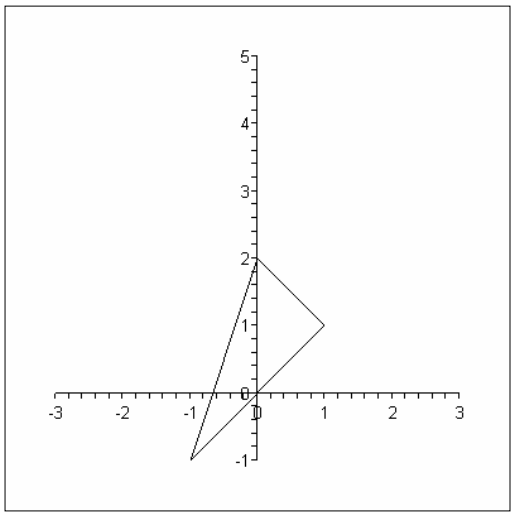
The choice of (-1/2, 0) and (1/2, 0) as the baseline is somewhat arbitrary, and baseline choice varies in the literature. Recalling that shape is defined residually as all the geometrical information that remains when location, scale and rotational effects are filtered out from an object, one notes that all line segments are identical after the appropriate transformations. A planar example using a simple triangle demonstrates that the process of obtaining Bookstein coordinates does indeed change each of the three variants. The process of obtaining Bookstein coordinates is illustrated geometrically in the following sequence of plots. Exact expressions were geometrically derived for the coordinates, and equivalent decimal values are used for plotting purposes.

First, location is changed by translation, the object is then rotated appropriately, and finally the triangle is scaled so that the resulting triangle is geometrically similar to the original triangle. Each of the transformations performed is invertible, that is, for each of the transformations performed, there exists a second transformation that is unique, and through composition will “cancel” the effect of the original transformation. Thus, by retaining information on each transformation applied, the original coordinates may be recovered precisely and no information is lost.

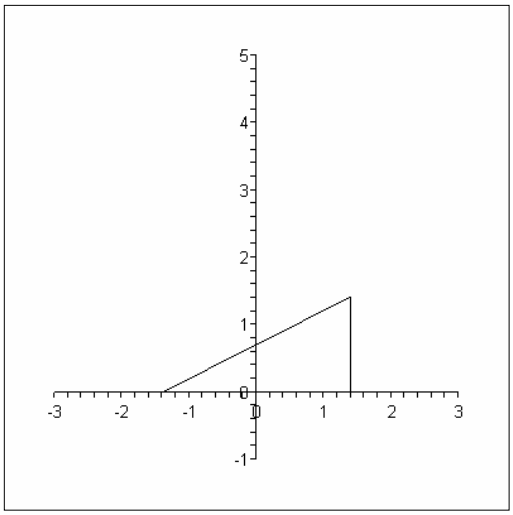
Plot (a) depicts the original triangle, in (b) it is translated, rotated in plot (c) and scaled in plot (d). Bookstein coordinates of (-1/2,0) and (1/2,0) for landmarks 1 and 2 are shown in plot (d), respectively, while the Bookstein coordinates for landmark 3 are determined completely by the shape of the original triangle regardless of size, orientation or position in the plane.



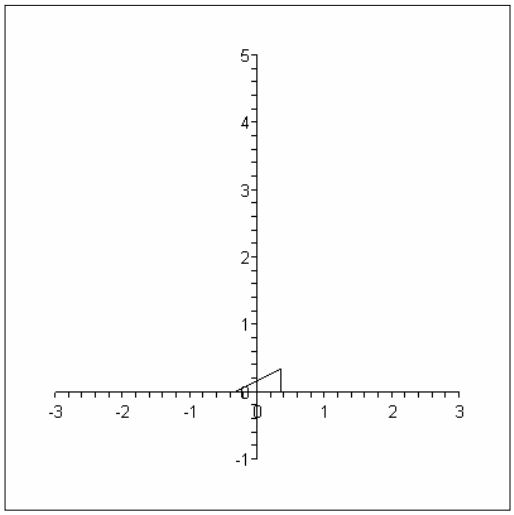

**(a) Coordinates: (1, 2), (3, 4), (2, 5)**
Starting with a triangle with vertices (1,2), (3,4) and (2,5). Landmarks (1,2) and (3,4) are the end points of the baseline.



