# Bootstrap and Empirical Likelihood Methods in Statistical Shape Analysis

by Getulio J. A. Amaral

Thesis submited to The University of Nottingham for the degree of

Doctor of Philosophy, August 2004

# Contents

**Abstract**

The aim of this thesis is to propose bootstrap and empirical likelihood confidence regions and hypothesis tests for use in statistical shape analysis.

Bootstrap and empirical likelihood methods have some advantages when compared to conventional methods. In particular, they are nonparametric methods and so it is not necessary to choose a family of distribution for building confidence regions or testing hypotheses.

There has been very little work on bootstrap and empirical likelihood methods in statistical shape analysis. Only one paper (Bhattacharya and Patrangenaru, 2003) has considered bootstrap methods in statistical shape analysis, but just for constructing confidence regions. There are no published papers on the use of empirical likelihood methods in statistical shape analysis.

Existing methods for building confidence regions and testing hypotheses in shape analysis have some limitations. The Hotelling and Goodall confidence regions and hypothesis tests are not appropriate for data sets with low concentration. The main reason is that these methods are designed for data with high concentration, and if this hypothesis is violated, the methods do not perform well.

On the other hand, simulation results have showed that bootstrap and empirical likelihood methods developed in this thesis are appropriate to the statistical shape analysis of low concentrated data sets. For highly concentrated data sets all the methods show similar performance.

Theoretical aspects of bootstrap and empirical likelihood methods are also considered. Both methods are based on asymptotic results and those results are explained in this thesis. It is proved that the bootstrap methods proposed in this thesis are asymptotically pivotal.

Computational aspects are discussed. All the bootstrap algorithms are implemented in

"R". An algorithm for computing empirical likelihood tests for several populations is also implemented in "R".

**Acnowledgements**

I would like to thank my first supervisor professor Andrew Wood who gave all necessary help in all the steps of this thesis. I also would like to thank my second supervisor professor Ian Dryden who supervised this work in the first two years of the PhD.

I also would like to thank Francisco Cribari-Neto and Gauss M. Cordeiro, for their help before the PhD.

I dedicate this thesis to Severina Amaral, Jose Amaral, Vitor, Debora e Daniel. I also would like to thank god.

**Notation**

| Symbol– | Meaning | Page Number |
|---|---|---|
| $k$ | Number of landmarks | 14 |
| $m$ | Number of dimensions | 14 |
| $\mathrm{Y} = \begin{pmatrix} y_{1,1} & y_{1,2} \\ \vdots & \vdots \\ y_{k,1} & y_{k,2} \end{pmatrix}$ | Configuration matrix | 14 |
| $\mathbb{R}^m$ | m-dimensional Euclidean space | 14 |
| $z^0$ | Complex coordinates for the landmarks | 15 |
| $H^F$ | Helmert matrix | 15 |
| $H$ | Helmert sub-matrix | 16 |
| $w = Hz^0$ | Helmertized configuration | 16 |
| $1_k$ | Vector of ones | 16 |
| $I_k$ | $k \times k$ Identity matrix | 16 |
| $n$ | Number of observations | 16 |
| $z_i = w_i/\|\|w_i\|\|$ | Pre-shapes | 17 |
| $CS^{k-1}$ | $(k-1)$ dimensional pre-shape space | 17 |
| $\mathbb{C}^k$ | Complex space | 17 |
| $\hat{S} = \sum_{i=1}^{n} z_i z_i^\star$ | Complex SSP matrix | 18 |
| $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_{k-1}$ | Eigenvalues of $\widehat{S}$ | 19 |
| $\widehat{\mu}$ | Full procrustes mean shape | 19 |

**Notation (Cont.)**

| Symbol– | Meaning | Page Number |
|---|---|---|
| $\widehat{\mu}_1, \ldots, \widehat{\mu}_{k-1}$ | Eigenvectors of $\widehat{S}$ | 19 |
| $\widehat{\mu} = \widehat{\mu}_1$ | Sample mean shape | 19 |
| $w_i^P = w_i^\star \widehat{\mu} w_i / (w_i^\star w_i)$ | Full procrustes coordinates | 20 |
| $v_i$ | Tangent coordinates | 21 |
| $N_k(.)$ | $k-$dimensional real multivariate normal | 22 |
| $CN_k(.,.)$ | Complex normal | 23 |
| $f(z) = \frac{1}{\pi^{k-1}\lvert\Sigma\rvert} e^{-(z-\mu)\Sigma^{-1}(z-\mu)}$ | Density of the complex normal distribution | 23 |
| $f(z) = c(A)^{-1} \exp(z^\star A z)$ | Density of the complex Bingham distribution | 24 |
| $Z$ | Random variable on the pre-shape space | 24 |
| $S_v$ | Covariance matrix of the tangent coordinates | 27 |
| $F$ | Hotelling's $T^2$ statistic | 30 |
| $S^+$ | Moore-Penrose inverse of $S$ | 30 |
| $d_F^2(z_i, z_j)$ | Squared procrustes distance | 31 |

**Notation (Cont.)**

| Symbol | Meaning | Page Number |
|---|---|---|
| $G$ | Goodall's statistic | 32 |
| $D$ | Mahalanobis distance | 33 |
| $H$ | Hotelling statistic (two sample case) | 33 |
| $G_T$ | Goodall statistic (two sample case) | 34 |
| $u^{(b)} = \{u_1^{(b)}, \ldots, u_n^{(b)}\}$ | Bootstrap sample | 42 |
| $R_\alpha = \{\nu : T_u^2(\nu) \leq t_\alpha^{(B)}\}$ | Bootstrap confidence region | 43 |
| $S^d$ | $d$ dimensional real sphere | 45 |
| $T(m)$ | Statistic of Fisher et al. (1996) | 45 |
| $\widehat{m}$ | Mean polar axis | 44 |
| $R_\alpha = \{m : T(m) \leqslant t_\alpha^{(B)}\}$ | Bootstrap confidence region for $m$ | 46 |
| $T(\mu)$ | Modified $T$ statistic in complex case | 49 |
| $\widehat{\Sigma}$ | Covariance matrix used in $T(\mu)$ | 49 |
| $\widehat{M}_{k-2} = [\widehat{\mu}_2, \ldots, \widehat{\mu}_{k-1}]^\star$ | Matrix of the orthogonal eigenvectors of $\widehat{S}$ | 49 |
| $\mathrm{R}_\alpha = \{\mu : T(\mu) \leqslant t_\alpha^{(B)}\}$ | Bootstrap confidence region for $\mu$ | 51 |

**Notation (Cont.)**

| Symbol | Meaning | Page Number |
|---|---|---|
| $F(\mu)$ | Statistic of Mahalanobis bootstrap method | 52 |
| $y^{[j]}$ | Sample of configurations from the $jth$ population | 82 |
| $\mathrm{z}^{[j]}$ | Sample of pre-shapes from the $jth$ population | 82 |
| $\widehat{\Sigma}^{[j]}$ | Covariance matrix for the $jth$ group | 84 |
| $\widehat{M}_{k-2}^{[j]}$ | Eigenvector matrix for the $jth$ group | 84 |
| $F_B(\widehat{\mu})$ | Statistic for bootstrap hypothesis test | 84 |
| $EL(m)$ | Empirical likelihood function | 107 |
| $R(\nu)$ | Empirical likelihood ratio | 111 |
| $EL(\mu)$ | Empirical likelihood for the mean shape | 118 |

# Chapter 1

# Introduction

In this chapter background on shape analysis is given and notation for describing shape data is presented. Extensive accounts of shape analysis are given in the monographs by Dryden and Mardia (1998), Small (1996) and Kendall et al. (1999).

In §1.1, the main ideas of statistical shape analysis are considered. A review of the literature about shape analysis is the topic of §1.2. The mathematical representation of shape and concepts such as the mean shape are reviewed in §1.3. In §1.4, coordinate systems including Procrustes coordinate systems and tangent coordinate systems are considered. Two relevant distributions, the complex normal and complex Bingham distributions, and techniques for their simulation, are studied in §1.5. How tangent coordinates can be used to obtain confidence regions for the mean shape via a normal approximation is reviewed in §1.6. Hypothesis tests for a single population are considered in §1.7 and for several populations in §1.8.

Readers who are familiar with shape analysis may wish to skip foward to §1.9.

## 1.1 Main Ideas of Shape Analysis

The study of the shape of random objects has received increasing attention in several disciplines. Advances in computer technology have made easier the capture and manipulation of images of objects. This information can be used to answer relevant questions in many disciplines including biology, medicine, archeology and computer vision. Some examples of objects which have been studied are mouse vertebrae, gorilla skulls and magnetic resonance brain scans.

The concept of the shape of an object plays an essential role in this study. Statistical shape analysis is concerned with summaries and comparisons of shapes of objects.

Some steps have to be carried out in order to represent the shape of an object in a mathematically convenient way. A convenient approach is to place landmarks on the object, which are points for identifying special locations on the object. The numerical coordinates of the landmarks are then used to represent an object. These coordinates belong to a space which is called the landmark space. The information about the shape of an object is what is left after allowing for the effects of translation, scale and rotation (Kendall, 1984).

A new set of coordinates of an object, which will be called pre-shape coordinates, can be obtained from the coordinates of that object in the landmark space. Suitable transformations are used to remove the effects of scale and translation. The new coordinate system also represents a mapping from the landmark space to the a new space. The new space is called pre-shape space.

We shall primarily concentrate on shapes of objects in two dimensions, i. e. planar shapes.

Two important summaries of a random sample of objects, the mean shape and the product matrix (or ssp), can be calculated using the pre-shape coordinates. The product matrix represents the variation of the pre-shape coordinates and the mean shape is defined as the eigenvector associated to the largest eigenvalue of this matrix.

The shape is finally obtained by removing the rotation information in the pre-shape coordinates of an object. The rotation information is eliminated by rotating an object to be as close as possible to a template. The new set of coordinates of the object are inside a new space, which is called shape space.

The pre-shape and shape spaces are non-Euclidean spaces. It is therefore difficult to perform standard statistical analyses on those spaces. To avoid the difficulties of non-Euclidean spaces it is possible to define a linear approximation to the space. A tangent space is a local linear approximation to the space at a particular point. For a given random sample of objects, the pre-shape coordinates of those objects can be projected on the tangent space at the sample mean shape. The new coordinates are called tangent coordinates.

Inference methods in shape analysis are often carried out in the tangent space. Such methods work better when the data are highly concentrated. In the tangent space many commonly used procedures of standard linear multivariate analysis are available. For example, shape variability can be studied by applying principal components analysis to the tangent coordinates.

There are some other possible approaches to statistical shape analysis which are not considered in this thesis. Possibilities include size-and-shape analysis, reflection shape analysis and reflection size-and-shape analysis. In the size-and-shape statistical analysis of objects, the information about size is retained, and the information about rotation and location is discarded.

If one wants to perform a reflection shape study of objects, the information about reflection should be removed from the shapes of those objects. Similarly, if one wants to perform a reflection size-and-shape study of objects, the information about reflection should be removed from the size-and-shapes of those objects (see Dryden and Mardia, 1998, p. 57).

## 1.2 Literature Review

The first work on statistical shape analysis was done by Kendall (1977). In a later paper, Kendall (1984) gives a more complete description of the research field. Several important concepts including shape spaces, shape manifolds, Procrustes analysis and shape densities are presented and discussed in depth. He also clarifies the differences between statistical shape analysis and the theory of shape which is studied by topologists.

In Kendall (1984) a system of coordinates is also introduced; we refer to this later as Kendall's coordinate system. One interesting fact about this system is that the location is removed by the use of a special matrix, the Helmert matrix. An important contribution of Kendall (1984) was the mathematical definition of shape, where he defines a mathematical space to represent the shape of a labelled set of $k$ points in $m$ dimensions.

On the other hand, Bookstein (1984, 1986) presents a mathematical basis for the study of morphometrics. In this case the objects under consideration are from disciplines such as biology and medicine, and have landmarks chosen according to some biological or medical features. He also introduces what is known as Bookstein's coordinate system, which removes the effects of translation, rotation and scale by manipulating two of the landmarks in such a way that they will be in fixed position.

When invited to comment the paper of Bookstein (1986), Kendall (1986) established the connection between their two theories. Kendall's labelled set of $k$ points in $m$ dimensions corresponds to Bookstein's landmarks. Even though they use different ways of calculating size and different coordinates systems, their ideas are quite similar in the sense of representing the shape of an object as a point in a manifold.

Procrustes analysis can be considered as a methodology for estimating for, a particular set of objects, the "optimal" scaling transformation, rotation transformation and translation transformation. The topic of Procrustes analysis was fully studied by Goodall (1991) who defined the mean shape in terms of Procrustes analysis. If the sum of squared distances between a point and the pre-shapes is minimal, then this point is said to be the mean shape.

A Gaussian model for the landmarks is also introduced by Goodall (1991). This model has a parameter for each transformation: scale, rotation and translation. Goodall (1991) also presented some algorithms to perform Procrustes analysis including an algorithm for ordinary procrutes analysis which minimizes the sum squares of the distances between two observations, and a more general method using weighted least squares. He also presented an iterative algorithm for estimating the transformations with several observations. This second algorithm is called the generalized Procrustes analysis.

After applying the transformations to the pre-shapes, the Procrustes fit coordinates are obtained. The mean shape also can be obtained as the mean of those coordinates.

Goodall also defined tests for shapes in the one and two population cases. Those tests were based on statistics of F-ratio and Hotelling's $T^2$ type. The F-ratio test is called Goodall's test in the literature.

Mardia and Walder (1994) considered tests for paired landmark data. They used a Gaussian model for the landmarks, where for each object there are two observations. The case of two x-rays for the same object was given as an example. They proposed a paired shape density, and they used this density to perform inference. They estimated the parameters of this distribution by maximum likelihood and they derived a likelihood ratio statistic, which can be used for testing hypotheses and for building confidence regions.

An important probabilistic model for statistical shape analysis is presented by Kent (1994). This model was the complex Bingham distribution, a complex version of the real Bingham distribution. One important property of the complex Bingham distribution is complex symmetry. This complex symmetry means that a vector and any rotated version of this vector will have the same distribution. This property is useful because shape analysis can be performed while working with pre-shapes.

The complex Watson distribution, which is a special case of the complex Bingham distribution, was discussed by Mardia and Dryden (1999). Maximum likelihood estimation and hypothesis testing procedures are considered, and they also illustrate how to use this distribution in shape analysis.

Kent (1997) introduced a method for calculating the mean shape which is resistant to outliers for landmark data in two dimensions. His model uses an angular central Gaussian distribution for the pre-shapes. The mean shape is calculated by maximum likelihood estimation using the EM algorithm.

The geometry of the shape space is studied by Kendall (1984), Le and Kendall (1993) and Kendall et. al (1999). See also Dryden and Mardia (1998, Ch 5, 7).

## 1.3 Mathematical Representation of Shape

Let $Y$ be a $k \times m$ matrix of Cartesian coordinates of $k$ landmarks in $m$ dimensions which is given by

$$Y = \begin{pmatrix} y_{1,1} & \cdots & y_{1,m} \\ \vdots & \ddots & \vdots \\ y_{k,1} & \cdots & y_{k,m} \end{pmatrix}. \tag{1.1}$$

A configuration are a set of landmarks on a particular object and the matrix $Y$ is usually called a configuration matrix.

The shape of a configuration matrix is obtained by removing the information about isotropic scaling, location and rotation. The shape space is the set of all possible shapes. The dimension of the shape space associated to objects with $k$ landmarks in $m$ dimension is

$$km - m - 1 - m(m-1)/2.$$

The term $km$ is the total dimension of the configuration matrix $Y$ and we subtract $m$, 1 and $m(m-1)/2$ as a consequence of removing location, scale and rotation respectively (see Dryden and Mardia, 1998, p. 56).

The landmark space is a real space $\mathbb{R}^m$ where the Cartesian coordinates of each landmark are represented. For example, for two dimensional objects, $m = 2$, and the landmark space is $\mathbb{R}^2$. In this thesis, the focus is exclusively on the case $m = 2$.

Some transformations need to be performed on the matrix $Y$ in order to remove the effects

of location, scale and rotation. When $m = 2$, the configuration matrix may be written as a complex vector. Define a $k \times 1$ complex vector

$$z^0 = (y_{1,1} + iy_{1,2}, \ldots, y_{k,1} + iy_{k,2})^T = (z^0_{(1)}, \ldots, z^0_{(k)})^T, \tag{1.2}$$

which corresponds to complex coordinates for the landmarks. The superscript $^0$ is used to indicate that the configuration retains the effects of location, scale and rotation. The details of each transformation in the case $m = 2$ will be given below.

The first step is to remove location. This can be done in various ways, depending on the coordinate system. Kendall's coordinates will be used here. Details about the Helmert matrix and Helmert sub-matrix are needed for Kendall's coordinate system. The Helmert sub-matrix provides a particular linear transformation which removes location by pre-multiplying $z^0$ (see Small, 1996, p. 130, and Dryden and Mardia, 1998, p. 34).

The full Helmert matrix $H^F$ is a $k \times k$ orthogonal matrix whose first row has all elements equal to $1/\sqrt{k}$, and has row $j + 1$ for $j \geq 1$ given by

$$(h_j, \ldots, h_j, -jh_j, 0, \ldots, 0), \;\; h_j = -\{j(j+1)\}^{-1/2},$$

with $j = 1, \ldots, k-1$, where the number of zeros elements in the row $j+1$ is equal to $k-j-1$. For example, if the number of landmarks is 5, the full Helmert matrix is given by

$$H^F = \begin{pmatrix} 1/\sqrt{5} & 1/\sqrt{5} & 1/\sqrt{5} & 1/\sqrt{5} & 1/\sqrt{5} \\ -1/\sqrt{2} & 1/\sqrt{2} & 0 & 0 & 0 \\ -1/\sqrt{6} & -1/\sqrt{6} & 2/\sqrt{6} & 0 & 0 \\ -1/\sqrt{12} & -1/\sqrt{12} & -1/\sqrt{12} & 3/\sqrt{12} & 0 \\ -1/\sqrt{20} & -1/\sqrt{20} & -1/\sqrt{20} & -1/\sqrt{20} & 4/\sqrt{20} \end{pmatrix}.$$

It can be shown by direct calculation that the Helmert matrix $H^F$ is an orthogonal matrix. The location of the complex configuration $z^0$ is removed by multiplying it by the $(k-1) \times k$ Helmert sub matrix, which is the Helmert matrix $H^F$ with the first row removed. The Helmert sub-matrix will be called $H$. The Helmertized configuration is given by

$$w = Hz^0. \tag{1.3}$$

A configuration is said to be *centered* if $1_k^T z^0 = 0$ where $1_k$ is a $k \times 1$ vector of ones. Helmertized configurations are connected to the centered configurations by the following property of the Helmert matrix (see Dryden and Mardia, 1998, p. 54):

$$H^T H = I_k - \frac{1}{k} 1_k 1_k^T,$$

where $I_k$ is a $k \times k$ identity matrix and $1_k$ is a $k \times 1$ vector of ones. Moreover, since $H^F$ is orthogonal, it follows that $H^T H = I_{k-1}$. Thus, if the $(k \times 1)$ vector $z^0 = (z^0_{(1)}, \ldots, z^0_{(k)})^T$ is a complex configuration, then

$$(I_k - \frac{1}{k} 1_k 1_k^T) z^0 = z^0 - \bar{z}^0 1_k,$$

where $\bar{z}^0 = k^{-1}\sum_{i=1}^{k} z^0_{(i)}$. Therefore, since $\mathrm{z}^0 - \bar{z}^0 1_k$ is a centered configuration, it means that the centered configurations are equal to the Helmertized configurations multiplied by $H^T$. So it always possible to obtain the Helmertized configurations from the centered configurations and vice versa.

The scale can be removed from the Helmertized configuration $w$ using

$$z = w/\sqrt{w^\star w} = H\mathrm{z}^0/\sqrt{(H\mathrm{z}^0)^\star H\mathrm{z}^0}, \tag{1.4}$$

where $w^\star$ is the complex conjugate transpose of $w$. The vector $z$ is called the pre-shape of the complex configuration $\mathrm{z}^0$. This name was coined by Kendall (1984). Note that a pre-shape is a shape with rotation information retained.

The concept of pre-shape space will be reviewed because it plays an important role (see Dryden and Mardia, 1998, p. 59 and Small, 1996, p. 9). The pre-shape space is the space of all possible $k-1$ complex vectors that do not have translation and scale information. Thus the pre-shape space is a unity complex hypersphere in $(k-1)-$dimensional complex dimensions; i.e.

$$CS^{k-1} = \{z \in \mathbb{C}^{k-1} : z^\star z = 1\}, \tag{1.5}$$

where $\mathbb{C}^{k-1}$ is $(k-1)-$dimensional complex space.

The shape space can be thought of as the pre-shape space with rotation information removed. The rotation information in the pre-shape vector $z$ can be eliminated by defining the equivalence class

$$[z] = \{e^{i\theta} z : \theta \in [0, 2\pi)\}, \tag{1.6}$$

where $[z]$ is identified with any of its rotated versions. Kendall (1984) notes that the shape space when $m = 2$ is the complex projective space $\mathbb{C}P^{k-2}$, the space of complex lines passing thought the origin.

An important problem of shape analysis is to estimate the average shape of a random sample of configurations. Consider $z_1^0, \ldots, z_n^0$ as a random sample of complex configurations from a population of objects $\Pi$, where each $z_i^0$ is defined by $(1.2)$.

Let $z_1, \ldots, z_n$ be the pre-shapes of $z_1^0, \ldots, z_n^0$, where $z_i$ is defined via $(1.4)$ and $z_i \in CS^{k-1}$. The full Procrustes mean shape $\widehat{\mu}$ can be found as the eigenvector corresponding to the largest eigenvalue of the complex sum of squares and product (SSP) matrix which is defined by (see Kent, 1994)

$$\hat{S} = \sum_{i=1}^{n} z_i z_i^{\star}.$$

Since the complex matrix $\widehat{S}$ satisfies the condition that $\widehat{S} = \widehat{S}^{\star}$, this matrix is Hermitian. Provided that the underlying distribution of the pre-shapes has a density with respect to the uniform distribution on the pre-shape sphere and $n \geq k - 1$, as opposed to being concentrated on a subspace, then $\widehat{S}$ has full rank with probability 1. So, applying the spectral decomposition theorem for Hermitian matrices which is given in Theorem $(A.1)$ in appendix A , $\widehat{S}$ is written as

$$\hat{S} = \sum_{j=1}^{k-1} \widehat{\lambda}_j \widehat{\mu}_j \widehat{\mu}_j^{\star}, \tag{1.7}$$

where $\widehat{\lambda}_1 \geqslant \widehat{\lambda}_2 \ldots \geqslant \widehat{\lambda}_{k-1} \geqslant 0$ are the eigenvalues, and $\widehat{\mu}_1, \ldots, \widehat{\mu}_{k-1}$ the corresponding eigenvectors of $\hat{S}$.

Provided that $\widehat{\lambda}_1 > \widehat{\lambda}_2, \ldots$, which will usually be the case in practice, the mean shape $\widehat{\mu}$ is well defined and is given by

$$\widehat{\mu} = \widehat{\mu}_1. \tag{1.8}$$

## 1.4 Coordinate Systems

In statistical shape analysis there several coordinate systems in common use. Each coordinate system is useful for some aspects of the analysis. Two coordinate systems will be considered here: full Procrustes coordinates and the tangent coordinates.

Procrustes analysis is a technique to match two objects up. When two or more objects are considered, they may have different rotations, translations and scales. So the technique of Procrustes analysis is used to match one object into the other. It is done using the pre-shapes of those objects since the pre-shapes have the same translation and scale.

For a given sample of pre-shapes, Procrustes analysis is performed by fitting the pre-shape of each object onto the mean shape. The new coordinates are called Procrustes fits or Procrustes coordinates and they will be defined below.

Let $z_1, \ldots, z_n$ be a random sample of pre-shapes, and also let $w_1, \ldots, w_n$ be a random sample of Helmertized configurations.

The configurations have an arbitrary rotation (see Dryden and Mardia, 1998, pp. 44-45). Thus, before proceeding with statistical shape analysis, it is necessary to rotate all the config-

urations in such way that they will be as close as possible of the sample mean shape. This is done by calculating

$$w_i^P = w_i^\star \widehat{\mu} w_i / (w_i^\star w_i),\ i = 1, \ldots, n. \tag{1.9}$$

Thus $w_1^P, \ldots, w_n^P$ are called the full Procrustes fits or full Procrustes coordinates.

Since the pre-shapes can be written as $z_i = w_i / ||w_i||$, where each $z_i$ is defined in $(1.4)$ and $||w_i|| = \sqrt{w_i^\star w_i}$ , the Procrustes coordinates can also be calculated from

$$w_i^P = z_i^\star \widehat{\mu} z_i,\ \ i = 1, \ldots, n.$$

Another useful system of coordinates is the tangent space coordinates. The concepts of tangent vectors and tangent space need to be presented before the definition of tangent coordinates (see Small, 1996, pp. 42-46). The tangent space of the shape space $\mathbb{C}P^{k-2}$ at the point z is the vector space of all the tangent vectors to $\mathbb{C}P^{k-2}$ at the point $z$. When performing tangent space inference, the tangent space at the sample mean pre-shape is often used.

The analysis of shape variability may be carried out in the tangent space. This space is a linearized version of the shape space. One of the main advantages of the tangent space is that standard multivariate techniques can be used directly.

There are several different types of tangent space coordinates. Here we use the partial Procrustes tangent coordinates, which are given by

$$t_i = e^{i\widehat{\theta}} [I_{k-1} - \widehat{\mu}\widehat{\mu}^\star] z_i,\ \ i = 1, \ldots, n, \tag{1.10}$$

where $z_i$ is a pre-shape vector defined in $(1.4)$ and $\widehat{\theta}$ minimizes $||\widehat{\mu} - z e^{i\widehat{\theta}}||^2$ and $||z|| = \sqrt{z^\star z}$.

Suppose that $z_1, \ldots, z_n$ is a random sample of pre-shapes and $t_1, \ldots, t_n$ their tangent coordinates, where each $z_i$ and $t_i$ are calculated using $(1.4)$ and $(1.10)$, respectively. Let $v_i$ be a $2k-2$ vector which is obtained by stacking the real and imaginary coordinates of each $t_i$. If $t_i = x_i + iy_i$, this operation is represented by $cvec$ where

$$v_i = cvec(t_i) = (x_i^T, y_i^T)^T, \tag{1.11}$$

where $x_i = \mathbf{Re}(t_i)$ is the real part of $t_i$ and $y_i = \mathbf{Im}(t_i)$ is the imaginary part of $t_i$. If the number of landmarks is $k$, a pre-shape vector $z_i$ has dimension $(k-1)$ and its corresponding vector of tangent coordinates $v_i$, where $v_i$ is given in $(1.11)$, has dimension $(2k-2)$.

Standard multivariate methods can be applied to the real tangent coordinates $v_i$. When the data are highly concentrated, methods based on the multivariate normal distribution can be applied for the real tangent coordinates $v_i$ (see Dryden and Mardia, 1998, p. 151). Some of these methods will be considered in the next sections.

## 1.5 Definition and Simulation of Shape Distributions

This section aims to review two distributions relevant to shape analysis: the complex normal distribution and the complex Bingham distribution. Methods for simulating these distributions are also discussed. The complex Bingham distribution is suitable for modelling pre-shapes and shapes and it will be used to evaluate the computer intensive methods of the next chapters.

### 1.5.1 Complex Normal Distribution

Since the multivariate complex normal and multivariate normal distribution are related, it is necessary to review the multivariate normal distribution.

The multivariate normal distribution is an extension of the univariate normal distribution to $(2k-2)$ variables (see Mardia et. al, 1979, p. 37), where the number of variables is chosen as $(2k-2)$ to make a connection with the shape context. The probability density function (pdf) of the multivariate normal of a $(2k-2)$ real vector $x$ is given by

$$f(x|\mu, V) = \frac{1}{(2\pi)^{(k-1)}}|V|^{-1/2}\exp\{-\frac{1}{2}(x-\mu)^T V^{-1}(x-\mu)\}, \tag{1.12}$$

where $V$ is $(2k-2)\times(2k-2)$ positive definite matrix, $|V| =$ det $V$, and $\mu$ is a $(2k-2)$ real vector.

A multivariate complex normal distribution can be represented as a real multivariate normal distribution (see Dryden and Mardia, 1998, p. 112). To clarify this relationship, consider the $(k-1)$ complex vector $z = (z_1, \ldots, z_{k-1})^T$ and the $(2k-2)$ real vector

$$v = (x^T, y^T)^T = (x_1, \ldots, x_{k-1}, y_1, \ldots, y_{k-1})^T, \tag{1.13}$$

where $x_j = \mathbf{Re}\{z_j\}$ is the real part of $z_j$ and $y_j = \mathbf{Im}\{z_j\}$ is the imaginary part of $z_j$. Suppose that

$$v \sim N_{2k-2}\left((\mu_1^T, \mu_2^T)^T, \frac{1}{2}\begin{bmatrix} \Sigma_1 & -\Sigma_2 \\ \Sigma_2 & \Sigma_1 \end{bmatrix}\right), \tag{1.14}$$

where $N_{2k-2}(\mu, \Sigma)$ denotes a $2k-2$ multivariate normal distribution with mean vector $\mu$ and covariance matrix $\Sigma$, $\Sigma_2 = -\Sigma_2^T$ is skew-symmetric and $\Sigma_1$ is symmetric positive definite.

The distribution of the complex vector $z$ is known as the complex normal distribution, which is denoted by $CN_{k-1}(\mu, \Sigma)$, where $\nu = \mu_1 + i\mu_2$ and $\Sigma = \Sigma_1 + i\Sigma_2$, (see Dryden and Mardia, 1998, p. 112). The pdf of $z$ is given by

$$f(z) = \frac{1}{\pi^{k-1}|\Sigma|} e^{-(z-\mu)^\star \Sigma^{-1}(z-\mu)}. \tag{1.15}$$

In the real case, it is well-known that the quadratic form $(x-\mu)^T V^{-1}(x-\mu)$ in $(1.12)$ has a $\chi^2_{2k-2}$ distribution. However, in the complex case, it is $2(z-\nu)^\star \Sigma^{-1}(z-\nu)$ which has a $\chi^2_{2k-2}$ distribution. The need for this factor 2 is explained in appendix C.

### 1.5.2 Simulation Method for the Complex Normal Distribution

Consider the problem of generating a vector $z$ which has a complex normal distribution with complex mean $\mu$ and Hemitian covariance matrix $\Sigma$.

The complex Gaussian vector $z$ will be represented as a real multivariate Gaussian vector $v$; see $(1.14)$. Then $v$ is simulated using a standard method (See Bratley et al, 1983, p. 152), and $z$ is obtained from $v$ by the inverse operation to $cvec$ in $(1.11)$.

The procedure to generate a $2(k-1)$ real Gaussian vector $v$ in $(1.14)$ is defined as follows. Let $A$ be a $(2k-2) \times (2k-2)$ upper triangular matrix such that

$$A^T A = \frac{1}{2} \begin{bmatrix} \Sigma_1 & -\Sigma_2 \\ \Sigma_2 & \Sigma_1 \end{bmatrix},$$

where $\Sigma_1$ and $\Sigma_2$ are $(k-1) \times (k-1)$ real matrices.

Let $u \sim N_{2k-2}(0_{2k-2}, I_{2k-2})$, where $0_{2k-2}$ is a $2k-2$ vector of $0$. Then the vector $v$ is given by

$$v = (\mu_1^T, \mu_2^T)^T + A^T u,$$

where $\mu_1$ and $\mu_2$ are real $(k-1)$ vectors, and $z$ is obtained by applying the inverse operation to $(1.11)$ to $v$.

So the $k-1$-dimensional vector $z$ has complex normal distribution with mean vector $\mu = \mu_1 + i\mu_2$ and covariance matrix $\Sigma = \Sigma_1 + i\Sigma_2$.

### 1.5.3 Complex Bingham Distribution

One of the most useful distributions for two dimensional landmark datasets is the complex Bingham distribution. A detailed account of this distribution is given by Kent (1994). This is a distribution on the space of complex unit vectors, or equivalently, the complex unit sphere.

If $z$ is a random complex unit vector with complex Bingham distribution, the pdf of $z$ is given by

$$f(z) = c(A)^{-1} \exp(z^\star A z), \;\; z \in \mathbb{C}S^{k-1}, \tag{1.16}$$

where $A$ is a $(k-1) \times (k-1)$ Hermitian matrix and $c(A)$ is a normalizing constant. If $A = I$, $f(z)$ becomes a uniform distribution on $\mathbb{C}S^{k-1}$, due to the constraint $z^\star z = 1$.

The complex Bingham distribution has the property of complex symmetry, which means that $z$ and $e^{(i\theta)}z$, where $\theta \in [0, 2\pi)$, have the same distribution (see Kent, 1994, p. 290). This is an important reason for using this distribution as a plausible model for the analysis of landmark

data in two dimensions, since a shape distribution should respect the definition of shape given in (1.6).

### 1.5.4 Simulation Method for the Complex Bingham Distribution

To simulate from the complex Bingham distribution, which is defined in (1.16), one of the methods proposed by Er (1998) is reviewed. Initially, $(k-2)$ truncated exponentials are generated subject to a linear constraint, and then these random variables are expressed in polar coordinates to deliver a complex Bingham distribution.

Let $TE(\lambda)$ denote the $\exp(\lambda)$ distribution conditioned to lie in $[0,1]$. A simple algorithm for simulating the $TE(\lambda)$ distribution is as follows.

It should be noted that $\lambda$ here is the rate.

**Algorithm 1.1.** *Simulation of* $TE(\lambda)$

1 - Simulate a uniform random variable $u \in [0,1]$.

2 - Calculate $X = -(1/\lambda)\log(1-u(1-\exp^{-\lambda}))$.

The method for simulating the complex Bingham distribution uses $(k-2)$ truncated exponentials to generate a $(k-1)$ vector with a complex Bingham distribution. Suppose the eigenvalues of $A$ are $\tilde{\lambda}_1 \leq \ldots \leq \tilde{\lambda}_{k-2} < \tilde{\lambda}_{k-1}$, and write $\lambda_j = \tilde{\lambda}_{k-1} - \tilde{\lambda}_j,\ j = 1,\ldots,k-2$. The input is a $(k-2)$-vector

$$\tilde{\lambda} = (\lambda_1, \ldots, \lambda_{k-2}). \tag{1.17}$$

**Algorithm 1.2.** *Simulation of Complex Bingham Distribution; Er (1998)*

1 - Generate $S = (S_1, S_2, \ldots, S_{k-2})^T$ where $S_j \sim TE(\lambda_j)$ are independent random variables simulated using Algorithm 1.1.

2 - If $\sum_{j=1}^{k-2} S_j < 1$, write $S_{k-1} = 1 - \sum_{j=1}^{k-2} S_j$. Otherwise, return to step 1.

3 - Generate independent angles $\theta_j \sim U[0, 2\pi),\ j = 1, \ldots, k-1$.

4 - Calculate $z_j = S_j^{1/2} \exp(i\theta_j)$, $j = 1, \ldots, (k-1)$.

The algorithm delivers a $(k-1)$ vector $z = (z_1, \ldots, z_{k-1})^T$, which has a complex Bingham distribution. Note that $(S_j^{1/2}, \theta_j)$ are essentially polar coordinates for complex number $z_j$.

If the parameter matrix $A$ has spectral decomposition $A = \Gamma\Lambda\Gamma^\star$ (see appendix A), with $\Gamma \neq I_{k-1}$, then $\Gamma z$ rather than $z$ should be returned.

## 1.6 Confidence Regions based on Normal Approximation

The tangent coordinates can be used for building confidence regions based on a normal approximation. First, it is necessary to study the variability on the tangent space. This variability can be studied using the method of principal components. The principal component method can also be used for building approximate normal-based confidence regions on the landmark space. These issues will be considered in this section.

Consider a random sample of complex configurations $z_1^0, \ldots, z_n^0$, where $z_i^0$ was defined in $(1.2)$. Suppose that $v_1, \ldots, v_n$ are the tangent coordinates of those complex configurations, where $v_i$ is defined in $(1.11)$. The variability in the tangent space is measured by the sample covariance matrix of the tangent coordinates $v_i$, given by the $(2k-2) \times (2k-2)$ matrix

$$S_v = \frac{1}{n}\sum_{i=1}^{n}(v_i - \bar{v})(v_i - \bar{v})^T, \tag{1.18}$$

where $\bar{v} = \sum_{i=1}^{n} v_i/n$.

The method of principal components can be used to summarize the variability of a random vector (see Mardia, et. al, 1979, p. 213). The idea of the principal component method is to reduce the dimension of the sample by focusing on the most important directions of variability. In the shape analysis context, the idea is to apply the principal component method to the sample covariance matrix of the tangent coordinates, to obtain the first few principal components and to project those components back to the landmark space (see Dryden and Mardia, 1998, pp. 47-51).

The matrix $S_v$ can be written in terms of the spectral representation

$$S_v = \sum_{i=1}^{p} \phi_i u_i u_i^T, \tag{1.19}$$

where $p = min(2k-4, n-1)$ is the total number of principal components , $\phi_1 \geq \ldots \geq \phi_p$ are eigenvalues and $u_1, \ldots, u_p$ the eigenvectors of $S_v$ (see Mardia et al, 1979, pp. 469).

The shape variability on the tangent space is studied using the principal components via the equations

$$v = \bar{v} + c\sqrt{\phi_j}u_j, \;\; j = 1, \ldots, p, \tag{1.20}$$

where $c$ is a constant, $\bar{v}$ is defined below $(1.18)$ and $\phi_j$ and $u_j$ were defined below (1.19).

Insight can be gained by giving different values to the constant $c$. Under the assumption that the tangent coordinates follow a multivariate normal distribution, it can be shown that $c$ is

approximately $N(0,1)$ (see Dryden and Mardia, 1998, p. 49). On this basis, a plausible range of values for c is $[-3,3]$.

The principal components method can also be used for building confidence regions based on a normal approximation (NA). The idea of using principal components for building a confidence region for the mean shape is particularly appealing when the observations on the landmark space for each landmark follow a bivariate normal distribution. The assumption of normality is more plausible for highly concentrated data.

The confidence regions obtained by normal approximation, referred to below as NA confidence regions, are calculated using the principal components for tangent coordinates. The NA method uses those principal component, conveniently relocated by replacing $\bar{v}$ by the mean shape $\widehat{\mu}$ (see $(1.8)$) in (1.20), to obtain the coordinates of the objects in the landmark space. Only the first and the second principal components are used since with those components it is possible to construct an ellipse for each landmark and represent it in a $2D$ plot. The axes of this ellipse are determined by the eigenvectors, and the relative scale along each axes is determined by the eigenvalues, corresponding to the two leading principal components. Thus NA confidence regions can be represented graphically by a plot of

$$\widehat{\mu} + c\sqrt{\phi_1}u_1 \text{ and } \widehat{\mu} + c\sqrt{\phi_2}u_2 \tag{1.21}$$

where usually $c \in (-3,3)$, $\widehat{\mu}$ is given in $(1.8)$ and $\phi_j$ and $u_j$ were defined below (1.19). See Dryden and Mardia (1998, p. 50).

## 1.7 Tests for One Group of Objects

We consider two methods in current use for testing if the mean shape is equal to a particular value. One is the one-sample Hotelling's $T^2$ test and the other is the one-sample Goodall test. The first one is less restrictive than the second but more complex. The Goodall test assumes the joint distribution on the landmark space is complex normal and isotropic (see Dryden and Mardia, 1998, p. 160), which means that the variance for each landmark is the same. On the other hand, the Hotelling's $T^2$ test assumes normality for the observations on the tangent space and isotropy is not assumed.

### 1.7.1 Hotelling's $T^2$ Test for a Specified Mean Shape

Consider the assumptions of the one sample Hotelling's $T^2$ test. Let $z_1^0, \ldots, z_n^0$ be a random sample of complex configurations, $z_1, \ldots, z_n$ be the pre-shapes of those configurations, where $z_i$ is calculated from $(1.4)$, and let $\widehat{\mu}$ be the mean shape of this sample, calculated using $(1.8)$. Let $v_1, \ldots, v_n$ be the partial Procrustes tangent coordinates of those pre-shapes, where $v_i$ is obtained from $(1.11)$. Recalling the tangent sample mean $\bar{v}$ and tangent sample covariance matrix $S_v$ from $(1.18)$, suppose that the $v_i$ have a multivariate normal distribution.

The aim of the Hotelling's $T^2$ test is to evaluate the hypotheses

$$H_0 : [\mu] = [\mu_0] \; versus \; H_1 : [\mu] \text{ unrestricted},$$

where $[\mu_0]$ is a pre-specified value for the mean shape. Here $[\mu]$ can be thought of as an equivalent class of pre-shapes. The partial tangent coordinates $\gamma_0$ for the mean pre-shape $\mu_0$

are given by

$$\gamma_0 = (I_{2k-2} - cvec(\hat{\mu})cvec(\hat{\mu}^T))cvec(\mu_0^P/||\mu_0^P||), \tag{1.22}$$

where $cvec(.)$ was defined in $(1.11)$, and $\mu_0^P$ is the procrustes fit of $\mu_0$, which is calculated using $(1.9)$. The statistic used for this test is given by

$$F = \frac{(n-M)}{M}(\bar{v} - \gamma_0)^T S_v^+ (\bar{v} - \gamma_0), \tag{1.23}$$

where $\gamma_0$ is given in $(1.22)$, $S_v^+$ is the Moore-Penrose generalized inverse (see appendix (A)) of $S_v$, and $M$ is the dimension of the tangent space and calculated as $2k - 4$.

This statistic has an $F_{M,n-M}$ distribution under $H_0$. The hypothesis $H_0$ is rejected at the level $\alpha$ if $F \geq F(M, n-M, \alpha)$, where $F(M, n-M, \alpha)$ is the quantile of the F distribuition with numerator $M$ and denominator $n - M$ for the $\alpha$ significance level.

### 1.7.2 Goodall's Test for a Specified Mean Shape

The situation is similar to Hotelling's test but isotropy is assumed. Let $z_1, \ldots, z_n$ a random sample of pre-shapes, where each $z_i$ is given by $(1.4)$. Also consider the tangent coordinates $v_1, \ldots, v_n$ of those pre-shapes, where $v_i$ is defined in $(1.11)$.

Goodall's test has the assumption that the tangent coordinates follow an isotropic normal model. So the $v_i$ have a multivariate normal distribution with mean vector $\mu$ and covariance matrix $\Sigma = \sigma^2 I_{2k}$, where $\sigma^2$ is a constant and $I_{2k}$ is the $2k \times 2k$ identity matrix (see Goodall, 1991, p. 314 and Dryden and Mardia, 1998, p. 160).

As in Hotelling's $T^2$ test, the hypotheses under consideration are

$$H_0 : [\mu] = [\mu_0] \;\; versus \;\; H_1 : [\mu] \neq [\mu_0].$$

The Goodall test is based on the squared Procrustes distances. For the pre-shapes $z_i$ and $z_j$, defined in $(1.4)$, this distance is given by

$$d_F^2(z_i, z_j) = 1 - z_i^\star z_j z_j^\star z_i, \tag{1.24}$$

for $i = 1, \ldots, n$ (see Dryden and Mardia, 1998, p. 41).