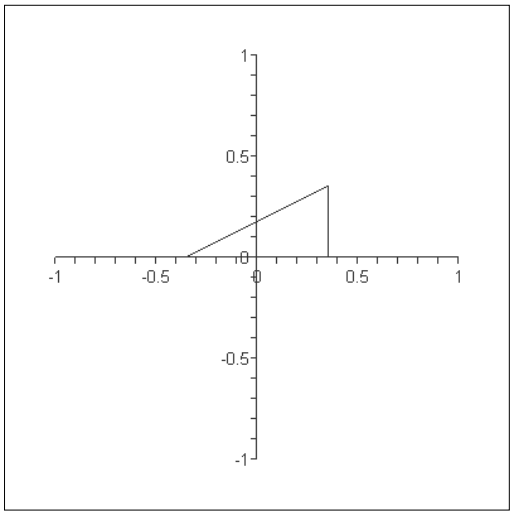

**(b) Coordinates: (-1, -1), (1,1), (0,2)**
Translated in the plane so that the baseline midpoint is at the origin.




**(c) Coordinates: (-2^(1/2), 0), (2^(1/2), 0), (2^(1/2),2^(1/2))**
Rotated so that the baseline points lie along the x-axis and the third vertex is above the x-axis.



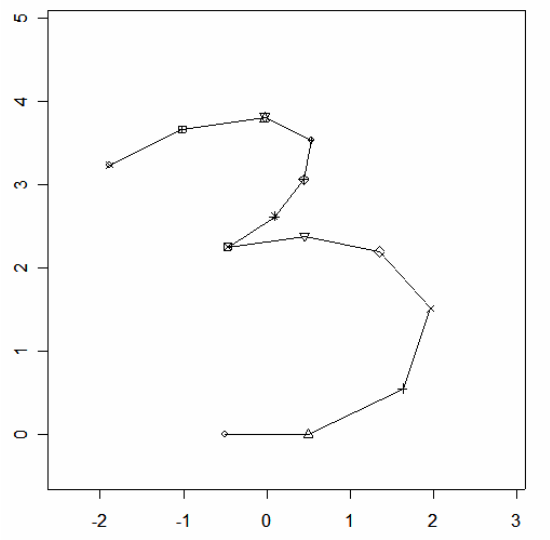

**(d) Coordinates: (-1/4*2^(1/2),0), (1/4*2^(1/2),0), (1/4*2^(1/2),1/4*2^(1/2))**
Scaled to a similitude so that the baseline points are at (-1/2,0) and (1/2,0).



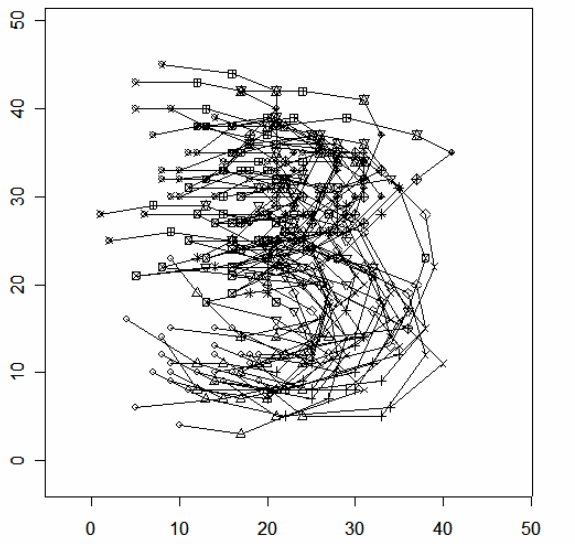

Unrelated to calculating the Bookstein coordinates, resizing the view port may render a more readable graphic, as in this particular case.

SAS® programs for obtaining Bookstein coordinates are widely available. In Marcus and Corti's SAS® code (Marcus and Corti) for a morphometrics workshop, several SAS/IML® procedures were developed for deriving and analyzing the Bookstein coordinates.

Returning to the mean shape concept mentioned in discussion of the handwriting example, a mean shape of a collection of 30 sets of Bookstein coordinates is shown. A plot of the raw data is shown on the right (Dryden, 2004).




**Mean shape for 30 observations**




**Joined lines for 30 observations**

**PROCRUSTES ANALYSIS**

Having derived suitable shape coordinates, further methods are needed in order to perform useful analysis on the transformed landmark data. With shape coordinates in hand a variety of statistical techniques and tools can be applied to the shape data to accomplish several tasks:

- Obtaining a measure of distance between the shape objects
- Estimating mean shapes from a random sample of shape objects
- Estimating shape variability from a random sample

Constraints of space and time permit only a brief sketch of a few of the core techniques regularly applied to shape data. Principal among these are methods of Procrustes analysis.

**Planar Procrustes analysis** is an approach that entails fitting objects by superimposition and is often a suitable procedure for investigation. Procrustes methods have a long history, having been employed as early as 1939 in psychometric applications (Mosier).