If $\widehat{\mu}$, the estimator of $\mu$, is close to $\mu$, and $\sigma$ is small, the approximate analysis of variance (ANOVA) is given by

$$\sum_{i=1}^{n} d_F^2(z_i, \mu) = \sum_{i=1}^{n} d_F^2(z_i, \widehat{\mu}) + n d_F^2(\mu, \widehat{\mu}),$$

(see Dryden and Mardia, 1998, p. 160).

Under the null hypothesis $H_0$, the distribution of the squared Procrustes distances are approximately chi-squared distributions, e. g.,

$$d_F^2(z_i, \mu_0) \sim \tau_0^2 \chi_M^2,$$

where $\tau_0 = \sigma / ||\mu_0||$ and $M = 2k - 4$. The proof of this result is derived using a Taylor series expansion (see Dryden and Mardia, 1998, p. 161).

Using this result and the additive property of independent chi-squared distributions,

$$\sum_{i=1}^{n} d_F^2(z_i, \mu_0) \sim \tau_0^2 \chi_{(n-1)M}^2.$$

Thus the test statistic (see Dryden and Mardia, 1998, pp. 160-161) is given by

$$G = (n-1)n\frac{d_F^2(\mu_0, \widehat{\mu})}{\sum_{i=1}^{n} d_F^2(x_i, \widehat{\mu})} \sim F_{M,(n-1)M}. \tag{1.25}$$

## 1.8 Tests for Several Populations

Two tests to compare the mean shape of two populations are considered in this section. The first one is the Goodall test and the second is the Hotelling's $T^2$. Those tests are extended versions of the tests of §1.7.

### 1.8.1 Hotelling's $T^2$ Test to Compare the Mean Shape of Two Populations

The test is used to compare the mean of two populations on the pre-shape space. However, the quantities being used are from the tangent space. This aspect will be clarified after the definitions of these quantities.

Consider an independent identically distributed (IID) random sample $z_{1j}^0, \ldots, z_{n_j j}^0$ of complex configurations from the population $\Pi^{[j]}$, where $j = 1, 2$. Let $z_{1j}, \ldots, z_{n_j j}$ and $v_{1j}, \ldots, v_{n_j j}$ be the pre-shapes and the tangent coordinates of $z_{1j}^0, \ldots, z_{n_j j}^0$, where $z_{lj}$ and $v_{lj}$ are calculated from $z_{lj}^0$ using $(1.4)$ and $(1.10)$.

The main assumptions of Hotelling's $T^2$ test are normality and homogeneity across populations of covariances matrices for the tangent coordinates. Suppose that the tangent coordinates $v_{1j}, \ldots, v_{n_j j}$ for population $j$ are IID, and approximately normally distributed with mean $\mu^{[j]}$ and common covariance matrix $V$.

The null and alternative hypothesis are given by

$$H_0 : [\mu^{[1]}] = [\mu^{[2]}] = [\mu] \; versus \; H_1 : [\mu^{[1]}], \; [\mu^{[2]}] \text{ unrestricted}, \tag{1.26}$$

where $[\mu]$ is the common mean shape.

Let $\widehat{\mu}^{[j]}$ and $\widehat{V}^{[j]}$ be the estimated mean and estimated covariance matrix of the tangent coordinates $v_{1j}, \ldots, v_{n_j j}$, where $\widehat{V}^{[j]}$ has divisor $n_j$. The Mahalanobis distance between $\widehat{\mu}^{[1]}$ and $\widehat{\mu}^{[2]}$ is given by

$$D = (\widehat{\mu}^{[1]} - \widehat{\mu}^{[2]})^T \widehat{V}^{+} (\widehat{\mu}^{[1]} - \widehat{\mu}^{[2]}),$$

where $\widehat{V} = (n_1 \widehat{V}^{[1]} + n_2 \widehat{V}^{[2]})/(n_1 + n_2 - 2)$, and $\widehat{V}^{+}$ is the Moore-Penrose generalized inverse of $\widehat{V}$, which is defined in (A.3) in appendix $A$.

The test statistic is

$$H = \frac{n_1 n_2 (n_1 + n_2 - M - 1)}{(n_1 + n_2)(n_1 + n_2 - 2)M} D \tag{1.27}$$

which, under $H_0$, has an $F_{M, n_1 + n_2 - M - 1}$ distribution, where $M = 2k - 4$ (see Dryden and Mardia, 1998, p. 154).

### 1.8.2 Goodall's Test to Compare the Mean Shape of Two Populations

Goodall's test assumes that the tangent coordinates have a jointly Gaussian distribution with an isotropic covariance matrix.

It should be noted that these assumptions are reasonable for data sets for which the variances of each landmark are small and similar. The hypotheses are

$$H_0 : [\mu^{[1]}] = [\mu^{[2]}] = [\mu] \; versus \; H_1 : [\mu^{[1]}], \; [\mu^{[2]}] \text{ unrestricted}, \tag{1.28}$$

where $[\mu]$ is the common mean.

To obtain the statistic of the test some results about the distribution of some Procrustes distances need to be used. These results are valid under $H_0$ and with $\sigma$ small. Therefore this test is appropriate for highly concentrated data. Set

$$\tau_0 = \sigma / ||\mu_0||,$$

where $||\mu_0|| = \sqrt{\mu_0^\star \mu_0}$.

The distribution of the Procrustes distances for each sample is given by

$$\sum_{i=1}^{n} d_F^2(z_{i1}, \widehat{\mu}^{[1]}) \sim \tau_0^2 \chi^2_{(n_1-1)M}, \tag{1.29}$$

where $d_F^2(.,.)$ is defined in $(1.24)$, and

$$\sum_{i=1}^{n} d_F^2(z_{i2}, \widehat{\mu}^{[2]}) \sim \tau_0^2 \chi^2_{(n_2-1)M}. \tag{1.30}$$

The Procrustes distance between the sample mean of the groups is given by

$$\sum_{i=1}^{n} d_F^2(\widehat{\mu}^{[1]}, \widehat{\mu}^{[2]}) \sim \tau_0^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right) \chi^2_M. \tag{1.31}$$

Thus, under $H_0$ and with $\sigma$ small, using (1.29), (1.30) and (1.31), the statistic

$$G_T = \frac{n_1 + n_2 - 2}{(n_1)^{-1} + (n_2)^{-1}} \frac{d_F^2(\widehat{\mu}^{[1]}, \widehat{\mu}^{[2]})}{\sum_{i=1}^{n} d_F^2(z_{i1}, \widehat{\mu}^{[1]}) + \sum_{i=1}^{n} d_F^2(z_{i2}, \widehat{\mu}^{[2]})} \tag{1.32}$$

has the approximate distribution $F_{M,(n_1+n_2-2)M}$ (see Dryden and Mardia, 1998, p. 162), where $M = 2k - 4$ as before.

.

## 1.9 Scope of the Thesis and Motivation

The contents of the following chapters are explained below. Some motivations for the thesis are given at the end of this section.

Chapter 2 explains how the bootstrap method of Fisher et al. (1996) for building confidence regions for directional data can be adapted to the shape context. It is proved that the distribution of the test statistic is asymptotically $\chi^2$ under the null hypothesis and is therefore asymptotically pivotal. The coverage accuracy of the bootstrap confidence region is compared numerically to Goodall and Hotelling confidence regions.

Chapter 3 introduces a bootstrap hypothesis test of a common mean shape across several populations. A proof that the statistic test is asymptotically pivotal under the null hypothesis is presented. This bootstrap test is compared to corresponding tests based on Goodall and Hotelling statistics using numerical simulation.

Chapter 4 presents both empirical likelihood confidence regions and hypothesis tests for shape data. First, it is explained how the empirical likelihood confidence regions of Fisher et. al. (1996) can be constructed in the shape context. Subsequently, an empirical likelihood hypothesis test of a common mean shape is introduced. Numerical simulations are carried out in order to compare these empirical likelihood methods to Goodall and Hotelling procedures.

Conclusions and some ideas for future work are presented in Chapter 5. Bootstrap and

empirical likelihood methods are compared, and numerical and methodological aspects are considered. How to apply the methods of this thesis in other areas of shape analysis is also discussed briefly.

The Goodall and Hotelling's $T^2$ tests work well under the assumption of high concentration, but they perform poorly when applied to data with low concentration. Even though the majority of shape datasets are highly concentrated, some datasets have low concentration. This provides motivation for using bootstrap and empirical likelihood methods in the shape analysis context, because they work well when applied to data having either high or low concentration. A second motivation is that bootstrap and empirical likelihood methods are nonparametric and only require weak assumptions about the underlying population.

# Chapter 2

# Bootstrap Confidence Regions for the Mean Shape

The aim of this chapter is to explain how the bootstrap confidence regions developed by Fisher et al. (1996) can be extended to the statistical shape analysis context. Fisher et al. (1996) proposed some bootstrap methods for building confidence regions for directional and axial data. Since there is a relationship between axial data and shape data for landmarks in two dimensions, it is possible to adapt bootstrap methods for axial data to shape data.

The sections are organized as follows. The main ideas and a literature review of the bootstrap are given in §2.1. Methodology for constructing bootstrap confidence regions is reviewed in §2.2. In §2.3 the bootstrap method of Fisher et al. (1996) for axial data is reviewed. The connection between axial and shape data is explained in §2.4. In §2.5 an asymptotically pivotal statistic for a sample of $n$ complex unit vectors is described. The bootstrap method for shape data, which is adapted from the bootstrap method for axial data, is explained in §2.6. In §2.7

the asymptotic distribution of the statistic we propose for shape data is derived. Some practical examples are considered in §2.8. In §2.9 some simulation experiments are performed in order to compare the bootstrap method for shape data with Goodall and Hotelling procedures.

## 2.1 Main Ideas and Literature Review of Bootstrap Methods

The main ideas about the bootstrap were introduced by Efron (1979). Efron (1979) presented the bootstrap as a more general method than the Quenouille-Tukey jackknife. According to Efron (1979), the jackknife can be considered as a linear expansion method for approximating the bootstrap.

Before explaining the bootstrap idea it is worth explaining what a functional is. A functional is a function of a function. Thus the notation

$$\nu(F) \text{ where } \nu : \{\text{space of distribution functions}\} \rightarrow \mathbb{R}^d \tag{2.1}$$

means that $\nu(F)$ is function of the distribution function $F$. For example, if $\nu(F)$ is the variance function and $F$ is the distribution function of the normal distribution $N(\mu, \sigma^2)$, then $\nu(F)$ is equal to $\sigma^2$.

To explain Efron's (1979) original idea, let $\mathrm{u} = \{u_1, \ldots, u_n\}$ be a random sample from a distribution with cumulative distribution function (CDF) $F$. Suppose that we are interested in an unknown parameter $\nu = \nu(F)$, and let $\widehat{F}_n(u) = n^{-1} \sum_{i=1}^{n} I(u_i \leq u)$, where $I(.)$ is the indicator function, denote the empirical distribution function based on the sample $\mathrm{u}$. The bootstrap estimator $\nu$ is given by $\widehat{\nu} = \nu(\widehat{F}_n)$. The bootstrap idea is to approximate the sampling

distribution of $\widehat{\nu}$ by drawing resamples randomly with replacement from the original sample $\mathrm{u}$, the main point being that the resampling distribution can be estimated to arbitrary accuracy using computer simulation.

Efron (1979) also considered the parametric bootstrap. The bootstrap mentioned above is nonparametric. But it is possible to define a parametric bootstrap by estimating $F$ by its parametric maximum likelihood estimator. For example, it is possible to assume that $F$ has a normal or any other particular distribution. The resamples with replacement will be not generated from the sample but from the parametric distribution $F$, with the parameters estimated from the sample.

Asymptotic properties of bootstrap methods can be examined using Edgeworth expansions. A seminal paper was Singh (1981). Singh (1981) showed theoretically that the bootstrap approximation for a distribution function of a sample mean is generally more accurate than the limiting normal distribution function approximation. For the case of quantiles he showed that the bootstrap approximation is as good as the normal approximation.

Bickel and Freedman (1981) showed some examples where the bootstrap approximation does not work so well. They conclude that for the majority of models with many parameters the bootstrap typically fails.

Hall (1992) presents very detailed information about bootstrap methods and Edgeworth expansions. Among other things, he explained the advantage of using an asymptotically pivotal statistic for bootstrapping (see Hall, 1992, pp. 83-91). A statistic is asymptotically pivotal if its limit distribution does not depend on unknown quantities (see Hall, 1992, p. 14). Considering an asymptotically normally distributed statistic $T$, Hall (1992) showed that bootstrapping $T$

reduces the error in the distribution function approximation from order $n^{-1/2}$ to order $n^{-1}$. However, if an asymptotically non-pivotal statistic is used the error does not reduce, its size is $n^{-1/2}$. Hall's (1992) discussion is very relevant for the bootstrap methods of this and the next chapter.

A number of authors have discussed bootstrap methods for confidence regions. A variety of methods for constructing nonparametric confidence intervals were introduced in Efron (1982). Some other important results can be found in Abramovitch and Singh (1985), Beran (1988), Hinkley (1988), Fisher and Hall (1990), and Hall and Wilson (1990), Hall (1988a), Hall (1988b) and Hall (1990).

Hall (1988a) compares five bootstrap confidence intervals. They come from both parametric and nonparametric contexts. Among the five methods, percentile-t and accelerated bias correction were identified as being superior. He also found that there is not a conclusive difference between the two methods: they achieve similar accuracy in both theoretical and numerical performance. Hall's (1988a) theoretical comparisons were made using Edgeworth expansions.

Asymptotic results clearly demonstrate the advantage of bootstrapping an asymptotically pivotal statistic for both hypotheses tests and confidence regions. Some papers supporting the use of pivotal statistics are Beran (1987), Liu and Sing (1987), Hall (1986), Hall (1988a) and Fisher et. al. (1996).

Bootstrap methods can be applied in many different areas of statistics, including generalized linear models, time series, sample surveys and statistical quality control, to name a few. These applications are covered in textbooks such as Efron and Tibshirani (1983), Davison and Hinkley (1997) and Chernick (1996).

In the directional data context, several papers have considered the use of the bootstrap for constructing confidence regions for the mean direction or mean axis of a population. See Ducharme et al. (1985), Fisher and Hall (1989) and Fisher et al. (1996). We now describe the developments in these papers in more detail.

Ducharme et al. (1985) developed a bootstrap method for directional data analysis for building confidence cones. They reviewed some parametric methods which are based on the assumption that the underlying distribution is a Fisher distribution. They presented a new bootstrap method which makes assumptions about the underlying distribution. In particular, the method of Ducharne et al. (1985) is not asymptotically pivotal except in relatively special circumstances, e. g. when the underlying population has rotational symmetry.

Fisher and Hall (1989) presented an asymptotic pivotal statistic for constructing confidence regions for directional data. However, this statistic leaves the sphere in its first step of calculation. Thus rescaling is needed to return to the surface of the unit sphere.

Fisher et al. (1996) introduced some asymptotically pivotal methods which involve projecting the true mean direction or mean axis onto the tagent space at the sample mean direction or axis. This approach has the advantages that it is simply to apply and (unlike the Fisher and Hall (1990) approach) no rescaling is required.

## 2.2 Bootstrap Confidence regions

Fisher et al. (1996)'s method for constructing confidence regions for an axis using axial data is based on the percentile-t method, one of the two methods identified by Hall (1988a) as being superior. The percentile-t method generalizes to higher dimensions more easily than the

accelerated bias correction method, the other superior method identified by Hall (1988a) in the scalar case.

The percentile-t method has some particular steps which need to be reviewed before considering the method of Fisher et al. (1996) for axial data. The case of a scalar parameter is considered initially.

Consider the problem of building a confidence interval for a unknown parameter $\upsilon$ of an unknown population based on the random sample $\mathrm{u} = \{u_1, \ldots, u_n\}$. Let $\widehat{\upsilon}$ be an estimator of $\upsilon$ and $\widehat{se}$ an estimator of its standard deviation which is denote by $se$.

The percentile-t method for building a confidence region for $\upsilon$ has the following steps (see, Efron and Tibshirani, 1993, pp. 160-161). First, consider $B$ resamples

$$\mathrm{u}^{(b)} = \{u_1^{(b)}, \ldots, u_n^{(b)}\}, \qquad b = 1, \ldots, B \tag{2.2}$$

each sampled randomly with replacement, from $\mathrm{u}$. For each $\mathrm{u}^{(b)}$ calculate

$$T_u^{(b)}(\widehat{\upsilon}) \equiv T_u^{(b)} = \frac{\widehat{\upsilon}^{(b)} - \widehat{\upsilon}}{\widehat{se}^{(b)}}, \tag{2.3}$$

where $\widehat{\upsilon}^{(b)}$ is the estimator $\widehat{\upsilon}$ calculated for the $b$-th bootstrap sample and $\widehat{se}^{(b)}$ is the estimated standard error of $\widehat{\upsilon}^{(b)}$.

The statistic $T_u^{(b)}$ is used to calculate a confidence interval for $\upsilon$ as follows. Set $T_u[1] < T_u[2] < \ldots < T_u[B-1] < T_u[B]$ to be the ordered values of $T_u^{(b)}$, $b = 1, \ldots, B$. Then a confidence interval for $\upsilon$ is given by

$$(\widehat{\upsilon} - t^{(1-\alpha)}\widehat{se}, \widehat{\upsilon} - t^{(\alpha)}\widehat{se}),$$

where $1 - 2\alpha$ is the confidence level of the interval, and $t^{(\alpha)}$ is the $\alpha$ percentile of $T_u[i]$. For example, if $\alpha = 0.05$ and $B = 100$, $t^{(0.05)} = T_u[5]$ and $t^{(0.95)} = T_u[95]$.

The percentile-t method is named thus because the pivot $T_u$ corresponds to the studentized version of $\widehat{\upsilon}$; see (Hall, 1992, p. 15).

This method can also be used for vector parameters. In this case $\upsilon$ is an unknown parameter vector, and $\widehat{\upsilon}$ is the estimator of $\upsilon$ and $\widehat{V}$ is the estimator of the covariance matrix of $\widehat{\nu}$ based on a random sample $u = \{u_1, \ldots, u_n\}$. The procedure above is used with the multivariate analogue of the square of $(2.3)$, which is given by

$$T_u^{2(b)}(\widehat{\nu}) = (\widehat{\upsilon}^{(b)} - \widehat{\upsilon})^T (\widehat{V}^{(b)})^{-1} (\widehat{\upsilon}^{(b)} - \widehat{\upsilon}). \tag{2.4}$$

The confidence region for the mean vector $\upsilon$ is built in a similar way to the confidence interval. Given $B$ resamples, $\mathrm{u}^{(1)}, \ldots, \mathrm{u}^{(b)}$, selected randomly with replacement, from $u$, calculate $T_u^{2(b)}$ for $b = 1, \ldots, B$. Let $T_B[1] \leq T_B[2], \ldots, T_B[B-1] \leq T_B[B]$ be the ordered values of $T_u^{2(b)}$, where $b = 1, \ldots, B$. then the confidence region is given by

$$R_\alpha = \{\upsilon : T_u^2(\upsilon) \leq t_\alpha^{(B)}\},$$

where $T_B[B(1-\alpha)]$ and $1 - \alpha$ is the nominal coverage level.

For some particular types of statistical analysis such as directional data analysis and statistical shape analysis, it is more difficult to find a pivotal statistic. The difficulties appear because these kinds of data are non-Euclidean.

## 2.3 The Method of Fisher et al. (1996) for Axial Data

Fisher et al. (1996) presented bootstrap confidence regions for directional and axial data which are based on asymptotically pivotal statistics. The method for axial data is explained in this section; in §2.4 we explain the relationship between axial data and shape data.

Some notation for axial data is now introduced. Let $x$ be a random vector on the unit sphere $S^d = \{x \in \mathbb{R}^d : ||x|| = 1\}$, where $\mathbb{R}^d$ is $d$ dimensional real space.

For axial data, $x$ and $-x$ are identified as equivalent. A relevant population characteristic is the mean polar axis, which is the unit vector $m$ that is defined to be the eigenvector associated to the largest eigenvalue of $S = E(XX^T)$. Thus for a sample of axes

$$\mathrm{x} = \{x_1, \ldots, x_n\}, \tag{2.5}$$

the parameter $S$ is estimated by $\hat{S} = n^{-1} \sum x_i x_i^T$. If $\hat{S}$ is written in spectral form (see appendix A)

$$\hat{S} = \sum_{j=1}^{d} \hat{\eta_j} \hat{m}_j \hat{m}_j^T, \tag{2.6}$$

where $\hat{\eta_1} \geqslant \hat{\eta_2} \geqslant \ldots \geqslant \hat{\eta_d}$ are the eigenvalues, and $\hat{m}_1, \ldots, \hat{m}_d$ the corresponding eigenvectors, the mean polar axis is given by

$$\hat{m} = \hat{m_1}. \tag{2.7}$$

Fisher et al. (1996) indicate how to construct a pivotal percentile-t method for axial data. In a non-Euclidean space addition and subtraction of vectors is not well-defined, so it is not clear

at the outset how to studentize directional or axial data. Fisher et al. (1996) used the statistic

$$T(m) = nm^T \widehat{M}_d^T \widehat{\Sigma}^{-1} \widehat{M}_d m, \tag{2.8}$$

where the elements of the $(d-1) \times (d-1)$ matrix $\widehat{\Sigma}$ are given by

$$\hat{\Sigma}_{jk} = n^{-1}(\hat{\eta}_1 - \hat{\eta}_j)^{-1}(\hat{\eta}_1 - \hat{\eta}_k)^{-1} \times \sum_{i=1}^{n} (\hat{m}_j^T x_i)(\hat{m}_k^T x_i)(\hat{m}^T x_i)^2, \tag{2.9}$$

where $\hat{\eta}_1 \geq \hat{\eta}_2 \geq \ldots \geq \hat{\eta}_d$ are the eigenvalues, and $\hat{m}, \hat{m}_2, \ldots, \hat{m}_d$ the corresponding eigenvectors of $\hat{S}$ in $(2.6)$, and the $(d-1) \times d$ matrix $\widehat{M}_d$ is given by

$$\widehat{M}_d = [\hat{m}_2, \ldots, \hat{m}_d]^T. \tag{2.10}$$

Fisher et al. (1996) use the idea of pivoting on the tangent space to the sphere $S^d$ at the sample mean axis $\widehat{m}$. The tangent plane for this case can be represented by the hyperplane $\mathcal{T}_{\widehat{m}} = \{t \in \mathbb{R}^d : t^T \widehat{m} = 0\}$, which is the space of all vectors orthogonal to $\widehat{m}$. Thus the rows of the matrix $M$ lie in the tangent space at $\widehat{m}$, and $\widehat{M}_d m = 0_{d-1}$. The product $\widehat{M}_d m = \widehat{M}_d(m - \widehat{m})$ projects $m$ onto the tangent plane at $\widehat{m}$. The matrix $\widehat{\Sigma}$ is the asymptotic covariance matrix of $\widehat{M}_d m$. Thus $T$ in (2.8) can be considered an asymptotically pivotal statistic for axial data, which is an analogue of $(2.4)$ for multivariate data. Further details about how to bootstrap these statistics will be given in §2.6.

Using the statistic (2.8), Fisher et al. (1996) present the following bootstrap algorithm, referred to as *Algorithm 2.1,* for building a confidence region for the mean axis given in (2.7).

**Algorithm 2.1.** *Bootstrap Method for Building Confidence Regions for the Mean Axis*

*Step 1* - For a sample x of axial data, defined in (2.5), calculate the matrix $\widehat{\Sigma}$ and the matrix $\widehat{M_d}$, which were defined in $(2.9)$ and $(2.10)$, respectively.

*Step 2* - Generate $B$ resamples $\mathrm{x}^{(b)} = \{x_1^{(b)}, \ldots, x_n^{(b)}\}$, $b = 1, \ldots, B$, randomly with replacement, from the original sample x.

*Step 3* - For each resample, calculate the quantities $\widehat{\Sigma}$, $\widehat{M_d}$ and $T$ using $\mathrm{x}^{(b)}$. Those quantities will be denoted $\widehat{\Sigma}^{(b)}$ , $\widehat{M}_d^{(b)}$ and $\widehat{T}^{(b)}$, respectively. The statistic $\widehat{T}^{(b)}$ is given by

$$\widehat{T}^{(b)} = T^{(b)}(\widehat{m}) = n\widehat{m}^T(\widehat{M}_d^{(b)})^T(\widehat{\Sigma}^{(b)})^{-1}\widehat{M}_d^{(b)}\widehat{m}.$$

*Step 4* - After the step (3), the values $\{\widehat{T}^{(b)}; b = 1, \ldots, B\}$ are sorted, into order, giving

$$\widehat{T}^{(b)}[1] \leqslant \widehat{T}^{(b)}[2] \leqslant \ldots \leqslant \widehat{T}^{(b)}[B-1] \leqslant \widehat{T}^{(b)}[B],$$

and let $t_\alpha^{(B)}$ be the chosen value corresponding to the level $\alpha$. For instance, if $B = 100$ and $\alpha = .1$, the chosen value is $t_{0.1}^{(100)} = T_{100}[90]$.

*Step 5* - The confidence region based on $(2.8)$ with coverage probability $1 - \alpha$ is given by

$$R_\alpha = \{m : T(m) \leqslant t_\alpha^{(B)}\}. \tag{2.11}$$

The method of Fisher et al. (1996) has some advantages when compared with the methods of Fisher and Hall (1989) and Ducharme et al. (1985). The Fisher and Hall (1989) method is asymptotically pivotal, but involves some awkward scaling while typically the method of Ducharme et. al. (1985) is not asymptotically pivotal. In contrast, the statistic $T$ is asymptotically pivotal and this is achieved without leaving the surface of the sphere.

Fisher et al. (1996) showed that the coverage error, which is defined by

$$coverage\ error\ =\ coverage\ probability\ -\ nominal\ coverage\ probability,$$

of the confidence region given by $(2.11)$ is of size $O(n^{-2})$; the order notation $O(.)$ is reviewed in appendix B. Equivalently, if $m_0$ is the true mean shape, then under mild conditions, $Pr[m_0 \in R_\alpha] = 1 - \alpha + O(n^{-2})$. The details of the proof about the theoretical coverage accuracy are given in appendix B of that paper. Edgeworth expansions for bootstrap quantities are used in this proof (see also Hall, 1992, Chap. 5). The proof of those results will not be explained here since they are beyond the objectives of this thesis.

## 2.4 Relationship Between Axial data and Shape Data

In this section the connections between axial and shape data will be explained (see Kent, 1992, pp. 118-9). Let $S^k = \{x \in \mathbb{R}^k : ||x|| = 1\}$ denote the real unit sphere in $\mathbb{R}^k$. Define $\aleph_k = \{uu^T : u \in S^k\}$. Note that the real unit vectors $u$ and $-u$ are mapped onto the same element of $\aleph_k$, and that $\aleph_k$ is the space of $k \times k$ symmetric, rank 1, projection matrices.

A $p \times p$ matrix $R$ is called orthogonal if $R^T R = I_p$. Let $O(p)$ be the space of $p \times p$ orthogonal matrices and define $SO(p) = \{R \in O(p) : |R| = 1\}$, the space of $p \times p$ rotation matrices.

Axial data can be understood in three distincts ways:

(a) an equivalence class of vectors in $\mathbb{R}^k$ in which a non-zero vector $x$ is identified with the axis $\{rx : r \neq 0\}$;

(b) an unsigned unit vector on the real sphere $\pm u \in S^k$;

(c) a projection matrix $uu^T \in \aleph_k$.

On the other hand, shape data with $k$ landmarks and $m$ dimensions can be represented either as

(a) an equivalence class of $k \times m$ matrices in which we identify $X \neq 0$

with $\{rXR : r > 0 \text{ and } R \in SO(m)\}$ or

(b) an equivalence class of standardized $k \times m$ matrices in which we identify $U$ with $\{UR : R \in SO(m)\}$. Note: we say that $U$ is standardized if $tr(U^T U) = 1$.

Shape data in two dimensions can be represented as:

(a) an equivalent class of complex $k-$vectors in which we identify $z$ with $\{rze^{i\theta} : r > 0, \theta \in [0, 2\pi)\}$;

(b) an equivalent class of rotated unit vectors in $CS^{k-2}$, $[u] = \{e^{i\theta}u : \theta \in [0, 2\pi)\} \subset CS^{k-2}$;

(c) a projection matrix $uu^\star$.

In the planar case, a pre-shape $z \in \mathbb{C}S^{k-1}$ can be written as $z = x + iy$ where $||z||^2 = ||x||^2 + ||y||^2 = 1$ but $z$ can be embedded in the real sphere $S^{2k}$ by stacking $x$ and $y$ forming a vector $(x^T, y^T)^T$.

## 2.5 Modified T-statistic for Complex Unit Vectors

A modified version of the statistic (2.8) can be used for complex vectors. The quantities (2.8), (2.9) and (2.10) need to be redefined. Let $z_1, \ldots, z_n$ be a random sample of pre-shapes, where each $z_i$ was defined in $(1.4)$, and let $\widehat{\mu}_1, \ldots, \widehat{\mu}_{k-1}$ denote the unit eigenvectors associated with the eigenvalues $\widehat{\lambda}_1$, $\widehat{\lambda}_2$,..., $\widehat{\lambda}_{k-1}$ of the product matrix which were defined below (1.7).

For complex unit vectors, which correspond to pre-shapes, the modified form of the statistic of Fisher et al. (1996) is defined by

$$T(\mu) = 2n\mu^\star \widehat{M}_{k-2}^\star \widehat{\Sigma}^{-1} \widehat{M}_{k-2}\mu, \tag{2.12}$$

where, for a sample of pre-shapes z, it is necessary to calculate a $(k-2) \times (k-2)$ matrix $\widehat{\Sigma} = (\widehat{\Sigma}_{jl})$ and a $(k-2) \times (k-1)$ matrix $\widehat{M}_{k-2}$, which are defined as follows:

$$\widehat{\Sigma}_{jl} = n^{-1}(\widehat{\lambda}_1 - \widehat{\lambda}_j)^{-1}(\widehat{\lambda}_1 - \widehat{\lambda}_l)^{-1} \times \sum_{i=1}^{n} (\widehat{\mu}_j^\star z_i)(z_i^\star \widehat{\mu}_l)(z_i^\star \widehat{\mu})(\widehat{\mu}^\star z_i), \tag{2.13}$$

where $\widehat{\mu} = \widehat{\mu}_1$, and

$$\widehat{M}_{k-2} = [\widehat{\mu}_2, \ldots, \widehat{\mu}_{k-1}]^\star. \tag{2.14}$$

**Comments**

1. Under the null hypothesis $H_0 : [\mu] = [\mu_0]$, where $[\mu_0]$ is the true population mean shape, the asymptotic distribution of the statistic $T(\mu_0)$ in $(2.12)$ is $\chi^2_{2k-4}$ under mild conditions. A proof is given in §2.7.

2. The statistic $T(\mu)$ in $(2.12)$ is invariant with respect to the choice of pre-shape $\mu$ from the shape equivalence class $[\mu]$.

3. The need for the factor 2 in $(2.12)$ follows from appendix C.

4. The tabular version of the test of the null hypothesis $H_0 : [\mu] = [\mu_0]$ based on the statistic $(2.12)$ is performed as follows. The null hypothesis $H_0$ is rejected, at the level $\alpha$, if

$T(\mu_0)$ is larger than $\chi^2_{2k-4}(\alpha)$, which is the quantile of a $\chi^2$ distribution with $(2k-4)$ degrees of freedom associated to the level $\alpha$. In the next section we explain how $T(\mu_0)$ can be used to construct bootstrap confidence regions.

## 2.6 Bootstrap Confidence Regions for the Mean Shape

There some points to note about how to adapt the method of Fisher et al. (1996) for axial data to shape data. Pre-shapes are complex unit vectors while axial data consists of real unit vectors. In both cases there is information which is discarded. In the axial case, this corresponds to the sign of the unit vector; and in the shape case, this corresponds to the rotation information.

*A***lgorithm 2.2.** *Pivotal bootstrap Confidence Regions for the Mean Shape*

The bootstrap method for building a confidence region for the mean shape $\mu$ can be described as follows:

*Step 1* - For a sample of pre-shapes z, defined previously, calculate the matrix $\widehat{\Sigma}$ and the matrix $\widehat{M}_{k-2}$ , which were defined in $(2.13)$ and $(2.14)$, respectively.

*Step 2* - Generate $B$ resamples $\mathrm{z}^{(b)} = \{z_1^{(b)}, \ldots, z_n^{(b)}\}$, randomly with replacement, from the original sample $\mathrm{z} = \{z_1, \ldots, z_n\}$.

*Step 3* - For each resample, calculate the quantities $\widehat{\Sigma}$, $\widehat{M}_{k-2}$ and $T$, where $T$ was defined in $(2.12)$, using $\mathrm{z}^{(b)}$. Those quantities will be denoted $\widehat{\Sigma}^{(b)}, \widehat{M}_{k-2}^{(b)}$ and $\widehat{T}^{(b)}$, respectively. So, for each bootstrap sample, calculate

$$\widehat{T}^{(b)} = T^{(b)}(\widehat{\mu}) = 2n\widehat{\mu}^{\star}(\widehat{M}_{k-2}^{(b)})^{\star}(\widehat{\Sigma}^{(b)})^{-1}\widehat{M}_{k-2}^{(b)}\widehat{\mu}, \tag{2.15}$$

where $\widehat{\mu}$ is the mean shape of the original sample.

*Step 4* - After concluding Step (3), the values of $\widehat{T}^{(b)}$ are sorted into order

$$\widehat{T}^{(b)}[1] \leqslant \widehat{T}^{(b)}[2] \leqslant \ldots \leqslant \widehat{T}^{(b)}[B-1] \leqslant \widehat{T}^{(b)}[B].$$

Let $t_{\alpha}^{(B)}$ be the chosen value corresponding to the level $\alpha$.

*Step 5* - Thus the region based on $(2.12)$ with nominal coverage probability $1-\alpha$ is given by

$$R_{\alpha} = \{\mu : T(\mu) \leqslant t_{\alpha}^{(B)}\}. \tag{2.16}$$

To represent graphically the bootstrap confidence regions obtained using *Algorithm 2.2*, the information about rotation in the bootstrap samples should be accounted for. In Step 3 of *Algorithm 2.2*, the rotation of the bootstrap mean shapes is arbitrary. If we wish to represent the bootstrap confidence regions graphically, as will be illustrated in §2.8, then the bootstrap mean pre-shapes should be rotated so that they are as close as possible to the sample mean pre-shape $\widehat{\mu}$. However, as noted above, the value of the statistic $T(\mu)$ in $(2.12)$ does not change if the rotation of $\mu$ changes. So the rotation information only needs to be removed when graphical representations are being considered.

### 2.6.1 Monte Carlo Simulation Design

If $n_M$ Monte Carlo samples are generated, and $B$ bootstrap samples are obtained for each Monte Carlo sample, this experiment will deliver $n_M$ confidence regions. Let $C_P$ denote the coverage probability of the region $(2.16)$, then $\widehat{C}_P$ is estimated by

$$\widehat{C}_P = \#\{i : T^{\langle i \rangle}(\mu_0) \leqslant t^{(B)}_{(\alpha,i)}, i = 1, \ldots, n_M\}/n_M,$$

where $\mu_0$ is the true mean, and $T^{\langle i \rangle}(\mu_0)$ is $T(\mu_0)$ in $(2.12)$ ,and $t^{(B)}_{(\alpha,i)}$ is $t^{(B)}_{\alpha}$, both based on the $i$th Monte Carlo sample.

### 2.6.2 Mahalanobis Bootstrap Method

An alternative bootstrap scheme is now described. The steps of the *Algorithm 2.2* are the same except that the statistic $(2.12)$ is replaced by the Mahalanobis statistic which is given in $(1.23)$. The main difference between this method and the previous one is that the Hotelling statistic $T^2$ is used in the bootstrap process. Recall from §1.7.1 that $v$ contains the partial Procrustes tangent coordinates and $S_v$ is the covariance matrix of the sample of tangent vectors.

*A***lgorithm 2.3.** *Bootstrap Confidence Region Using Hotelling $T^2$ Statistic*

*Step 1* - For a sample of pre-shapes z, calculate the statistic

$$F(\mu) = \frac{(n-M)}{M}(\bar{v} - \gamma_0(\mu))^T S_v^+ (\bar{v} - \gamma_0(\mu)), \tag{2.17}$$

where $\gamma_0(\mu) = (I_{km-m} - vec(\hat{\mu})vec(\hat{\mu}^T))vec(\mu^p/||\mu^p||)$, $vec(.)$ is defined in $(1.11)$, $S_v^-$ is the Moore-Penrose generalized inverse of $S_v$ and $M = 2k - 4$ is the dimension of the tangent space.

*Step 2* - Generate $B$ resamples $\mathrm{z}^{(b)} = \{z_1^{(b)}, \ldots, z_n^{(b)}\}$, randomly with replacement, from the original sample z.

*Step 3* - For each resample, calculate the statistic $F_B$.

*Step 4* - Sort the vector $F_B$ so that $F_B[1] \leqslant F_B[2], \ldots, F_B[B-1] \leqslant F_B[B]$, and let $f_\alpha^{(B)}$ be the chosen value corresponding to the level $\alpha$.

*Step 5* - Thus the region based on $(2.17)$ with nominal coverage probability $1-\alpha$ is given by

$$R_\alpha = \{\mu : F(\mu) \leqslant f_\alpha^{(B)}\}.$$

## 2.7 Asymptotic Distribution of the Statistic T

The purpose of this section is to prove that the statistic $T$, which was defined in $(2.12)$, has a null asymptotic $\chi^2_{(2k-4)}$ distribution, under mild conditions on the underlying population.

To prove this theorem two results, closely related to results which can be found in Watson (1983, pp. 216-217), will be assumed. It should be noted that the theorems presented by Watson (1983, pp. 216-217) are valid for real axial data. However, the type of data which are treated here are pre-shape data, i.e., they involve complex unit vectors. The results in Watson (1983) can also be derived for complex vectors.

To present the results and the theorem of this section some quantities need to be defined. Assume $z_1, \ldots, z_n$ are complex unit vectors from a population F and let $\widehat{S}_n = n^{-1}\sum_{i=1}^n z_i z_i^\star$ be a $(k-1)\times(k-1)$ Hermitian matrix and $S = E(\widehat{S}_n)$ its population analogue. The eigenvalues and eigenvectors and other terms associated with $\widehat{S}_n$ are denoted with a *hat*.

The first result is a central limit theorem for $\widehat{S}_n$ which states that

$$n^{1/2}(\widehat{S}_n - S) \xrightarrow{d} G, \tag{2.18}$$

where $G$ is a $(k-1) \times (k-1)$ Hermitian matrix. Its entries are jointly normally distributed with zero mean.

The second result, which can be derived from (2.18) using results analogous to those in Watson (1983, p. 216), states that

$$n^{1/2}(\widehat{P}_1 - P_1) \xrightarrow{d} \sum_{k>1} \frac{P_k G P_1 + P_1 G P_k}{\lambda_1 - \lambda_k}, \tag{2.19}$$

where $P_1 = \mu_1 \mu_1^\star = \mu\mu^\star$, $P_j = \mu_j \mu_j^\star$, $\widehat{P}_j = \widehat{\mu}_j \widehat{\mu}_j^\star$ and $\lambda_1 > \lambda_2 \ldots \geqslant \lambda_{k-1}$ are the eigenvalues of $S$. Note the assumption that the largest eigenvalue, $\lambda_1$, is simple.

Define

$$a(z) = (\mu^\star z)(z^\star \mu) \tag{2.20}$$

where $\mu = \mu_1$. Then $a(z) \in [0, 1]$ when $z$ and $\mu$ are both complex unit vectors. Also define

$$R = diag\{(\lambda_1 - \lambda_2)^{-1}, \ldots, (\lambda_1 - \lambda_{k-1})^{-1}\},$$

where the $\lambda_j$ were defined after $(2.19)$. In addition, let $M$ denote the population analogue of $\widehat{M}_{k-2}$ defined in $(2.14)$.

The following lemma will be used in the proof of the theorem.

***Lemma 2.1.*** *Let $X_1, X_2, \ldots$ be an IID sequence of random p-vectors such that $E[X_1] = \gamma$ is well defined, and $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$. Let $\widehat{A}_n(p \times 1)$ be a sequence of random vectors such that $||\widehat{A}_n - A|| \xrightarrow{p} 0$ as $n \to \infty$, where $||B|| = \sqrt{B^T B}$. Then $\widehat{A}_n^T \bar{X} \xrightarrow{p} A^T \gamma$ as $n \to \infty$.*

**Proof** We have

$$\hat{A}_n^T \bar{X} - A^T \gamma \equiv (\widehat{A}_n - A)^T \gamma + A^T(\bar{X} - \gamma) + (\widehat{A}_n - A)^T (\bar{X} - \gamma).$$

It is sufficient to show that all the terms on the right hand side go to zero. The first term on the right hand side goes to zero by assumption since

$$||(\hat{A}_n - A)^T \gamma|| \leq ||\hat{A}_n - A||\, ||\gamma|| \xrightarrow{p} 0.$$

The second term goes to zero by the weak law of large numbers, i.e.,

$$||A^T(\bar{X} - \gamma)|| \leq ||A||\, ||\bar{X} - \gamma|| \xrightarrow{p} 0.$$

The third term goes to zero because

$$||(\widehat{A}_n - A)^T (\bar{X} - \gamma)|| \leq ||\widehat{A}_n - A||\, ||\bar{X} - \gamma||,$$

where both terms on the right hand side go to zero in probability.

**Theorem 2.1.** *Suppose the underlying population $F$ is such that $(i)$ the largest eigenvalue of $S$ is distinct (so that the corresponding eigenvector is well-defined); $(ii)$ $a(z) = ||\mu^\star z||^2$ does not have a point mass at 0, i.e. $P[a(z) = 0] = 0$ and $(iii)$ the smallest eigenvalue of $S$ is positive. Then if $[\mu] = [\mu_1]$ is the true mean shape, $T(\mu)$, defined in (2.12), has an asymptotic $\chi^2_{2k-4}$ distribution.*

**Proof** The proof of the theorem is organized in 4 steps.

**Step 1** - Show $||n^{1/2}\widehat{M}_{k-2}\mu + n^{1/2} M_{k-2}(\widehat{\mu} - \mu)|| \xrightarrow{d} 0$ as $n \to \infty$.