**Generalized Procrustes Analysis** (GPA) refers to the situation in which several objects are fitted using Procrustes superimposition.

**Ordinary Procrustes Analysis** (OPA) applies when a single object is fitted to another. OPA is NOT symmetrical in the ordering of the objects, whereas GPA is invariant under re-orderings of the objects.

**FULL PROCRUSTES DISTANCE**

Let $X_1$ and $X_2$ be a pair of configuration matrices with pre-shapes $Z_1$ and $Z_1$.
The **full Procrustes distance** between $X_1$ and $X_2$ is

$$d_F(X_1,X_2) = \inf_{\Gamma \in SO(m), \beta \in \mathbb{R}} \| Z_2 - \beta Z_1 \Gamma \|$$

**Result**: The full Procrustes distance is

$$d_F(X_1,X_2) = \left\{ 1 - \left( \sum_{i=1}^{m} \lambda_i \right)^2 \right\}^{1/2},$$

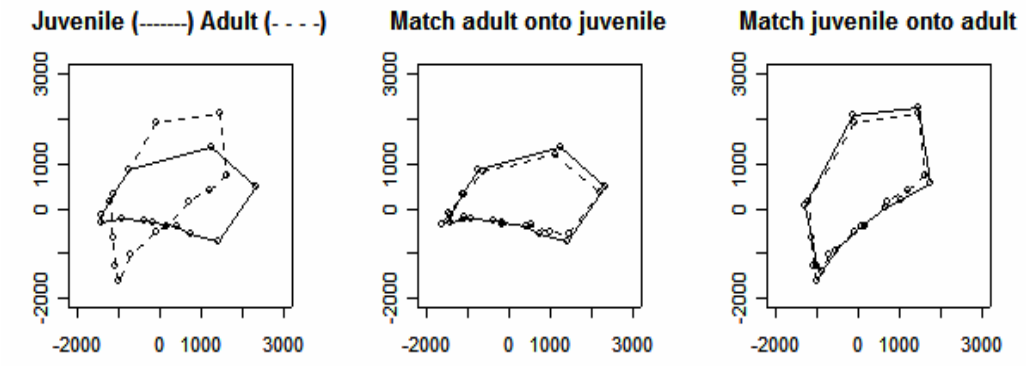
where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{m-1} \geq |\lambda_m|$ are the square roots of the eigenvalues of $Z_1^T Z_2 Z_2^T Z_1$, and the smallest value $\lambda_m$ is the negative square root iff $\det(Z_1^T Z_2) < 0$.
The minimizing rotation is given by

$$\hat{\Gamma} = UV^T$$

where $U, V \in SO(m)$ and $Z_2^T Z_1 = V \Lambda U^T$ with $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_m)$.

The minimizing scale is $\hat{\beta} = \sum_{i=1}^{m} \lambda_i$.

**Ordinary Procrustes analysis** matches one configuration to another using translation, rotation and (possibly) scale. Reflections can also be included if desired. The function matches one configuration onto another by least squares.

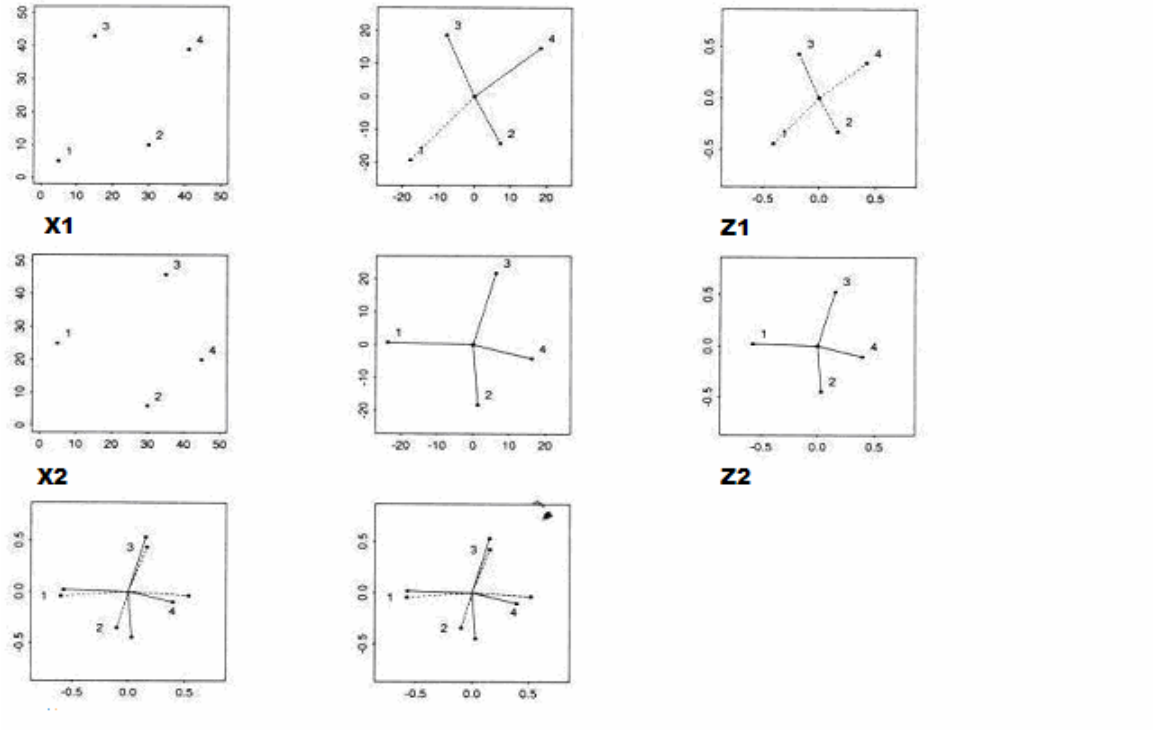



Adult onto Juvenile: Rotation angle: -45.52717 Scale: 0.8745017 OSS: 231146.5
Juvenile onto Adult: Rotation angle: 45.52717 Scale: 1.130936 OSS: 298926.7

Dryden and Mardia geometrically illustrate the difference between notions of partial and full Procrustes fits. From left to right in the first row figure X1 is illustrated then centered and finally rescaled on the right. The second row illustrates the same centering and rescaling for figure X2.

While the figures in the third row are difficult to distinguish visually, the figure on the left of the third row represents the partial Procrustes distance, a rotation of Z2 to Z1 that minimizes the sum of squared distances between pairs of landmarks. In the middle (right) figure on the third row Z2 is both rotated and rescaled to minimize the sum of squared distances between pairs of landmarks, representing the full Procrustes distance.

**Geometry of Procrustes Fits in Calculating Procrustes Distances (Dryden and Mardia)**




**MEAN SHAPE, COMPARING SAMPLES, AND SHAPE VARIABILITY**

Either Hotelling's $T^2$ or Goodall's F test can be carried out to examine differences in mean shape between two independent populations.

Principal component analysis of the sample covariance matrix in Procrustes tangent space coordinates provides an effective means of analyzing the main modes of variation in shape. Though not illustrated here, a frequently used method of visualizing the effect of each principal component is to evaluate and plot an icon for a few values of the standardized PC scores, c, in [-3,3] where c = 0 corresponds to the full Procrustes mean shape.

## OUTLINE ANALYSIS

Outline or contour analysis is based on digitizing a large number of points around the boundary of an object. For those situations in which landmarks are difficult to identify or obtain, if the outline can be represented by a closed curve or boundary, then **outline analysis** is the preferred approach to shape analysis. Many shape objects that do not have clearly identifiable landmarks nevertheless can be analyzed successfully under such conditions. Considering the planar outlines of certain familiar unicellular organisms such as an amoeba and a paramecium, the paucity of landmarks is immediately apparent, although the representations of the organisms can be easily differentiated visually.

Literature on Fourier analysis evolved primarily in the fields of mathematics and engineering rather than within the classical domains of shape analysis. Lestrel's (1997) text begins with an introduction and overview of Fourier techniques that is very helpful in bringing the reader up to speed quickly on topics needed for boundary and outline work.

The outline analytic approach is exemplified in Rohlf and Archie's (1984) classical study comparing Fourier methods in a study of the shape of mosquito wings. Digitizing the outlines of wings from 127 species of mosquitoes, the authors compared several methods of data description for use in multivariate analysis. The wings were oriented horizontally and a single anatomical landmark was chosen as center. Coordinates were computed for 100 equally spaced radii, and the data set then consisted of angle and radius pairs. Calculating the centroid and translating the data accordingly, a matrix of Fourier coefficients for 16 harmonics (0 through 15) was calculated. Using this method, the zeroth harmonic describes the contribution of a centered circle, the first harmonic is an offset circle, the second is a figure 8, and so on.

**Elliptic Fourier Analysis (EFA)** is a Fourier method that interpolates the outline to get a large number of points but relaxes the constraint of sampling at equal intervals. Traveling counterclockwise around the outline from a given starting point, we take note of the x and y increments from point to point, thereby defining two periodic functions which are independently subjected to Fourier analysis.