We have the identity

$$\begin{aligned} -n^{1/2}\widehat{M}_{k-2}\mu &= n^{1/2}\widehat{M}_{k-2}(\widehat{\mu} - \mu) \\ &= n^{1/2} M_{k-2}(\widehat{\mu} - \mu) + n^{1/2}(\widehat{M}_{k-2} - M_{k-2})(\widehat{\mu} - \mu). \end{aligned}$$

Thus

$$||n^{1/2}\widehat{M}_{k-2}\mu + n^{1/2}M_{k-2}(\widehat{\mu}-\mu)|| = ||n^{1/2}(\widehat{M}_{k-2}-M_{k-2})(\widehat{\mu}-\mu)|| \xrightarrow{p} 0. \qquad (2.21)$$

To obtain $(2.21)$, note that $(i)$ $\widehat{S}_n \xrightarrow{p} S$ implies that $\widehat{M}_{k-2} \xrightarrow{p} M_{k-2}$, and $(ii)$ $||n^{1/2}(\widehat{\mu}-\mu)|| = O_p(1)$ as a consequence of $(2.18)$, where the last statement means the following: given $\epsilon > 0$, there exists a constant C, independent of $n$, such that

$$\limsup_{n\to\infty} P[n^{1/2}||\hat{\mu}-\mu|| > C] < \epsilon.$$

**Step 2** - Show that

$$n^{1/2}M_{k-2}(\widehat{\mu}-\mu) \xrightarrow{d} RM_{k-2}G\mu \sim \mathbb{C}N_{k-2}(0_{k-2},\Sigma),$$

where $\Sigma = cov(RM_{k-2}G\mu) = Rcov(M_{k-2}G\mu)R$. Pre-multiplying the left hand side of $(2.19)$ by $M_{k-2}$ and postmultiplying by $\mu$ we obtain

$$n^{1/2}M_{k-2}(\widehat{P}_1 - P_1)\mu \;\; = \;\; n^{1/2}M_{k-2}\widehat{\mu}\widehat{\mu}^{\star}\mu,$$

since $\widehat{P}_1 = \widehat{\mu}\widehat{\mu}^{\star}$ and $M_{k-2}P_1\mu = 0_{k-2}$ by definition of the quantities involved. Moreover,

$$n^{1/2}M_{k-2}\widehat{\mu}\widehat{\mu}^{\star}\mu \;\; = \;\; n^{1/2}M_{k-2}\widehat{\mu} + n^{1/2}M_{k-2}(\widehat{\mu}^{\star}\mu - 1)\widehat{\mu}.$$

Therefore

$$||n^{1/2}M_{k-2}\widehat{\mu}\widehat{\mu}^{\star}\mu - n^{1/2}M_{k-2}(\widehat{\mu}-\mu)|| = ||n^{1/2}M_{k-2}(\widehat{\mu}^{\star}\mu - 1)\widehat{\mu}|| \xrightarrow{p} 0 \qquad (2.22)$$

since $\widehat{\mu}^{\star}\mu \xrightarrow{p} 1$ and $M_{k-2}\mu = 0_{k-2}$.

Now consider the right hand side of $(2.19)$, pre-multiplying by $M_{k-2}$ and post-multiplying by $\mu$ we obtain

$$\begin{aligned} M_{k-2}\left[\sum_{j=2}^{k-1}\frac{P_jGP_1+P_1GP_j}{\lambda_1-\lambda_j}\right]\mu &= M_{k-2}\left[\sum_{j=2}^{k-1}\frac{1}{\lambda_1-\lambda_j}P_j\right]G\mu \\ &= RM_{k-2}G\mu \sim \mathbb{C}N_{k-2}(0,\Sigma), \end{aligned}$$

where $\Sigma$ is defined above. We have used the fact that, by definition, $M_{k-2}=\sum_{j=2}^{k-1}e_{j-1}\mu_j^\star$, where $e_j$ is the $(k-1)$-vector with the $jth$ element 1 and all other elements zero; and consequently

$$\begin{aligned} M_{k-2}\left[\sum_{j=1}^{k-1}\frac{1}{\lambda_1-\lambda_j}P_j\right] &= \sum_{j=2}^{k-1}\frac{1}{\lambda_1-\lambda_j}M_{k-2}P_j \\ &= \sum_{j=2}^{k-1}\frac{1}{\lambda_1-\lambda_j}e_{j-1}\mu_j^\star\mu_j\mu_j^\star \\ &= RM_{k-2}. \end{aligned}$$

Therefore the result follows from the fact that the left hand side of $(2.19)$, pre-multiplied by $M_{k-2}$ and post-multiplied by $\mu$ becomes $n^{1/2}M_{k-2}(\widehat{\mu}-\mu)$ and the corresponding right hand side is $RM_{k-2}G\mu$ which has distribution $CN_{k-2}(0,\Sigma)$.

**Step 3** - Show that $\widehat{\Sigma}=\frac{1}{n}\sum_{i=1}^{n}\widehat{R}(\widehat{M}_{k-2}z_i)(z_i^\star\widehat{M}_{k-2}^\star)(\widehat{\mu}^\star z_i)(z_i^\star\widehat{\mu})\widehat{R}\xrightarrow{p}\Sigma$. Lemma 2.1 can be applied to prove this.

The $jh$ element of matrix $\widehat{\Sigma}$ is given by

$$\widehat{\Sigma}_{jh}=\frac{1}{n}\sum_{i=1}^{n}\{\widehat{R}(\widehat{M}_{k-2}z_i)(z_i^\star\widehat{M}_{k-2}^\star)(\widehat{\mu}^\star z_i)(z_i^\star\widehat{\mu})\widehat{R}\}_{jh}$$

and can be written in the form

$$\widehat{A}_{jh}^{T} n^{-1} \sum_{i=1}^{n} X_i = \widehat{A}_{jh}^{T} \bar{X},$$

where $X_i$ is a vector whose components are of the form

$$z_{i\alpha} \bar{z}_{i\beta} z_{i\gamma} \bar{z}_{i\delta} : \alpha, \beta, \gamma, \delta \in [1, \ldots, k-1]$$

and $\hat{A}_{jh}$ is a vector whose components are polynomial functions of the components of $\hat{R}$, $\hat{M}_{k-2}$ and $\hat{\mu}$. By the law of large numbers, $\frac{1}{n} \sum X_i$ converges in probability to $E(X_1) = \gamma$, say, where the components of $\gamma$ are of the form $E(z_{1\alpha} \bar{z}_{1\beta} z_{1\gamma} \bar{z}_{1\delta})$. Moreover, since $\hat{R}$,$\hat{M}_{k-2}$ and $\hat{\mu}$ converge to their population analogues $R$, $M_{k-2}$ and $\mu$, say it follows that $\widehat{A}_{jh} \xrightarrow{p} A_{jh}$, where $A_{jh}$ is obtained from $\hat{A}_{jh}$ by replacing $\hat{R}$, $\hat{M}_{k-2}$ and $\hat{\mu}$ by their population values.

**Step 4 -** $\Sigma$ has full rank.

Define $a(z)$ as in $(2.20)$ and let $y = M_{k-2} z$, where $z \sim F$.

Then $a(z) \in [0, 1]$ since $\mu$ and $z$ are both complex unit vectors. Note that

$$E[yy^{\star}] = M_{k-2} S M_{k-2}^{\star} = diag[\lambda_2, \ldots, \lambda_{k-1}]$$

which is positive definite by assumption $(iii)$. Therefore the result will follow if we can show that, for some $\epsilon > 0$, $\Sigma \geq \epsilon E[yy^{\star}]$, where " $\geq$ " should be understood in terms of the partial ordering of non-negative definite matrices.

We have

$$\begin{aligned}\Sigma &= E[a(z)yy^{\star}] \\ &= E\left[a(z)I(a(z)>\delta)yy^{\star}\right] + E\left[a(z)I(a(z)\leq\delta)yy^{\star}\right]. \\ &\geq E\left[a(z)I(a(z)>\delta)yy^{\star}\right] \\ &\geq \delta E\left[I(a(z)>\delta)yy^{\star}\right].\end{aligned}$$

But it follows from assumption $(ii)$ of the theorem that

$$E[yy^{\star}] = lim_{\delta\to 0}E[I(a(z)>\delta)yy^{\star}],$$

and so there exists a $\tilde{\delta}>0$ such that

$$E\left[I(a(z)>\tilde{\delta})yy^{\star}\right] \geq \frac{1}{2}E\left[yy^{\star}\right],$$

from which it follows that

$$\Sigma \geq \frac{\tilde{\delta}}{2}E\left[yy^{\star}\right],$$

and therefore $\Sigma$ is positive definite.

Finally, by steps $(1)$ and $(2)$,

$$n^{1/2}\widehat{M}_{k-2}\mu \xrightarrow{d} \mathbb{C}N_{k-2}(0,\Sigma),$$

where $\mathbb{C}N$ is a complex normal, which was defined in (1.15). The steps $(3)$ and $(4)$ state that $\widehat{\Sigma} \xrightarrow{p} \Sigma$ and $\Sigma$ has full rank. Thus the inverse of $\widehat{\Sigma}$ exists in probability as $n\to\infty$ and the statistic

$$T(\mu) = 2n\mu^{\star}\widehat{M}_{k-2}^{\star}\widehat{\Sigma}^{-1}\widehat{M}_{k-2}\mu,$$

is well defined in the limit. Thus

$$T(\mu) \xrightarrow{d} \chi^2_{2k-4} \tag{2.23}$$

as required.

**Comments**

1. Condition (ii) of Theorem 2.1 is satisfied if the underlying population is continuous.

2. Condition (iii) is satisfied provided the population distribution is not concentrated on a subspace of lower dimension.

## 2.8 Practical Applications

*Algorithm 2.2* is applied to two real datasets. The first example is the dataset consisting of $T2$ mouse vertebra (see Dryden and Mardia, 1998, p. 9), which is explained in §2.8.1. The second example is a dataset of neural spines of $T2$ mouse vertebrae, which is discussed in §2.8.2.

### 2.8.1 Example 2.1

The method of §2.6 was applied to the real dataset consisting of $T2$ mouse vertebra, which is described by Dryden and Mardia (1998, p. 9). This dataset was obtained from an experiment whose purpose was to evaluate how the body weight of a mouse can affect the shape of its vertebra. The mice were divided into 3 differents groups of weight: control, large and small.

The proposed bootstrap methods are applied to the small group. The sample of shapes is highly concentrated around the mean shape.

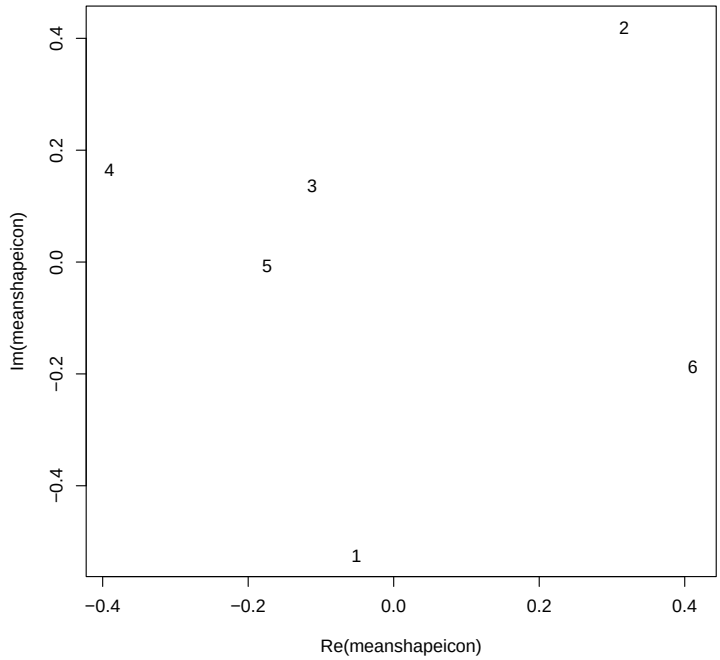


Figure 2.1: The labels of the landamrks which will be used in further comments

The labels of the landmarks are give in Figure 2.1. These labels will be used for the comments about the probability region which is consider in the next figure. The order of the labels is arbitrary.

Before applying this bootstrap method, a probability region for the small group was obtained using the NA confidence region method, which is explained in §1.6, and in particular, $(1.21)$. This is shown in Figure 2.2. It should be noted that in Figure 2.2 the NA confidence region is multiplied by $\sqrt{n}$ and so it is a probability region for the observations. The nominal levels of the NA confidence region is taken as 0.90. The dots represent the individuals and

the ellipses the NA confidence region. The NA probability region is not appropriate for the landmarks 3, 5 and 6 since many individuals are outside the ellipses of those landmarks. For the landmarks 1 and 4 the majority of the individuals are in the interior of its confidence ellipses. It means that the principal component technique is appropriate to describe the global variability of the landmarks, but it may not be suitable for representing the marginal variability, particularly those landmarks with less variability.

In Figure 2.3, a graphical comparison between bootstrap and NA confidence regions for the mean shape is shown. The nominal levels of the NA and bootstrap confidence regions are taken as 0.90.

The preceding discussion shows that the NA method can be inappropriate for some real data cases when too few PCs are used. Additionally, the bootstrap method is more robust in the sense that there is no degenerate confidence region for particular landmarks.

### 2.8.2 Example 2.2

The bootstrap method is applied to a second dataset consisting of sets of three landmarks which are obtained from twenty neural spines of $T2$ mouse vertebrae. The two main differences between this dataset and the previous one are that there is less concentration and the variances of the landmarks are more homogeneous.

In Figure 2.4, the bootstrap and NA confidence regions are shown. The two methods deliver similar results for the three landmarks. The procedures to obtain the NA and bootstrap regions were explained in §2.8.1. The two confidence regions have almost the same size as well. Thus one can conclude in this example, where the landmarks are homogeneous in relation to the

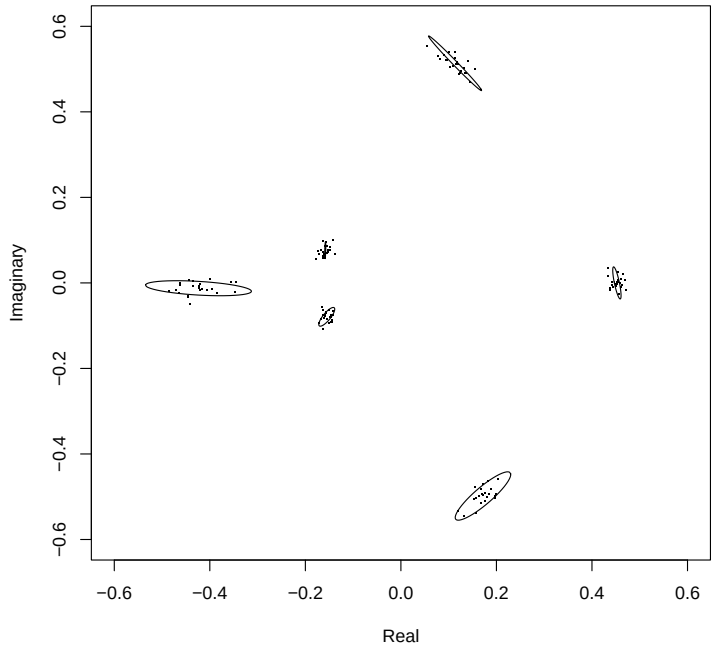


Figure 2.2: NA Probability Regions for the Individuals. These regions are obtained using the principal components method for the tangent coordinates and projecting those components back to the landmark space. The ellipses are the NA regions for the landmark and the dots are the observations. This dataset is highly concentrated and the variances of each landmark are very different.

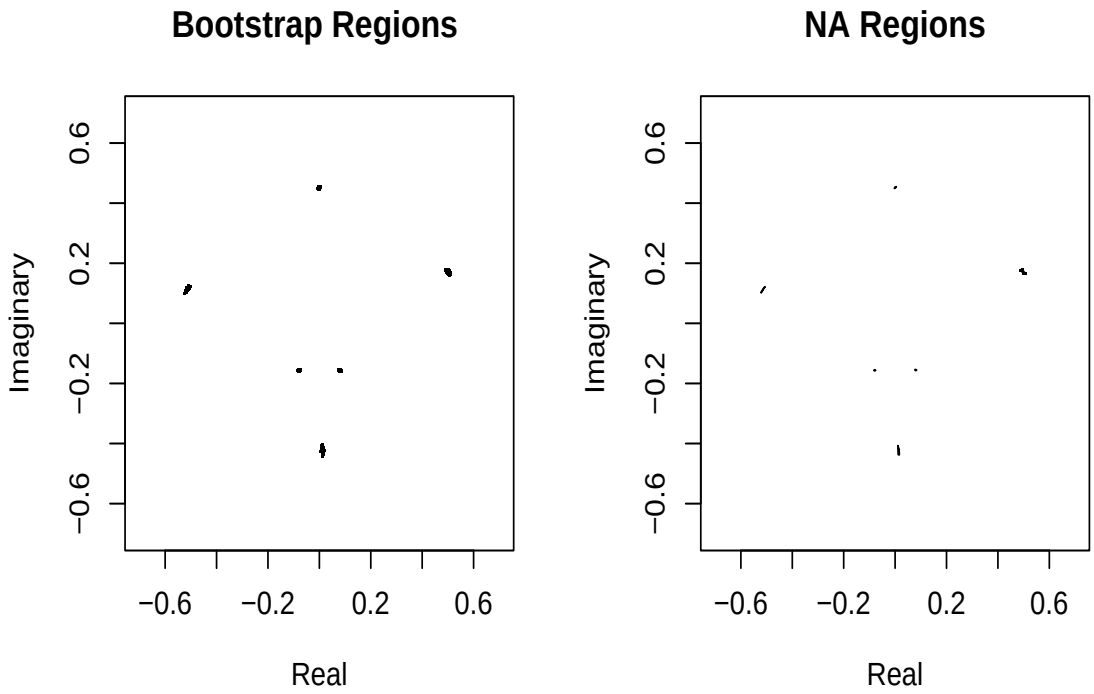


Figure 2.3: Bootstrap and NA Confidence Regions. Bootstrap regions are obtained by plotting the means of all bootstrap samples. The NA regions are smaller than the bootstrap regions

variability, the first two principal components capture nearly all variability and then the NA method is appropriate.

## 2.9 Simulation Results

In shape analysis a simulation experiment can be done either on the pre-shape space or on the landmark space. To evaluate the methods introduced here the results of some simulation experiments are presented. The first experiment is performed on the landmark space and the second on the pre-shape space. The landmark space is $\mathbb{R}^2$ and the pre-shape space was defined in $(1.5)$.

The experiment was conducted as follows: 1000 samples were generated from a complex normal distribution. The number of landmarks is $k$ and the number of observations is $n$. For each Monte Carlo sample, 200 bootstrap samples were generated. Thus the coverage probability for 4 methods were calculated from the 1000 Monte Carlo samples. The two tests described in §2.3 and the two bootstrap methods from §3 were evaluated.

In Table 2.1 the results of the coverage probability for different values of $\sigma$ and $n$ are shown. In this simulation experiment, for a chosen value for the mean shape, a multivariate complex normal is added. In this multivariate complex normal the components are not correlated and $\sigma$ is the standard deviation for each component. Note that the smaller $\sigma$ is, the more highly concentrated the data are. Two of the procedures, the Hotelling test and the Goodall test, work well only for the highly concentrated distributions. For instance, when $\sigma = 0.01$ and the sample size is bigger than 40, the estimated coverage probability of these procedures is equal to the nominal value. For distributions with low concentration, those methods do not work well at

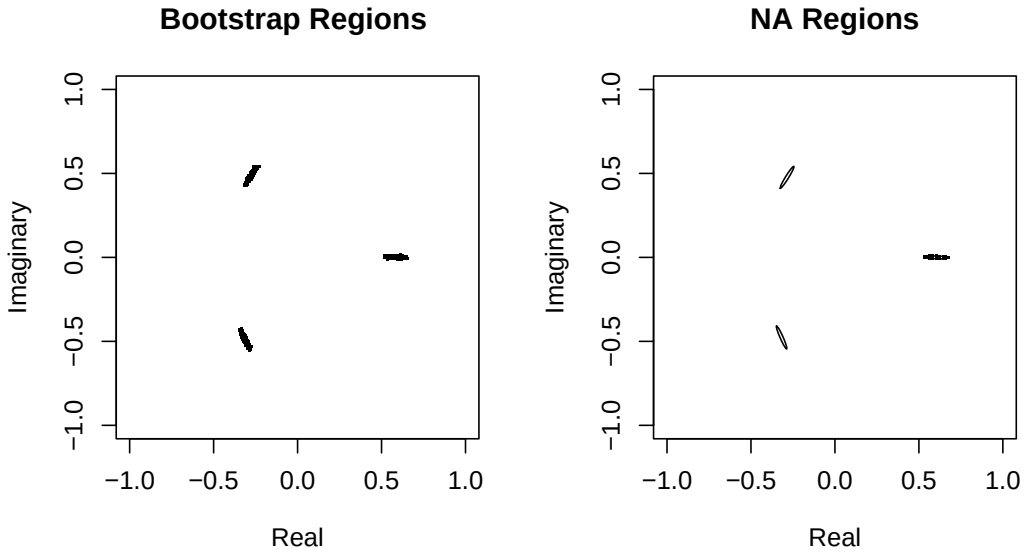


Figure 2.4: Bootstrap Confidence Regions and NA Confidence Regions. The bootstrap regions and NA regions are plotted using the same scale. The two regions are very similar for this dataset whose landmarks have similar variances.

all. For example, when $\sigma = 1$ and sample size is $30$, the coverage probability of both methods is $0.02$ which is very far from $0.90$. The two bootstrap methods work very well when the value of $\sigma$ is smaller than $0.5$. When $\sigma = 1$, the coverage probability of the pivotal bootstrap is not very close to the nominal value. However, it is considerably better than the Hotelling $T^2$ bootstrap described in §2.6.2. For example, when $n = 30$ the coverage probability of the pivotal bootstrap is $0.84$ and that for the Hotelling $T^2$ bootstrap is $0.74$.

Another simulation experiment using the complex Bingham distribution, which is defined in (1.16), was carried out. The method of §1.5.4 was used for simulating from this distribution. To use this method it is necessary to specify the eigenvalues of the matrix $A$ of (1.16). The vector of the eigenvelues of $A$ is called $\tilde{\lambda}$ and it is given in $(1.17)$.

The experiment is similar to the previous one since the number of Monte Carlo replications and bootstrap resamples were kept as 1000 and 200, respectively. The nominal value of the coverage probability is $0.90$ as before.