In such situations the shape of outlines are be represented with Fourier series. The Kuhl and Giardina functions are elliptical Fourier functions (EFF) parametrically defining the *x* and *y* positions in terms of a third variable, *t*. Here *n* is the number of the harmonic, and *N* denotes the maximum harmonic number:

$$x(t) = A_0 + \sum_{n=1}^{N} a_n \cos nt + \sum_{n=1}^{N} b_n \sin nt, \text{ and } y(t) = C_0 + \sum_{n=1}^{N} c_n \cos nt + \sum_{n=1}^{N} d_n \sin nt.$$

Rohlf and Archie's study determined that the EFF formulation produced the best results among several formulations studied.

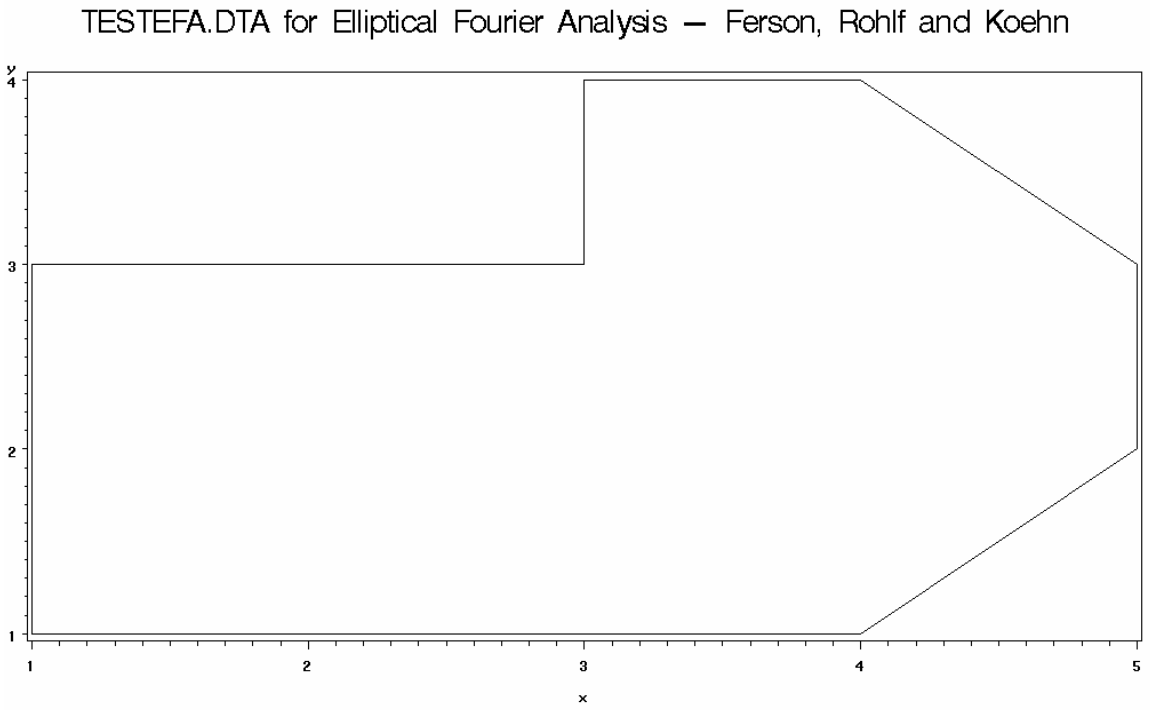
In contrast to the landmark methods, standardization of size, location and rotation are accomplished in outline methods through estimations obtained in the course of applying the algorithms. By calculating the *x* and *y* coordinates of the centroid of the enclosed region and subtracting these quantities from the input *x* and *y* coordinates, location is standardized. Through estimating the area of the ellipse determined by the first harmonic and dividing appropriate quantities by its square root, size is approximately standardized. The outline is rotated so that the major axis of the ellipse defined by the first harmonic is parallel to the *x*-axis.

Along with the Fortran program for obtaining EFA coefficients (Ferson, Rohlf and Koehn) the TESTEFA.DTA file is provided as a simple example of a data file. We are currently working on porting the EFA calculation component to SAS/IML®. Meanwhile, reading and plotting the data in SAS® is straightforward. Here the 13th data point is added as a repetition of the first in order to close the outline:

```
DATA testefa ;
INPUT PID $ x y ;
CARDS ;
1 1 1
2 1 2
3 1 3
4 2 3
5 3 3
6 3 4
7 4 4
8 5 3
9 5 2
10 4 1
11 3 1
12 2 1
13 1 1
;
RUN ;
```

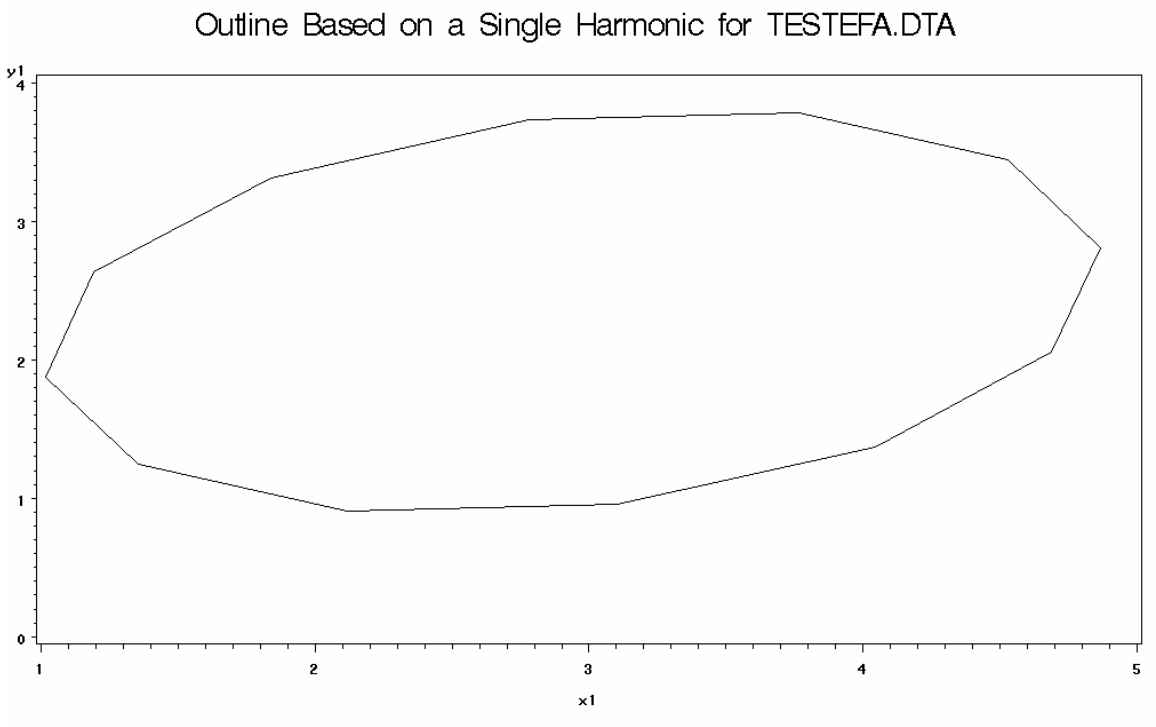
```
TITLE 'TESTEFA.DTA for Elliptical Fourier Analysis - Ferson, Rohlf and
Koehn' ;

PROC GPLOT DATA = testefa ;
   PLOT y*x ;
   RUN ;
   QUIT ;
```