In Table $2.2$ the results of the estimated coverage probability for different values of $\lambda$ and $n$ are shown. There are 4 fixed values for $\lambda$ and 3 different sample sizes $n$ which are 30, 50 and 100. The values of the eigenvalues are chosen in a way that 4 situations are considered. Those situations are combinations of the cases of low and high concentration, and isotropic and non-isotropic Bingham distribution. The Watson distribution is a special case of the complex Bingham distribution that is obtained when there is a dominant eigenvalue and all remaining eigenvalues are equal. The two bootstrap methods are much better than the other methods for the low concentrated cases of both Watson and complex Bingham distributions. For example, when $k = 4$, the eigenvalues are $0$, $1$ and $2$ and the sample size is $100$, the coverage probabilities

| $\sigma$ | Sample Size | Hotelling Test (1.23) | Goodall Test (1.25) | Pivotal Bootstrap (2.15) | Hotelling's $T^2$ Bootstrap (2.17) |
|---|---|---|---|---|---|
| 0.01 | 30 | 0.92 | 0.91 | 0.92 | 0.93 |
| | 40 | 0.90 | 0.90 | 0.90 | 0.90 |
| | 50 | 0.90 | 0.90 | 0.90 | 0.90 |
| | 80 | 0.91 | 0.90 | 0.90 | 0.90 |
| 0.2 | 30 | 0.83 | 0.80 | 0.93 | 0.92 |
| | 40 | 0.81 | 0.79 | 0.90 | 0.89 |
| | 50 | 0.79 | 0.78 | 0.89 | 0.89 |
| | 80 | 0.80 | 0.79 | 0.90 | 0.90 |
| 0.3 | 30 | 0.67 | 0.63 | 0.93 | 0.93 |
| | 40 | 0.66 | 0.63 | 0.90 | 0.89 |
| | 50 | 0.63 | 0.60 | 0.89 | 0.88 |
| | 80 | 0.63 | 0.60 | 0.90 | 0.90 |
| 0.4 | 30 | 0.46 | 0.40 | 0.93 | 0.91 |
| | 40 | 0.43 | 0.37 | 0.90 | 0.89 |
| | 50 | 0.41 | 0.39 | 0.89 | 0.88 |
| | 80 | 0.40 | 0.38 | 0.90 | 0.90 |
| 0.5 | 30 | 0.29 | 0.22 | 0.92 | 0.89 |
| | 40 | 0.25 | 0.21 | 0.90 | 0.89 |
| | 50 | 0.24 | 0.21 | 0.89 | 0.88 |
| | 80 | 0.23 | 0.21 | 0.91 | 0.90 |
| 1 | 30 | 0.02 | 0.02 | 0.84 | 0.71 |
| | 40 | 0.02 | 0.01 | 0.83 | 0.72 |
| | 50 | 0.01 | 0.01 | 0.84 | 0.74 |
| | 80 | 0.00 | 0.00 | 0.88 | 0.80 |

Table 2.1: *Coverage probabilities for the Confidence Region for the Mean Shape. The simulation is performed with an isotropic complex normal distribution on the landmark space. $\sigma$ is the variance of the complex normal distribution. The nominal coverage probability is 0.90. The number of landmarks is $k = 4$.*

| Parameters Eigenvalues of the Complex Bingham | Sample Size | Hotelling Test (1.23) | Goodall Test (1.25) | Pivotal Bootstrap (2.15) | Hotelling's $T^2$ Bootstrap (2.17) |
|---|---|---|---|---|---|
| 0 0 800 | 30 | 0.900 | 0.890 | 0.899 | 0.909 |
| | 50 | 0.900 | 0.892 | 0.898 | 0.894 |
| | 100 | 0.906 | 0.901 | 0.903 | 0.901 |
| 0 50 850 | 30 | 0.900 | 0.891 | 0.899 | 0.909 |
| | 50 | 0.900 | 0.892 | 0.898 | 0.894 |
| | 100 | 0.906 | 0.899 | 0.903 | 0.901 |
| 0 0 1 | 30 | 0.023 | 0.015 | 0.822 | 0.719 |
| | 50 | 0.013 | 0.008 | 0.864 | 0.745 |
| | 100 | 0.008 | 0.011 | 0.871 | 0.823 |
| 0 1 2 | 30 | 0.057 | 0.049 | 0.863 | 0.769 |
| | 50 | 0.036 | 0.032 | 0.870 | 0.811 |
| | 100 | 0.020 | 0.024 | 0.891 | 0.857 |

Table 2.2: *Coverage probabilities for the Confidence Region for the Mean Shape. Four different special cases of the complex Bingham distribution are considered: complex watson distribution, highly concentrated; Bingham distribution, highly concentrated; complex Watson distribution, low concentrated and Bingham distribution highly concentrated.*

for the Hotelling and Goodall tests are almost zero, but the results are $0.89$ and $0.86$ for the pivotal bootstrap and Hotelling $T^2$ bootstrap, respectively, both of which are reasonably close to the nominal value $0.90$. For low concentrated distributions, the results show that the coverage probability of the pivotal bootstrap is closer to $0.90$ than the Hotelling $T^2$ bootstrap.

In Table $2.3$, results of simulation experiments with a fixed sample size $n = 30$ and several values for the paramaters are presented. For very highly concentrated distributions all the methods produce similar results. For low concentrated distributions, the pivotal bootstrap and

| Parameters Eigenvalues of the Complex Bingham | Hotelling Test (1.23 ) | Goodall Test (1.25) | Pivotal Bootstrap (2.15) | Hotelling's $T^2$ Bootstrap (2.17) | Tabular $T$ Test (2.12) |
|---|---|---|---|---|---|
| 0 0 200 | 0.897 | 0.882 | 0.899 | 0.909 | 0.857 |
| 0 0 30 | 0.866 | 0.856 | 0.901 | 0.904 | 0.857 |
| 0 0 25 | 0.851 | 0.850 | 0.902 | 0.903 | 0.858 |
| 0 0 20 | 0.837 | 0.845 | 0.903 | 0.903 | 0.859 |
| 0 0 15 | 0.817 | 0.810 | 0.897 | 0.899 | 0.859 |
| 0 0 10 | 0.772 | 0.742 | 0.901 | 0.893 | 0.860 |
| 0 0 8 | 0.72 | 0.696 | 0.898 | 0.882 | 0.857 |
| 0 0 7 | 0.669 | 0.649 | 0.901 | 0.888 | 0.857 |
| 0 0 5 | 0.533 | 0.485 | 0.897 | 0.891 | 0.846 |
| 0 0 4 | 0.433 | 0.378 | 0.901 | 0.896 | 0.854 |
| 0 0 3 | 0.290 | 0.247 | 0.897 | 0.879 | 0.844 |
| 0 0 2 | 0.125 | 0.097 | 0.880 | 0.831 | 0.782 |
| 0 0 1 | 0.023 | 0.015 | 0.821 | 0.719 | 0.672 |

Table 2.3: *Coverage probabilities for the Confidence Region for the Mean Shape for the sample size 30. In this case, 1000 Monte Carlo samples and 200 bootstrap samples are generated from the complex Watson distribution.*

Hotelling's $T^2$ bootstrap perform well. For example, when the eigenvalues of the complex Bingham are $0, 0, 3$ the coverage probability of the pivotal bootstrap is $0.897$, and the Goodall and Hotelling Tests have coverage probabilities 0.290 and 0.247.

In Table 2.4, the results of a simulation experiment where the vector of parameters is $(0, 0, 1)$ are shown for several sample sizes. When the sample size increases, the pivotal bootstrap, Hotelling's $T^2$ bootstrap and the tabular $T$ test all improve in accuracy. The coverage probabilities of the Goodall and Hotelling's $T^2$ do not change much.

| Sample Size | Hotelling Test (1.23) | Goodall Test (1.25) | Pivotal Bootstrap (2.15) | Hotelling's $T^2$ Bootstrap (2.17) | Tabular $T$ Test (2.12) |
|---|---|---|---|---|---|
| 30 | 0.023 | 0.015 | 0.821 | 0.719 | 0.672 |
| 40 | 0.012 | 0.011 | 0.843 | 0.734 | 0.686 |
| 50 | 0.013 | 0.008 | 0.863 | 0.745 | 0.723 |
| 60 | 0.012 | 0.013 | 0.863 | 0.796 | 0.749 |
| 70 | 0.015 | 0.010 | 0.865 | 0.782 | 0.771 |
| 80 | 0.011 | 0.009 | 0.858 | 0.799 | 0.775 |
| 90 | 0.010 | 0.007 | 0.874 | 0.812 | 0.779 |
| 100 | 0.008 | 0.011 | 0.872 | 0.823 | 0.785 |
| 110 | 0.014 | 0.012 | 0.865 | 0.833 | 0.787 |
| 120 | 0.007 | 0.006 | 0.871 | 0.836 | 0.798 |
| 130 | 0.008 | 0.007 | 0.879 | 0.848 | 0.815 |
| 140 | 0.008 | 0.009 | 0.876 | 0.840 | 0.802 |
| 150 | 0.004 | 0.005 | 0.873 | 0.849 | 0.810 |
| 160 | 0.006 | 0.005 | 0.882 | 0.846 | 0.814 |
| 170 | 0.007 | 0.012 | 0.873 | 0.856 | 0.827 |
| 180 | 0.010 | 0.011 | 0.881 | 0.853 | 0.825 |
| 190 | 0.009 | 0.009 | 0.868 | 0.857 | 0.817 |
| 200 | 0.013 | 0.008 | 0.862 | 0.849 | 0.820 |
| 250 | 0.008 | 0.008 | 0.874 | 0.874 | 0.841 |
| 1000 | 0.005 | 0.004 | 0.895 | 0.893 | 0.888 |

Table 2.4: *Coverage probabilities for the confidence region for the mean shape for a very low concentrated distribution. The nominal value of the coverage probability is* $0.90$. *The values of the eigenvalues of the complex Bingham are (0, 0, 1).*

# Chapter 3

# Bootstrap Tests in Statistical Shape Analysis

A suitable bootstrap method for testing equality of the mean shape of $d$ distinct populations is introduced in this chapter. The method presented here follows general guidelines for bootstrap hypotheses tests given by Fisher and Hall (1990) and Hall and Wilson (1991). According to these authors, bootstrap tests should follow the same principles as bootstrap confidence regions in that, when possible, tests should be based on statistics which are pivotal under the null hypothesis. On the other hand, the resampling scheme is not the same as for confidence regions. Bootstrap resampling should be done under the null hypothesis even if the observed samples are far from satisfying the null hypothesis. The bootstrap test presented here is related to, but extends, the bootstrap approach of Chapter 2. An important feature is that the sample of each population should be rotated in a such way that the rotated mean shape of each sample will be equal to a common mean shape, so that resampling takes place under the null hypothesis.

The outline of the chapter is as follows. Basic concepts about bootstrap hypothesis tests

are reviewed in §3.1. §3.2 describes the procedure for calculating the unitary matrices which are used to rotate the observations of each group so that the rotated samples satisfy the null hypothesis of a common mean shape. The bootstrap method is described in §3.3. The null asymptotic distribution of the test statistic is derived in §3.4 and is shown to be $\chi^2$. An example is studied in §3.5 and simulation results are presented in §3.6.

## 3.1 Bootstrap Hypothesis Testing

This section explains some issues about bootstrap hypothesis tests. It was mentioned before that even though hypothesis tests and confidence regions are related, differences between the two techniques imply that bootstrap tests require separate study. Some particular situations are considered to explain bootstrap hypothesis testing methodology. The problem of calculating a bootstrap test for a one parameter hypothesis is addressed, and some guidelines are reviewed and applied to this problem. These guidelines are also applied to the nonparametric one-way analysis of variance, which will be explained further.

Let $\tilde{u}_1, \ldots, \tilde{u}_n$ be a random sample drawn from the population $F$, where $F$ has an unknown parameter $\nu = \nu(F)$. Suppose that a procedure to test

$$H_0 : \upsilon = \upsilon_0 \text{ versus } H_1 : \upsilon \neq \upsilon_0, \tag{3.1}$$

where $\upsilon_0$ is given, needs to be developed.

The bootstrap approach is as follows. The first step is to arrange that the random sample $\tilde{u}_1, \ldots, \tilde{u}_n$ satisfies the null hypothesis by applying a suitable transformation. This issue will be discussed later on in this section. Let $u_1, \ldots, u_n$ be the transformed sample, assumed to

satisfy $H_0$. Generate $B$ samples from $u_1, \ldots, u_n$. Those samples will be named $\mathrm{u}^{(b)}$, where $b = 1, \ldots, B$. Set $T_u$ to be a statistic for the sample $\mathrm{u}$ and $T_u(b)$ to be the corresponding statistic for the bootstrap sample $\mathrm{u}^{(b)}$. Assume that larger values of $T_u$ are "more extreme" with respect to the null hypothesis. The p-value of the test is calculated by

$$p\text{-}value = \frac{(\text{number of } T_u(i) \geq T_u) + 1}{B + 1}.$$

Nonparametric bootstrap tests have the advantage that it is not necessary to choose a particular parametric family of distributions for $F$. Bootstrap methods for hypothesis tests are studied by Beran (1988), Hinkley (1988), Fisher and Hall (1990) and Hall and Wilson (1991). Fisher and Hall (1990) and Hall and Wilson (1991) have presented two main guidelines for bootstrap hypothesis testing: use a statistic which is asymptotically pivotal under the null hypothesis; and resample under the null hypothesis.

The first guideline is similar to that which is used for constructing bootstrap confidence regions, and aims to keep the level error of bootstrap tests to a minimun.

The second guideline, resampling under the null hypothesis, is necessary because p-values are based on the distribution of the test statistic under the null hypothesis.

This issue will be discussed for a simple problem to give some intuition as to why it influences the level error of bootstrap tests. Consider $u_1, \ldots, u_n$ to be a random sample drawn from the population $F$, where $F$ has an unknown scalar mean $\nu = \nu(F)$, and suppose that one wants to test the hypotheses $(3.1)$, assuming without loss of generality that $\nu_0 = 0$. This condition means that the resampling should be done using a centred version of the sample which is given by $u_1 - \hat{v}, \ldots, u_n - \hat{v}$, where $\hat{v}$ is the sample mean. It is intuitively reasonable, particularly

if the null hypothesis is far from true. If the resampling is based on $u_1 - \upsilon_0, \ldots, u_n - \upsilon_0$, the bootstrap value of the statistic $T_u(i)$ will be bigger than the sample value $T_u$ in the majority of the cases, and so the bootstrap p-value will be unreliable.

Since the problem of a bootstrap one-way analysis of variance is closely related to the bootstrap test that will be introduced in this chapter, this case also will be considered in this section (see Fisher and Hall, 1990, p. 178). Suppose that $\{u_{ij}, 1 \leq i \leq n_j\}$ is a random sample from the population $\Gamma_j$, where the population $\Gamma_j$ has mean $\upsilon_j$ and variance $\varsigma_j$, $j = 1, \ldots, p$.

It should be noted that this is a very general situation. The populations $\Gamma_j$ can belong to a broad class of distributions. If the variances $\varsigma_j$ are assumed to be the same, then it is a homoscedastic problem.

The one-way analysis of variance has the hypothesis

$$H_0 : \nu_1 = \nu_2 = \ldots = \nu_p = \nu \text{ versus } H_1 : \nu_1,\ \nu_2,\ \ldots,\ \nu_p \text{ unrestricted} \tag{3.2}$$

Following the first guideline for a bootstrap hypothesis test, Fisher and Hall (1990) obtained an asymptotically pivotal statistic in the one-way analysis of variance. For the homoscedastic one-way analysis of variance, Fisher and Hall (1990) concluded that the F-ratio statistic is not asymptotically pivotal. The F ratio is given by

$$T_1 = (n-p)\frac{\{\sum_{j=1}^{p} n_{.j}(\bar{r}_{.j} - \bar{r}_{..})^2\}}{\{\sum_{i=1}^{n}\sum_{j=1}^{p}(r_{ij} - \bar{r}_{.j})^2)\}},$$

where the variable $r_{ij} = u_{ij} - \nu_j$ is used to simplify the mathematical expression of $T$, and $n \equiv \sum_{j=1}^{p} n_j$, $\bar{r}_{.j} \equiv n_j^{-1}\sum_{j=1}^{p} r_{ij}$ and $\bar{r}_{..} \equiv n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{p} r_{ij}$.

So Fisher and Hall (1990) considered another statistic, which was proposed by James (1951), and is given by

$$T = \sum_{j=1}^{p} \left\{ \frac{\left\{ n_j (n_j - 1)(\bar{r}_{.j} - \bar{r}_{..})^2 \right\}}{\sum_{i=1}^{n} (r_{ij} - \bar{r}_{.j})^2} \right\}, \tag{3.3}$$

Fisher and Hall (1990) showed that $T$ is asymptotically pivotal, and also that the level error of this bootstrap test is $O(n^{-2})$. The asymptotic distribution of $T$ is $\chi^2_{d-2}$ and it does not depend on the $\varsigma_j$'s. It should be noted that those results are obtained under the homoscedastic assumption for the $\varsigma_j$'s (see Fisher and Hall, 1990, p. 181).

The resampling scheme is also performed according to the bootstrap test's guidelines. Consider the bootstrap samples $\left(u^{[j]}\right)^{(b)} = u_{1j}^{(b)}, \ldots, u_{1j}^{(b)}$ for $\{b = 1, \ldots, B\}$. It is also necessary to calculate the bootstrap version of $T$, which is given by

$$T^{(b)} = \sum_{j=1}^{d} \frac{n_j (n_j - 1)(\bar{r}_{.j}^{(b)} - \bar{r}_{..}^{(b)})^2}{\sum_{i=1}^{n} \left(r_{ij}^{(b)} - \bar{r}_{.j}^{(b)}\right)^2}, \tag{3.4}$$

where the variable $r_{ij} \equiv u_{ij} - \bar{u}_{.j}$. It should be noted that the Fisher and Hall (1990) method uses $r_{ij} \equiv u_{ij} - \bar{u}_{.j}$ instead of $r_{ij} \equiv u_{ij} - \mu_{.j}$. This agrees with the second guideline for bootstrap tests.

The next sections present a bootstrap test in the shape context which is analogous to the one-way analysis of variance. This test satisfies both guidelines for bootstrap hypothesis testing.

## 3.2 Rotations Determined by Geodesics

In this section a method for rotating a sample in such way that the mean shape of this sample is equal to a fixed vector is described. The aim of the bootstrap method that will be introduced

in the next section is to test the null hypothesis that the mean shape of $d$ populations are equal against the alternative that there are no constraints. Since one of the main principles of bootstrap testing is that its resampling scheme should be performed under the null hypothesis, the method described in this section will be used to make the mean shape of the sample of each population be equal to a common mean.

The rotation procedure will be described for both real and complex vectors, starting with the real case. Suppose that $a$ and $b$ are unit vectors in $\mathbb{R}^d$, and that we wish to move $b$ to $a$ along the geodesic path which connects $a$ to $b$. If $|a^T b| < 1$, a rotation matrix can be determined as follows. Define a unit vector

$$c = \frac{b - a(a^T b)}{||b - a(a^T b)||},$$

where for any vector $d$, $||d|| = (d^T d)^{1/2}$. Provided $|a^T b| < 1$, the unit vector $c$ is well defined since $||b - a(a^T b)||^2 = 1 - (a^T b)^2 > 0$, where $||b|| = \sqrt{(b^T b)}$. Suppose that $\alpha = \cos^{-1}(a^T b)$ and $A = ac^T - ca^T$.

**Proposition 3.1.** *Rotation Matrix for Real Vectors. Assume that $a, b \in \mathbb{R}^d$ are unit vectors such that $|a^T b| < 1$, and let $\alpha$, $A$ and $c$ be as defined above. The matrix*

$$Q = \exp(\alpha A) = I_d + \sum_{j=1}^{\infty} \frac{\alpha^j}{j!} A^j$$

*satisfies the following.*

*a) $Q$ is a $d \times d$ rotation matrix.*

*b) $Q$ can be written as*

$$Q = I_d + (\sin \alpha)A + (\cos \alpha - 1)(aa^T + cc^T)$$

*c) $Qb = a$*

*d) for any $z \in \mathbb{R}^d$ such that $a^T z = 0$, $b^T z = 0$, we have $Qz = z$.*

**Comments**

1. The item $d)$ can be interpreted as saying that, on the orthogonal complement of the subspace

$$\{\lambda a + \mu b : \quad \lambda, \mu \in \mathbb{R}\},$$

the matrix $Q$ acts as the identity transformation.

2. The path of minimum distance on the unit sphere in $\mathbb{R}^d$ connecting $b$ to $a$ is given by $\{x(\theta) = \exp(\theta A)b : \theta \in [0, \alpha]\}$.

3. Matrix exponentials are discussed briefly in appendix A.

The rotation matrix for complex unit vectors is obtained in similar way. However, in the application to shape analysis, a pre-shape $\tilde{b}$ has to be chosen from the shape $[b]$ of $b$, where $[.]$ was defined in $(1.6)$, $b \in \mathbb{C}^d$ and $b^\star b = 1$. Then $\tilde{b}$ moves to $a$ along a horizontal geodesic in the pre-shape space, which corresponds to a geodesic in the shape space. For practical reasons, $b$ is replaced by

$$\tilde{b} = \frac{b(b^\star a)}{|b^\star a|},$$

so that $\tilde{b}^\star a = |b^\star a|$ is real. After this change the results are very similar to the real case. Define

$$\tilde{c} = \frac{\tilde{b} - a(a^\star \tilde{b})}{||\tilde{b} - a(a^\star \tilde{b})||},$$

$\tilde{A} = a\tilde{c}^\star - \tilde{c}a^\star$ and $\tilde{\alpha} = cos^{-1}(a^\star \tilde{b})$.

**Proposition 3.2.** *Unitary Matrix for Complex Vectors. Assume that* $a, b \in \mathbb{C}^d$ *satisfy* $||a||^2 = a^\star a = 1$ *and* $b^\star b = 1$*, and suppose* $|b^\star a| < 1$*. Let* $\tilde{c}$*,* $\tilde{A}$ *and* $\tilde{\alpha}$ *be defined as above. Then the matrix*

$$U = \exp(\tilde{\alpha}\tilde{A}) = I_p + \sum_{j=1}^{\infty} \frac{\tilde{\alpha}^j}{j!}\tilde{A}^j$$

*satisfies the following.*

*(a)* $U$ *is* $d \times d$ *unitary matrix.*

*(b)* $U$ *can be written*

$$U = I_d + (\sin\tilde{\alpha})\tilde{A} + (\cos\tilde{\alpha} - 1)(aa^\star + \tilde{c}\tilde{c}^\star). \tag{3.5}$$

*(c)* $U\tilde{b} = a$.