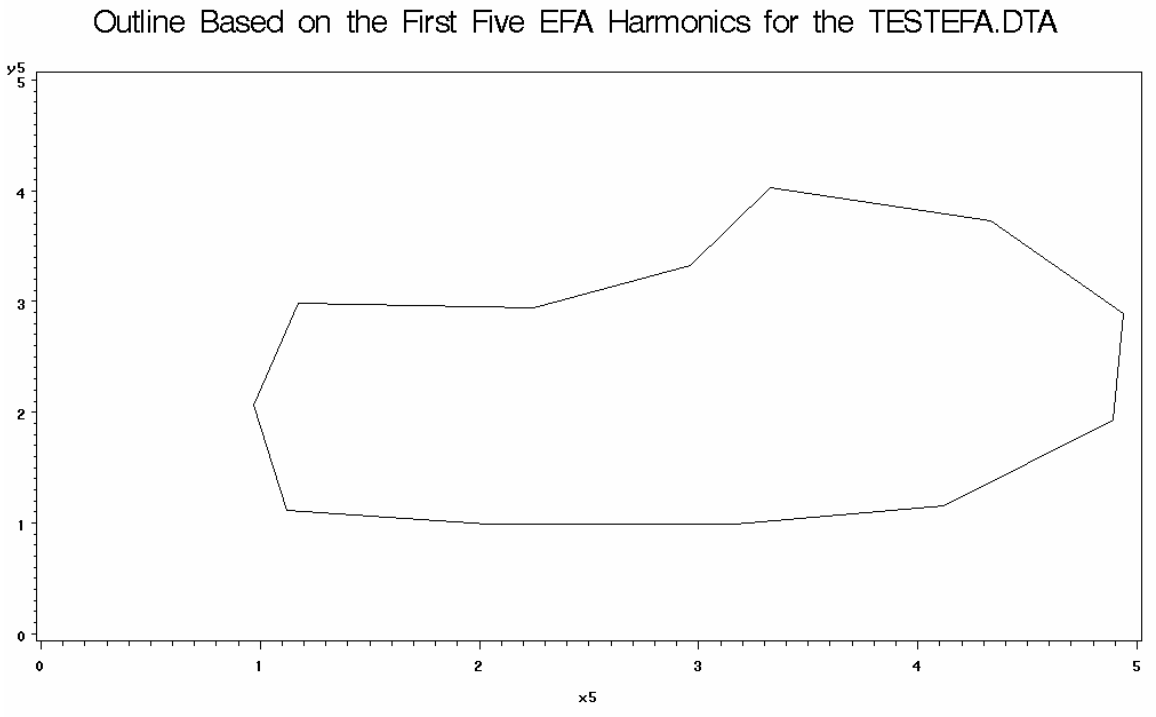



```
DATA efa1 ;
INPUT PID $ x1 y1 ;
CARDS ;
   1    1.35403900        1.24387200
   2    1.01572500        1.87695500
   3    1.19327700        2.63520900
   4    1.83912000        3.31545900
   5    2.78020100        3.73543400
   6    3.76435800        3.78260100
   7    4.52788700        3.44432200
   8    4.86620000        2.81123800
   9    4.68864900        2.05298500
  10    4.04280500        1.37273400
  11    3.10172400        .952759600
  12    2.11756800        .905592600
  13    1.35403900        1.24387200
;
RUN ;

TITLE Outline Based on a Single Harmonic for TESTEFA.DTA';
PROC GPLOT DATA = efa1 ;
   PLOT y1*x1 ;
   RUN ;
   QUIT ;
```

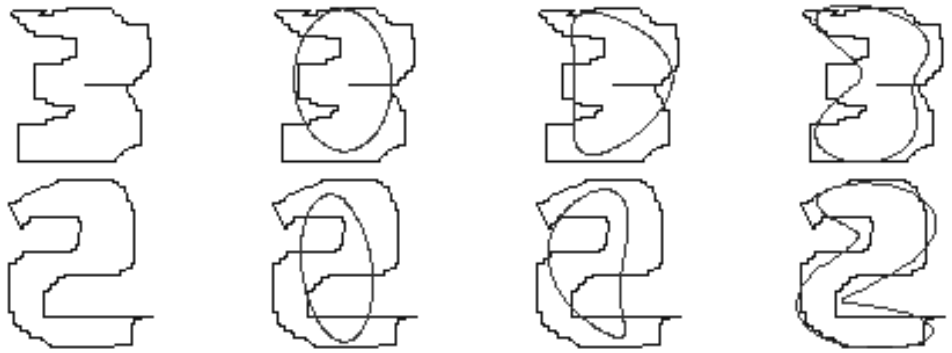



As expected, the first harmonic yields an ellipse above. Convergence to the outline of the original data is becoming very evident when five harmonics are used as shown below.

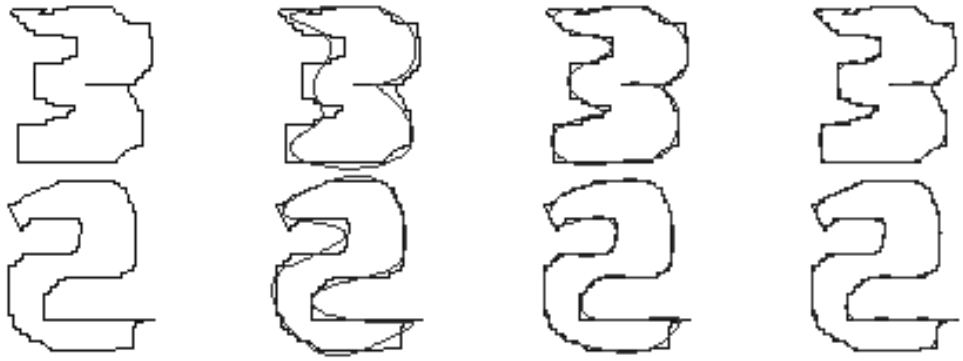


Using a large number of data points on the outline and higher order harmonics will of course lead to finer graphical representations and closer approximations. The next pair of graphics are based on 256 points from the sketch of the figure. In each the first harmonic representation is an obvious ellipse, and here the represenation using 20 harmonics can barely be distinguished from the original data plot.

**EFF Approximations to Crudely Hand Drawn Outlines of a "Three" and a "Two."**
Original drawings are on the left then each is summarized using 1, 2 and 3 harmonics.



**EFF Approximations to Crudely Hand Drawn Outlines of a "Three" and a "Two."**
The original drawings are on the left then each is summarized using 5, 10, and 20 harmonics



The EFA approximations above and associated numerical results were derived using the EFAWIN program (Isaev and Denisova).

## CONCLUSIONS

Statistical shape analysis may be thought of as a cross-disciplinary field lying at the theoretical foundation of domains of application such as facial recognition and pattern analysis. Additionally, shape data presents itself in garden-variety situations, and performing certain statistical analyses on fairly small data sets of simple shape data is well within the scope of the current desktop

Two key approaches to shape data representation, landmarks and outlines, have been examined. Rooted in 2-dimensional geometry, the landmark approach may be more accessible to most programmers, while outline analysis requires familiarity with elements of Fourier analysis. Although outline analysis does require a bit more mathematical background, the payoff is that the techniques may be applied to shape data that is ill suited for landmark analysis. While we have barely scratched the surface of statistically analyzing the shape data, the methods in Procrustes analysis lead to meaningful information rather easily.

A significant portion of shape analysis program development began within SAS® and additional programs have been ported to the language. For example, the SAS® code for obtaining Bookstein coordinates is widely available. It is likely that as interest in outline analysis grew, most researchers chose not to reinvent the wheel and turned to Fortran implementations of Fourier techniques that have been solidly in place for decades.

Much of the existing shape analysis code for coordinate systems makes use of SAS/IML SAS® which would also be the natural habitat for porting code snippets that may not be readily available to those experimenting with shape analysis algorithms.

Finally, several significant topical areas of shape analysis have not been addressed here, but are worthy of mention:

- Shape coordinate systems other than Bookstein
- Allometry, the relationship between shape and size
- Deformation analysis including relative warps, smoothing splines and tangent space methods
- Shape data embedded in images
- High-level Bayesian image analysis
- Extension of the methods presented to problems in three dimensions

## REFERENCES


Adams, Dean C., Rohlf, F. James, Slice, Dennis E. (2002), *Geometric Morphometrics: Ten Years of Progress Following the 'Revolution'*, Italian Journal of Zoology, 71:5-16.

Bookstein, Fred L (1991). *Morphometric Tools for Landmark Data,* Cambridge University Press, Cambridge.

Dryden, Ian L. and Mardia, Kanti V. (1998). *Statistical Shape Analysis*, West Sussex, England: John Wiley & Sons, Ltd.

Dryden, Ian (2004). shapes: Statistical shape analysis, R package version 1.0-8.

Ferson, S. F., F. J. Rohlf, and R. K. Koehn (1985). *Measuring shape variation of two-dimensional outlines*. Systematic Zoology, 34:59-68.

Isaev M.A., Denisova, L.N. (1995). *The computer programs for shape analysis of plant leaves*, Proceedings of the Mathematics. Computer. Education International conference, Pushchino.

Kuhl, F P and C R Giardina (1982). *Elliptic Fourier Features of a Closed Contour,* Computer Graphics and Image Processing, 18:236-258.

Lestrel, Pete E., (1997). *Fourier Descriptors and Their Applications in Biology,* Cambridge, Cambridge University Press.

Marcus,L.F. and Corti, M. (1992). *Data Analysis in Systematics*, workshop, I Congreso Latinoamericano de Teriologia), Caracas, Venezuela.

Mosier, C.I. (1939). *Determining a simple structure when loadings for certain tests are known*, Psychometrika, 4:149-162.

Reyment, R. (1991). *Multidimensional Paleobiology.* New York, Pergamon Press: New York. SAS/IML®, code supplement by Les Marcus.

Rohlf, F. James, Archie, James W (1984). *A Comparison of Fourier Methods for the Description of Wing Shapes in Mosquitos (Diptera: Culicidae).* Systematic Zoology, 33: No 3 302-317.

Rohlf, F. James (1990). *Morphometrics,* Annual Review of Ecology and Systematics, 21:299-316.


## ACKNOWLEDGMENTS


I wish to thank University of Virginia economics graduate student, Yakup Asarkaya, for kindly contributing the handwriting samples and interesting discussion of these topics.

## CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Katherine Gerber
University of Virginia
ITC – Research Computing Support Group
PO Box 400779
Charlottesville VA 22904
Work Phone: 434-982-4986
E-mail: kmg5b@virginia.edu

## TRADEMARK CITATIONS