*(d) For any* $z \in \mathbb{C}^d$ *such that* $a^\star z = 0$*,* $b^\star z = 0$*, the matrix* $U$ *is the identity transformation, i.e. ,* $Uz = z$.

The proofs of Proposition 3.1 and 3.2 are similar but only the proof of the Proposition 3.2 is presented since this is the result which is relevant to the formulation of the bootstrap test of §3.3.

**Proof of the Proposition 3.2**

**Proof of (a)** Since

$$\tilde{A}^\star = (a\tilde{c}^\star - \tilde{c}a^\star)^\star = \tilde{c}a^\star - a\tilde{c}^\star = -\tilde{A},$$

it follows that

$$\begin{aligned} U^\star &= \left(I_p + \sum_{j=1}^{\infty} \frac{\tilde{\alpha}^j}{j!}\tilde{A}^j\right)^\star \\ &= I_p + \sum_{j=1}^{\infty}(-1)^j \frac{\tilde{\alpha}^j}{j!}\tilde{A}^j \\ &= \exp(-\tilde{\alpha}\tilde{A}). \end{aligned}$$

Therefore

$$UU^\star = \exp(\tilde{\alpha}\tilde{A})\exp(-\tilde{\alpha}\tilde{A}) = I_d,$$

and so $U$ is unitary.

**Proof of (b)** The proof of this result has a few steps. Since $a^\star\tilde{c} = 0$, the following result about the matrix $\tilde{A}^2$ is derived.

$$\begin{aligned}
\tilde{A}^2 &= (a\tilde{c}^\star - \tilde{c}a^\star)(a\tilde{c}^\star - \tilde{c}a^\star) \\
&= a\tilde{c}^\star a\tilde{c}^\star - a\tilde{c}^\star\tilde{c}a^\star - \tilde{c}a^\star a\tilde{c}^\star + \tilde{c}a^\star\tilde{c}a^\star \\
&= -(aa^\star + \tilde{c}\tilde{c}^\star)
\end{aligned}$$

In addition to that, the matrix $\tilde{A}$ has the property that $\tilde{A}^3 = -\tilde{A}$, because

$$\begin{aligned}
\tilde{A}^3 &= \tilde{A}^2\tilde{A} \\
&= -(aa^\star + \tilde{c}\tilde{c}^\star)(a\tilde{c}^\star - \tilde{c}a^\star) \\
&= -aa^\star a\tilde{c}^\star + aa^\star\tilde{c}a^\star + \tilde{c}\tilde{c}^\star a\tilde{c}^\star + \tilde{c}\tilde{c}^\star\tilde{c}a^\star \\
&= -a\tilde{c}^\star + \tilde{c}a^\star \\
&= -\tilde{A}
\end{aligned}$$

Thus the matrix $\tilde{A}$ follows the general order

$$\tilde{A}^k = (-1)^j(aa^\star + \tilde{c}\tilde{c}^\star) \qquad k = 2j$$

and

$$\tilde{A}^k = (-1)^j \tilde{A} \qquad k = 2j+1,$$

where $j = 1, 2, \ldots$

Using the results above we have

$$\begin{aligned} U &= I_d + \sum_{j=1}^{\infty} \frac{\tilde{\alpha}^j}{j!} \tilde{A}^j \\ &= I_d + \left( \sum_{j=0}^{\infty} \frac{\tilde{\alpha}^{2j+1}}{(2j+1)!} (-1)^j \right) \tilde{A} \;+\; \left( \sum_{j=1}^{\infty} \frac{\tilde{\alpha}^{2j}}{(2j)!} (-1)^j \right) (-\tilde{A}^2) \\ &= I_d + (\sin\tilde{\alpha})\tilde{A} + (\cos\tilde{\alpha} - 1)(aa^\star + \tilde{c}\tilde{c}^\star). \end{aligned}$$

**Proof of (c)** To prove that $U\tilde{b} = a$ it is necessary to use the fact that

$$\tilde{b} = (\cos\alpha)a + (\sin\alpha)\tilde{c}$$

and

$$\begin{aligned} \tilde{A}\tilde{b} &= (a\tilde{c}^\star - \tilde{c}a^\star)(\cos\alpha)a + (\sin\alpha)\tilde{c} \\ &= -(\cos\alpha)\tilde{c} + (\sin\alpha)a. \end{aligned}$$

So the product $U\tilde{b}$ is calculated as

$$\begin{aligned} U\tilde{b} &= \left( I_d + (\sin\tilde{\alpha})\tilde{A} + (\cos\tilde{\alpha} - 1)(aa^\star + \tilde{c}\tilde{c}^\star) \right) \tilde{b} \\ &= (\cos\tilde{\alpha})a + (\sin\tilde{\alpha})\tilde{c} + \sin\tilde{\alpha}(-(\cos\tilde{\alpha})\tilde{c} + (\sin\tilde{\alpha})a) + (\cos\tilde{\alpha} - 1)((\cos\tilde{\alpha})a + (\sin\tilde{\alpha})\tilde{c}) \\ &= (\sin^2\tilde{\alpha} + \cos^2\tilde{\alpha})a + (\sin\tilde{\alpha}\cos\tilde{\alpha} - \sin\tilde{\alpha}\cos\tilde{\alpha})\tilde{c} \\ &= a. \end{aligned}$$

**Proof of (d)** The product $Uz$ can be written as

$$\begin{aligned} Uz &= (I_d + (\sin\tilde{\alpha})A + (\cos\tilde{\alpha} - 1)(aa^\star + \tilde{c}\tilde{c}^\star))\, z \\ &= z + (\sin\tilde{\alpha})Az + (\cos\tilde{\alpha} - 1)(aa^\star z + \tilde{c}\tilde{c}^\star z). \end{aligned}$$

On the other hand, since $a^\star z = 0$, $\tilde{b}^\star z = 0$, for any $z \in \mathbb{R}^p$, it is seen that $\tilde{c}^\star z = 0$ and $\tilde{A}z = 0$. So all the terms of $Uz$ are zero apart from the first and then $Uz = z$.

**Comment**

The set $\{\tilde{x}(\theta) = \exp(\theta\tilde{A})\tilde{b} : \theta \in [0, \tilde{\alpha}]\}$ is a horizontal geodesic in the pre-shape sphere, and therefore corresponds to a geodesic in the shape space.

## 3.3 Description of the Bootstrap Test

The bootstrap method from Chapter 2 and in particular the pivotal statistic $(2.12)$ can be extended to the problem of comparing the mean shapes of $d$ groups. The basic concepts from Chapter 2 need to be defined for the case of several populations. Let $\mathrm{y}^{[j]} \equiv \{Y_{ij}, 1 \leq i \leq n_j\}$ be a random sample of configurations from population $\prod^j$, where $1 \leq j \leq p$ denotes the population.

Let $\mathrm{w}^{[j]} = \{w_{1j}, \ldots, w_{nj}\}$ be the Helmertized configurations of $\mathrm{y}$. The complex sum of squares and product matrix for $\mathrm{w}^{[j]}$ is defined by

$$\hat{S}^{[j]} = \sum_{i=1}^{n_j} w_{ij} w_{ij}^\star / (w_{ij}^\star w_{ij}) = \sum_{i=1}^{n} z_{ij} z_{ij}^\star,$$

where $z_{ij} = w_{ij}/||w_{ij}||,\ i = 1, \ldots, n_j$ are the pre-shapes for the $jth$ group.

The spectral form of $\hat{S}^{[j]}$ is given by

$$\hat{S}^{[j]} = \sum_{q=1}^{k-1} \widehat{\lambda}_q^{[j]} \widehat{\mu}_q^{[j]} \widehat{\mu}_q^{[j]\star},$$

where $\widehat{\lambda}_1^{[j]} \geqslant \widehat{\lambda}_2^{[j]} \ldots \geqslant \widehat{\lambda}_{k-1}^{[j]} \geqslant 0$ are the eigenvalues, and $\widehat{\mu}_1^{[j]}, \ldots, \widehat{\mu}_{k-1}^{[j]}$ the corresponding eigenvectors. Thus the full Procrustes mean shape for x is given by $\widehat{\mu}^{[j]} \equiv \mu_1^{[j]}$.

A bootstrap method is introduced to test

$$H_0 : \mu^{[1]} = \mu^{[2]} = \ldots = \mu^{[p]} = \mu \text{ versus } H_1 : \mu^{[1]}, \ldots, \mu^{[p]} \text{ unrestricted.} \tag{3.6}$$

The quantities $\widehat{\Sigma}$ and $\widehat{M}_{k-2}$, introduced in $(2.13)$ and $(2.14)$, have to be defined for the case of several populations. Let $\widehat{M}_{k-2}^{[j]}$ and $\widehat{\Sigma}^{[j]}$ be the matrices $\widehat{M}_{k-2}$ and $\widehat{\Sigma}$ for sample $j$. So for the Helmertized configurations $\mathrm{w}^{[j]}$, the $(k-2) \times (k-2)$ matrix $\widehat{\Sigma}^{[j]} = (\widehat{\Sigma}_{ql}^{[j]})$ and the $(k-2) \times (k-1)$ matrix $\widehat{M}_{k-2}^{[j]}$ are given by:

$$\widehat{\Sigma}_{ql}^{[j]} = n^{-1}(\widehat{\lambda}_1^{[j]} - \widehat{\lambda}_q^{[j]})^{-1}(\widehat{\lambda}_1^{[j]} - \widehat{\lambda}_l^{[j]})^{-1} \times \sum_{i=1}^{n} ((\widehat{\mu}_q^{[j]})^\star z_i^{[j]})((z_i^{[j]})^\star \widehat{\mu}_l^{[j]})((z_i^{[j]})^\star \widehat{\mu}_q^{[j]})((\widehat{\mu}_l^{[j]})^\star z_i^{[j]}) \tag{3.7}$$

and

$$\widehat{M}_{k-2}^{[j]} = [\widehat{\mu}_2^{[j]}, \ldots, \widehat{\mu}_{k-1}^{[j]}]^\star. \tag{3.8}$$

Before giving a detailed account of the bootstrap test, it is explained how the common mean is computed. Define

$$F_B(\mu) = \sum_{j=1}^{p} \mu^\star (\widehat{M}_{k-2}^{[j]})^\star (\widehat{\Sigma}^{[j]})^{-1} \widehat{M}_{k-2}^{[j]} \mu. \tag{3.9}$$

The common mean, $\widehat{\mu}$, is defined as the complex unit vector $\mu$ which minimizes $F_B(\mu)$. From Lemma A.1, $\widehat{\mu}$ is given by the unit eigenvector corresponding to the smallest eigenvalue of

$$\sum_{j=1}^{p} (\widehat{M}_{k-2}^{[j]})^{\star} (\widehat{\Sigma}^{[j]})^{-1} \widehat{M}_{k-2}^{[j]}. \tag{3.10}$$

The test statistic is defined as

$$min_{\mu:\mu^{\star}\mu=1} F_B(\mu) = F_B(\widehat{\mu}) = \sum_{j=1}^{p} \widehat{\mu}^{\star} (\widehat{M}_{k-2}^{[j]})^{\star} (\widehat{\Sigma}^{[j]})^{-1} \widehat{M}_{k-2}^{[j]} \widehat{\mu}. \tag{3.11}$$

It should be noted that $F_B(\widehat{\mu})$ is the eigenvalue corresponding to $\widehat{\mu}$. $F_B(\widehat{\mu})$ is also an extended version of $(2.12)$.

The statistic $F_B(\widehat{\mu})$ has an asymptotic $\chi^2$ distribution under the hypothesis $(3.6)$. The proof is given in §3.4.

The bootstrap test to compare the hypotheses given in $(3.6)$ using $\widehat{\Sigma}^{[j]}$ and $\widehat{M}_{k-2}^{[j]}$ has the following steps:

*A***lgorithm 3.1.** *Bootstrap Hypothesis Test of* $(3.6)$

**Step 1** - Obtain the values of $\widehat{\mu}^{[j]}$, $\widehat{M}_{k-2}^{[j]}$ and $\widehat{\Sigma}^{[j]}$ for the pre-shape samples $z^{[j]}$, where $1 \leq j \leq p$.

**Step 2** - Obtain the pooled estimate of the common mean shape $\widehat{\mu}$, defined as the eigenvector of $(3.10)$ corresponding to the smallest eigenvalue.

**Step 3** - Rotate the pre-shapes of each group using Proposition 3.2. After this step the new sample mean shapes $\{\widehat{\mu}^{[j]},\ j = 1, .., p\}$ will all be equal to $\widehat{\mu}$.

Applying $(3.5)$, we calculate a unitary matrix which rotates $\widehat{\mu}^{[j]}$ along a "horizontal" geodesic in the pre-shape space to coincide with $\widehat{\mu}$. Let $R(\widehat{\mu}^{[j]}, \widehat{\mu})$ be this matrix. So the new set of pre-shapes $Rz^{[j]} = R(\widehat{\mu}^{[j]}, \widehat{\mu})z^{[j]},\ j = 1, \ldots, p$ will satisfy the null hypothesis.

**Step 4** - Produce $B$ bootstrap resamples from $Rz^{[j]}$ for $j = 1, \ldots, p$ and let $Rz^{[j]}(b)$ denote those resamples for $j = 1, \ldots, p$ groups, where $b = 1, \ldots, B$. For each bootstrap sample $b$ calculate $\widehat{\mu}^{[j](b)}$, $\widehat{M}_{k-2}^{[j](b)}$ and $\widehat{\Sigma}^{[j](b)}$ as the bootstrap versions of $\widehat{\mu}^{[j]}$, $\widehat{M}_{k-2}^{[j]}$ and $\widehat{\Sigma}^{[j]}$, respectively. Set $\{F_B^{(b)}(\widehat{\mu}),\ b = 1, \ldots, B\}$ as the statistic value for the bootstrap samples.

**Step 5** - Compute the p-value of the boostrap test using

$$p\text{-}value = \frac{(\text{number of} F_B^{(b)} \geq F_B(\widehat{\mu})) + 1}{B + 1},$$

where $F_B(\widehat{\mu})$ is $(3.11)$ calculated using the original sample.

## 3.4 Asymptotic Distribution of $F_B(\mu)$

The asymptotic distribution of $F_B$ is now derived. Three lemmas will be stated and proved. After that, some assumptions are stated, and then a theorem about the asymptotic distribution of $F_B$ and its proof are given.

The notation is chosen to facilitate the proofs of the lemmas and of the theorem.

*L***emma 3.1.** *Suppose that $V$ $(c \times c)$ is a complex Hermitian matrix of rank $r < c$. Let $A$ be any complex $r \times c$ matrix such that the following holds: $(i)$ the columns of $A^\star$ lie in the orthogonal complement of the null space of $V$, where the orthogonal complement is defined in $(A.2)$ and the null space is defined in $(A.1)$; and $(ii)$ $AVA^\star$ is invertible. Then*

$$V^+ = A^\star (AVA^\star)^{-1} A,$$

*where $V^+$ is the Moore-Penrose inverse of $V$ (which, by uniqueness of $V^+$, must be independent of the particular choice of A.)*

**Proof.** Since V is Hermitian of rank $r$, it admits a spectral decomposition $V = U\Lambda U$ where $\Lambda$ $(r \times r)$ is a diagonal matrix with non-zeros entries, and $U$ $(p \times r)$ satisfies $U^\star U = I_r$. By assumption $(i)$, each column of $A^\star$ can be represented as a linear combination of the columns of $U^\star$, i.e. there exist an $r \times r$ matrix $R$ such that $A = RU^\star$. For such $A$,

$$AVA^\star = RU^\star U\Lambda U^\star UR^\star = R\Lambda R^\star$$

and, since $\Lambda$ is invertible and, by $(ii)$, $AVA^\star$ is invertible, it follows that $R$ is also invertible.

Therefore

$$\begin{aligned} A^\star(AVA^\star)^{-1}A &= UR^\star(R\Lambda R^\star)^{-1}RU^\star \\ &= UR^\star(R^\star)^{-1}\Lambda^{-1}R^{-1}RU^\star \\ &= U\Lambda^{-1}U^\star \\ &= V^+. \end{aligned}$$

***Lemma 3.2.*** *Suppose that for $j = 1, .., p,\ \ y_{j,n} = y_j$ are independent sequences (indexed by $n = 1, 2, \ldots$) of random vectors and suppose that, for each $j$, $n^{1/2}(y_j - \mu) \xrightarrow{d} N_k(0_k, \Omega_j)$, where each $\Omega_j$ has full rank $k$. Suppose that the symmetric matrix $\widehat{\Omega}_{j,n} = \widehat{\Omega}_j$ is a weakly consistent estimator of $\Omega_j$ for each $j$, and define*

$$\widehat{\mu}_0 = \left(\sum_{j=1}^{p} \widehat{\Omega}_j^{-1}\right)^{-1} \sum_{j=1}^{p} \widehat{\Omega}_j^{-1} y_j.$$

*Then, as $n \to \infty$,*

$$n(\widehat{\mu}_0 - \mu)^T \left( \sum_{j=1}^{p} \widehat{\Omega}_j^{-1} \right) (\widehat{\mu}_0 - \mu) \xrightarrow{d} \chi^2_k$$

*and*

$$\begin{aligned} n\sum_{j=1}^{p}(y_j - \widehat{\mu})^T \widehat{\Omega}^{-1}(y_j - \widehat{\mu}) &= n\sum_{i=1}^{p}(y_j - \mu)^T \widehat{\Omega}_j^{-1}(y_j - \mu) \\ &\quad -n(\widehat{\mu} - \mu)^T \left( \sum_{i=1}^{p} \widehat{\Omega}^{-1} \right) (\widehat{\mu} - \mu) \xrightarrow{d} \chi^2_{(p-1)k}. \end{aligned}$$

**Proof.** The basic idea of the proof is first stated. If for each $n$ the $y_j$ were exactly normal and the $\widehat{\Omega}_j$ were exactly equal to the true $\Omega_j$ for $j = 1, \ldots, p$, then the limiting results stated in the lemma would be exactly true by standard theory for the normal linear model. The limiting results follow directly from the fact that each statistic is a jointly continuous function of the $\widehat{\Omega}_j$.

Consider

$$y_j - \mu_j \sim N_n(0_k, n^{-1}\Omega_j).$$

The likelihood function is given by

$$L(\mu_1, \ldots, \mu_p) = const - \sum_{i=1}^{n} \frac{n}{2}(y_i - \mu_i)^T {\Omega_i}^{-1}(y_i - \mu_i),$$

where const denotes a constant term. Let $\hat{\mu}_0$ be the MLE under $H_0$. It is calculated by setting $\frac{\partial L}{\partial \mu} = 0_k$, where

$$\frac{\partial L}{\partial \mu} = \frac{\partial L(\mu, \ldots, \mu)}{\partial \mu} = n\sum_{i=1}^{p} \Omega_i^{-1}(y_i - \mu).$$

The equation $\frac{\partial L}{\partial \mu} = 0$ gives

$$\sum_{i=1}^{p} \Omega_i^{-1}(y_i - \hat{\mu}_0) = O_k.$$

Therefore

$$\sum_{i=1}^{p} \Omega_i^{-1} y_i = \left( \sum_{i=1}^{p} \Omega_i^{-1} \right) \hat{\mu}_0$$

and

$$\hat{\mu}_0 = \left( \sum_{i=1}^{p} \Omega_i^{-1} \right)^{-1} \sum_{i=1}^{p} \Omega_i^{-1} y_i.$$

The expectation and covariance of $\hat{\mu}_0$ are given by

$$\begin{aligned} E(\hat{\mu}_0) &= \left( \sum_{i=1}^{p} \Omega_i^{-1} \right)^{-1} \left( \sum_{i=1}^{p} \Omega_i^{-1} \right) E(y_i) \\ &= \mu \end{aligned}$$

and

$$\begin{aligned} Cov(\hat{\mu}_0) &= \frac{1}{n} \left( \sum_{i=1}^{p} \Omega_i^{-1} \right)^{-1} \left( \sum_{i=1}^{p} \Omega_i^{-1} \right) \left( \sum_{i=1}^{p} \Omega_i^{-1} \right)^{-1} \\ &= \frac{1}{n} \left( \sum_{i=1}^{p} \Omega_i^{-1} \right)^{-1} \end{aligned}$$

respectively.

Thus, under $H_0$,

$$\widehat{\mu}_0 \sim N_k \left( \mu, \frac{1}{n} \left( \sum_{i=1}^{p} \Omega_i^{-1} \right)^{-1} \right).$$

Also, under $H_0$, using the Fisher-Cochran theorem (see Rao, 1972, pp. 185-187), we have

$$(\hat{\mu}_0 - \mu)^T \left( \sum_{i=1}^{p} \Omega_i \right) (\hat{\mu}_0 - \mu) \ \sim \ \chi_k^2.$$

The proof in the general case follows from the fact that the $y_j$ are independent of each other and for $j = 1, \ldots, p$,

$$n^{1/2}\widehat{\Omega}_j^{-1/2}(y_j - \mu) \xrightarrow{d} N_k(0_k, I_k)$$

under the assumption of the lemma.

*L***emma 3.3.** *Let $A[\epsilon] = A_0 + \epsilon A_1 + \epsilon^2 A_2 + \ldots$ denote a Hermitian matrix defined in terms of a power series in the real variable $\epsilon$ (so, in particular, each member of the sequence $A_0, A_1, \ldots$ is Hermitian). Suppose that $A_0$ has an isolated eigenvalue $\lambda_0$ and corresponding unit vector $u_0$. Then for all $\epsilon$ sufficiently small, $A[\epsilon]$ has an isolated eigenvalue $\lambda[\epsilon] = \lambda_0 + \epsilon\lambda_1 + \epsilon^2\lambda_2 + \ldots$ and corresponding unit eigenvector $u[\epsilon] = u_0 + \epsilon u_1 + \epsilon^2 u_2 + \ldots$, with*

$$\lambda_1 = u_0^\star A_1 u_0, \tag{3.13}$$

$$\lambda_2 = u_0^\star \{A_2 - A_1(A_0 - \lambda_0 I)^+ A_1\} u_0, \tag{3.14}$$

*and*

$$u_1 = (A_0 - \lambda_0 I)^+ A_1 u_0, \tag{3.15}$$

*where $(A_0 - \lambda_0 I)^+$ is the Moore-Penrose inverse of $(A_0 - \lambda_0 I)$.*

**Proof.** Writing $A[\epsilon]u[\epsilon] = \lambda[\epsilon]u[\epsilon]$ in expanded form we obtain

$$(A_0 + \epsilon A_1 + \epsilon^2 A_2 + \ldots)(u_0 + \epsilon u_1 + \epsilon^2 u_2 + \ldots) = (\lambda_0 + \epsilon\lambda_1 + \epsilon^2\lambda_2 + \ldots)(u_0 + \epsilon u_1 + \epsilon^2 u_2 + \ldots),$$

and the expressions for $\lambda_1, \lambda_2$ and $u_1$ are obtained by equating the coefficients of $\epsilon^0 = 1$, $\epsilon$ and $\epsilon^2$ to zero. We obtain the following three equations.

$$\text{coefficient of } \epsilon^0 = 1 : \qquad A_0 u_0 = \lambda_0 u_0; \tag{3.16}$$

$$\text{coefficient of } \epsilon^1 = \epsilon : \qquad A_1 u_0 + A_0 u_1 = \lambda_1 u_0 + \lambda_0 u_1 \tag{3.17}$$

and

$$\text{coefficient of } \epsilon^2: \qquad A_1u_1 + A_2u_0 + A_0u_2 = \lambda_2u_0 + \lambda_0u_2 + \lambda_1u_1. \tag{3.18}$$

Since $u_0^\star u_1 = 0$, which follows by equating the coefficient of $\epsilon$ to zero in the constraint $||u[\epsilon]||^2 = 1$, it follows from $(3.16)$ that $u_0^\star A_0 u_1 = 0$. Taking the scalar product of each side of $(3.17)$ with $u_0^\star$, and using $(3.16)$ and the fact that $u_0^\star u_0 = 1$, we obtain

$$\begin{aligned} LHS = u_0^\star(A_1u_0 + A_0u_1) &= u_0^\star A_1u_0 + \lambda_0 u_0^\star u_1 = u_0^\star A_1u_0 \\ RHS = u_0^\star(\lambda_1u_0 + \lambda_0^\star u_1) &= \lambda_1 u_0^\star u_0 + \lambda_0 u_0^\star u_1 = \lambda_1, \end{aligned}$$

where LHS is the scalar product of $u_0^\star$ and the left hand side of $(3.16)$, and RHS is the scalar product of $u_0^\star$ and the right hand side of $(3.16)$. Therefore, equating the LHS and RHS,

$$\lambda_1 = u_0^\star A_1 u_0.$$

Using $(3.17)$ to obtain $u_1$, we have

$$(A_0 - I\lambda_0)u_1 = -(A_1 - \lambda_1 I)u_0$$

from which it follows that

$$\begin{aligned} u_1 &= -(A_0 - \lambda_0 I)^+(A_1 - \lambda_1 I)u_0 \\ &= -(A_0 - \lambda_0 I)^+ A_1 u_0, \end{aligned}$$

since $u_0$ is in the null space of $(A_0 - I\lambda_0)^+$.

Pre-multiplying both sides of $(3.18)$ by $u_0^\star$, we obtain

$$\lambda_2 = u_0^\star(A_0 - \lambda_0 I)u_2 + u_0^\star A_1 u_1 + u_0^\star A_2 u_0 - \lambda_1 u_0^\star u_1.$$

Since $u_0^\star u_1 = 0$ and $u_0^\star(A_0 - \lambda_0 I) = \lambda_0 u_o^\star - \lambda_0 u_0^\star = 0$, this equation can be written as

$$\begin{aligned} \lambda_2 &= u_0^\star A_1 u_1 + u_0^\star A_2 u_0 \\ &= u_0^\star \{A_2 - A_1 (A_0 - \lambda_0 I)^+ A_1\} u_0. \end{aligned}$$

This completes the proof.

Consider the $p$ samples $\mathrm{x}^{[j]}$ and other definitions of §3.3. Using Lemma 4.2, there exists a function defined locally, such that $\widehat{M}_{k-2}^{[j]} = f(\widehat{\mu}^{[j]})$.

We make two assumptions when deriving the asymptotic distribution of $F_B(\mu)$, which is defined in $(3.11)$. The first assumption can be called *asymptotically balanced sampling*. Let $n_i = n_i(n)$ denote the size of sample $i$ $(i = 1, \ldots, p)$, viewed as a function of the sample size index $n$. Then it is assumed that $n_i(n) = n w_i(n)$ where

$$\liminf_{n\to\infty} \min_{i=1,\ldots,k} w_i(n) > 0 \text{ and } \limsup_{n\to\infty} \max_{i=1,\ldots,k} w_i(n) < \infty. \tag{3.19}$$

If (3.19) fails then the contribution of those samples whose sample size is of smaller order than the largest sample size becomes asymptotically negligible.

In addition to (3.19) , suppose that, for $j = 1, \ldots, p$,

$$n^{1/2} \widehat{M}_j \mu \xrightarrow{d} CN_{k-1}(0, \Sigma^{[j]}), \;\; \Sigma^{[j]} \text{ of full rank}, n \to \infty \tag{3.20}$$

and assume $\widehat{\Sigma}^{[j]}$ is a consistent estimator of $\Sigma^{[j]}$. Note that, from the proof of Theorem 2.1, $(3.20)$ will hold provided that population $j$ satisfies conditions $(i)$, $(ii)$ and $(iii)$ of Theorem 2.1.

**Theorem 3.1.** *Assume all the conditions of (3.19) and (3.20). Then the statistic $F_B(\widehat{\mu})$ has an asymptotic $\chi^2_{(p-1)(2k-4)}$ distribution under the null hypothesis of a common mean shape.*

**Proof.** Before considering the details , a general idea of the proof is given. The statistic used is the smallest eigenvalue of $A[n^{-1/2}]$ in $(3.21)$, and the proof has two steps:

**Step 1**: Using Lemma 3.3, the smallest eigenvalue of $A[n^{-1/2}]$ is asymptotically equivalent to $\lambda_2$ in $(3.22)$ with $A_2$ and $A_1$ defined below, and $A_0^+$ given by Lemma $3.1$.

**Step 2**: Recognise that $\lambda_2$ has the same structure as the RHS of the final term in Lemma 3.2.

Define

$$A[n^{-1/2}] = \sum_{j=1}^{p} (\widehat{M}^{[j]})^\star \left(\widehat{\Sigma}^{[j]}\right)^{-1} \widehat{M}^{[j]}, \tag{3.21}$$

where $\widehat{M}^{[j]}$ and $\left(\widehat{\Sigma}^{[j]}\right)^{-1}$ were defined in §3.3. The term $A[n^{-1/2}]$ is going to represent a Hermitian matrix defined in terms of a power series in the real variable $n^{-1/2}$. The right side of the equation is the kernel of the statistic $(3.11)$ which is given in $(3.9)$. Thus $A$ can be written

$$A[n^{-1/2}] = \sum_{j=1}^{p} (M + \widehat{M}^{[j]} - M)^\star \left(\widehat{\Sigma}^{[j]}\right)^{-1} (M + \widehat{M}^{[j]} - M).$$

Expanding $A$,

$$A[n^{-1/2}] = A_0 + n^{-1/2} A_1 + n^{-1} A_2,$$

where

$$A_0 = \sum_{j=1}^{p} M^\star \left(\widehat{\Sigma}^{[j]}\right)^{-1} M = M^\star \left( \sum_{j=1}^{p} \left(\widehat{\Sigma}^{[j]}\right)^{-1} \right) M$$

$$A_1 = n^{1/2} \sum_{j=1}^{p} \left\{ \left(\widehat{M}^{[j]} - M\right)^\star \left(\widehat{\Sigma}^{[j]}\right)^{-1} M + M^\star \left(\widehat{\Sigma}^{[j]}\right)^{-1} \left(\widehat{M}^{[j]} - M\right) \right\}$$

and

$$A_2 = n\sum_{j=1}^{p}\left(\widehat{M}^{[j]} - M\right)^{\star}\left(\widehat{\Sigma}^{[j]}\right)^{-1}\left(\widehat{M}^{[j]} - M\right).$$

Note that $(i)$ for $l \geqslant 3$, $A_l$ is the matrix of zeros; and (ii) since

$$||n^{1/2}\left(\widehat{M}^{[j]} - M\right)|| = O_p(1),$$

where $||A_1||$ and $||A_2||$ are both $O_p(1)$. In the above, $||.||$ is any suitable matrix norm such as the Euclidean norm $\{tr(A^{\star}A)\}^{1/2}$.

We now use Lemma 3.3 to determine an expansion for the smallest eigenvalue of $A[n^{-1/2}]$. We have

$$\lambda_0 = \mu^{\star}A_0\mu = 0 \text{ and } \lambda_1 = \mu^{\star}A_1\mu = 0,$$

since $M\mu = O_{k-1}$. Therefore the leading term in the expansion of the smallest eigenvalue is

$$\lambda_2 = \mu^{\star}A_2\mu - \mu^{\star}A_1A_0^{+}A_1\mu. \tag{3.22}$$

Now, if we calculate $A_1\mu$, since $M\mu = 0$, all the terms with $M$ on the left side of $A_1$ are null, and then

$$A_1\mu = n^{1/2}M^{\star}\left(\sum_{j=1}^{p}\left(\widehat{\Sigma}^{[j]}\right)^{-1}\widehat{M}^{[j]}\right)\mu,$$

and so

$$\mu^{\star}A_1^{\star}A_0^{+}A_1\mu = n\mu^{\star}\left(\sum_{j=1}^{p}\widehat{M}^{[j]\star}\left(\widehat{\Sigma}^{[j]}\right)^{-1^{\star}}\right)MA_0^{+}M^{\star}\left(\sum_{j=1}^{p}\left(\widehat{\Sigma}^{[j]}\right)^{-1}\widehat{M}^{[j]}\right)\mu.$$

But using Lemma 3.1 with $A_0^+$ and $M$,

$$\begin{aligned} A_0^+ &= M^\star (M A_0 M^\star)^{-1} M \\ &= M^\star \left( M M^\star \sum_{j=1}^{p} \left(\widehat{\Sigma}^{[j]}\right)^{-1} M M^\star \right)^{-1} M \\ &= M^\star \left( \sum_{j=1}^{p} \left(\widehat{\Sigma}^{[j]}\right)^{-1} \right)^{-1} M, \end{aligned}$$

and thus

$$M A_0^+ M^\star = \left( \sum_{j=1}^{p} \left(\widehat{\Sigma}^{[j]}\right)^{-1} \right)^{-1}.$$

Consequently,

$$\mu^\star A_1^\star A_0^+ A_1 \mu = n v^\star \left( \sum_{j=1}^{p} \left(\widehat{\Sigma}^{[j]}\right)^{-1} \right) v$$

where

$$v = \left( \sum_{j=1}^{p} \left(\widehat{\Sigma}^{[j]}\right)^{-1} \widehat{M}^{[j]} \right) \mu.$$

Also, since $M\mu = 0$,

$$\mu^\star A_2 \mu = n \mu^\star \left( \sum_{j=1}^{p} \widehat{M}^{[j]\star} \left(\widehat{\Sigma}^{[j]}\right)^{-1} \widehat{M}^{[j]} \right) \mu$$

we may apply Lemma 3.2 as follows to obtain the limiting $\chi^2$ result for $F_B(\mu)$. Put

$$y_j = (\Re(\widehat{M}_{k-2}^{[j]}\mu)^T, \Im(\widehat{M}_{k-2}^{[j]}\mu)^T)^T, \;\; \Omega_j = \frac{1}{2} \begin{pmatrix} \Re\left(\Sigma^{[j]}\right) - \Im\left(\Sigma^{[j]}\right) \\ \Im\left(\Sigma^{[j]}\right) \Re\left(\Sigma^{[j]}\right) \end{pmatrix}$$

with a corresponding definition $\widehat{\Omega}_j$ in terms of $\widehat{\Sigma}^{[j]}$. By assumption $(3.20)$ we have $n^{1/2} y_j \xrightarrow{d} N_{2k-4}(0_{2p-2}, \Omega_j)$. Also, by assumption $\widehat{\Sigma}^{[j]}$ is a consistent estimator of $\Sigma^{[j]}$, so $\widehat{\Omega}_j$ is a consistent estimator of $\Omega_j$, and therefore Lemma 3.2 may be applied. This concludes the proof of the theorem.

## 3.5 Some Applications

The bootstrap test was applied to two datasets from Dryden and Mardia (1998). The p-values of the bootstrap tests are compared to those obtained using Goodall and Hotelling tests, which are reviewed in §1.7.1 and §1.7.2.

The first dataset considered was the Gorillas Skulls (see Dryden and Mardia,1998, p. 10), which has 8 landmarks from 29 male and 30 female Gorillas. The p-values of the bootstrap, Goodall and Hotelling test were less than 0.0001 in each case. The bootstrap agrees with the other two tests in this example, where there is a very significant difference between the means of the two populations.

The second dataset is related to schizophrenic patients (see Dryden and Mardia,1998, p. 11). For this dataset, 13 landmarks are placed on a 2D image of the brains of 14 schizophrenic and 14 normal patients. The p-values for the bootstrap, Goodall and Hotelling tests were 0.0004, 0.0007 and 0.6579. Thus the bootstrap test agrees with the Goodall test. Even though the Goodall test has very strong assumptions, which are not satisfied in this example, it does not mean that the bootstrap test gives the wrong answer. The assumptions of the Hotelling test are very strong for this case as well, and so one should not trust its results. A bigger sample size would allow a better comparison between the tests and so we have carried out a simulation study.

## 3.6 Simulation Study

We consider two additional methods to test if the mean shapes of two populations are equal or not. These tests are Hotelling's $T^2$ test and Goodall's test that were described in §1.8.

In all the simulation experiments the number of Monte Carlo runs is 1000, and for each run 200 bootstrap samples were used. In each Monte Carlo run, two samples from complex Bingham distributions are generated. To evaluate the power of the tests, the true mean of one of the populations is rotated by an angle $\phi$. The parameters of these distributions are changed in each experiment in order to study some situations of interest.

In Figure 3.1 a diagram of the Monte Carlo simulation is presented. This diagram shows one Monte Carlo experiment for the case of two populations. The steps of this diagram were repeated $1000$ times in each case. The output of a pass though the diagram is a p-value. So at the end of the process the algorithm will deliver a $1000 \times 1$ vector, and the final p-value will be the average of the components of this vector.

The case of low concentrated distributions is considered in Table 3.1. The variances of the two populations are very different since the eigenvalues of the second population are equal to the eigenvalues of the second times $15$. The tests are evaluated under the null hypotheses and the size of the test is chosen as $\alpha = 0.05$. The results show that the p-values of the bootstrap test are closer to $0.05$ than those from the Hotelling and Goodall tests. For example, when the parameter vector of the first population is $\lambda = (0, 1, 2)$ and the sample size is $100$, the observed significance level of the bootstrap test, Hotelling test and Goodall test were $0.057$, $0.201$ and $0.966$, respectively . So the Goodall test completely loses its precision and the Hotelling test is not accurate for the situation considered.

Since the p-values of the tests are very different for low concentrated distributions, their power will be comparable for highly concentrated distributions only.

The results of a simulation experiment with highly concentrated and isotropic distributions

are presented in Table 3.2. In each Monte Carlo iteration two samples from Complex Bingham distributions with parameters $\lambda = (0, 400, 400)$ were simulated. For this case the two methods are expected to work well and they do. In particular, all the assumptions of the Goodall test are satisfied and therefore this test is the most powerful for the situation considered. In all the cases the Goodall test is more powerful than the Hotelling test and the power of this one is smaller than that of the bootstrap test. For example, when $n = 30$ and $\phi = 0.126$ the power of the tests has the order Goodall $(0.981) >$ Bootstrap $(0.968) >$ Hotelling $(0.961)$.

The results for nonisotropic and highly concentrated distributions are shown in Table 3.3. Two complex Bingham samples with parameters $\lambda = (0, 50, 100)$ are generated in each Monte Carlo run. The Goodall test is less powerful than the other two test as expected since this test is designed for isotropic distributions. On the other hand, bootstrap and Hotelling tests have similar power. For example, when the sample size is $100$ and $\phi = 0.031$, the tests have the following order in relation to the power Hotelling $(0.662) >$ Bootstrap $(0.656) >$ Goodall $(0.541)$.

In Table 3.4 the simulation results for the case that the populations are highly concentrated and have different variances are presented. In this simulation experiment one complex Bingham sample is generated with the parameters $\lambda = (0, 50, 100)$ and the other one with parameters $\lambda = (0, 100, 200)$.

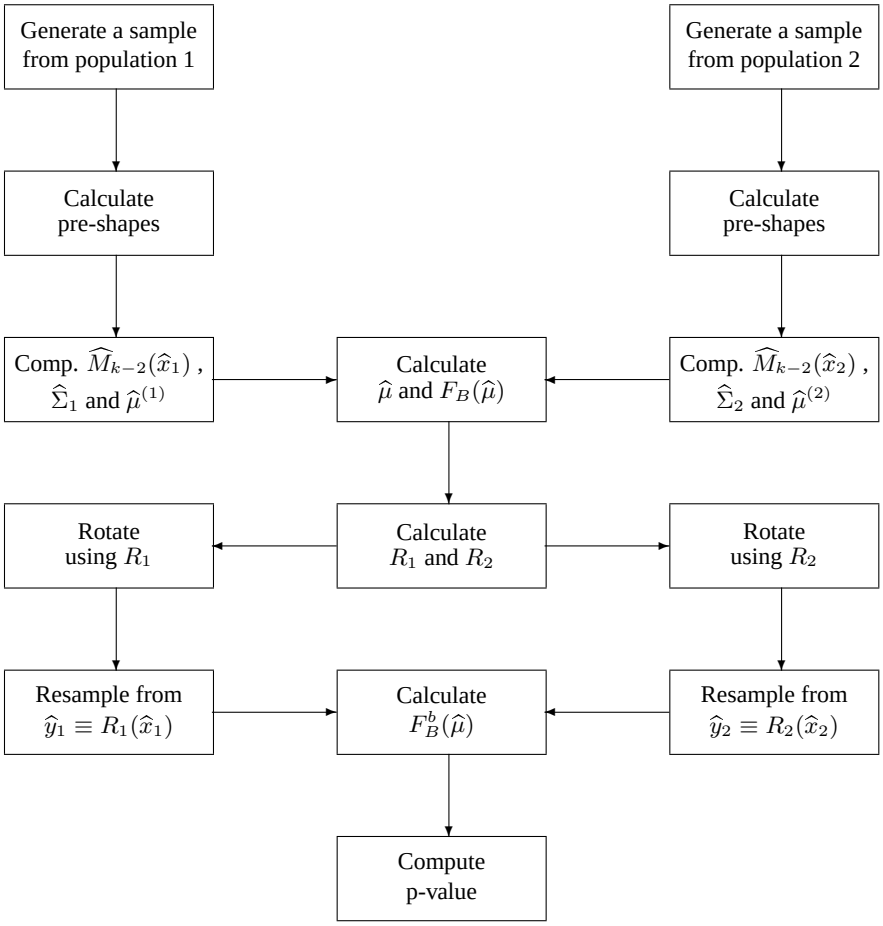


Figure 3.1: Simulation study diagram. This diagram is for the case of two populations. The details of each step are given in the algorithm. It corresponds to one iteration of a Monte Carlo simulation.

| Bingham eigenvalues | $n$ | Bootstrap $(3.11)$ | Goodall Test $(1.32)$ | Hotelling Test $(1.27)$ |
|---|---|---|---|---|
| 0, 1, 2 | 30 | 0.071 | 0.924 | 0.214 |
| | 50 | 0.066 | 0.954 | 0.204 |
| | 100 | 0.057 | 0.966 | 0.201 |
| 0, 2, 4 | 30 | 0.063 | 0.716 | 0.184 |
| | 50 | 0.052 | 0.743 | 0.192 |
| | 100 | 0.037 | 0.781 | 0.181 |
| 0, 4, 6 | 30 | 0.045 | 0.444 | 0.149 |
| | 50 | 0.049 | 0.425 | 0.149 |
| | 100 | 0.051 | 0.419 | 0.144 |
| 0, 6, 8 | 30 | 0.041 | 0.268 | 0.129 |
| | 50 | 0.057 | 0.246 | 0.115 |
| | 100 | 0.053 | 0.265 | 0.118 |

Table 3.1: *Observed significance level of the tests for populations with low concentration and heterogeneous variance structure. The vector of eigenvalues of the first and second populations are* $(0,\ 1,\ 2)$ *and* $(0,\ 15,\ 30)$, *respectively. The nominal significance level is* $0.05$.

| Sample Size | $\phi$ | Bootstrap $(3.11)$ | Goodall Test $(1.32)$ | Hotelling Test $(1.27)$ |
|---|---|---|---|---|
| 30 | 0.000 | 0.046 | 0.053 | 0.051 |
| | 0.003 | 0.047 | 0.060 | 0.055 |
| | 0.031 | 0.114 | 0.157 | 0.125 |
| | 0.063 | 0.420 | 0.483 | 0.425 |
| | 0.094 | 0.791 | 0.839 | 0.787 |
| | 0.126 | 0.968 | 0.981 | 0.961 |
| 50 | 0.000 | 0.035 | 0.041 | 0.043 |
| | 0.003 | 0.032 | 0.057 | 0.044 |
| | 0.031 | 0.180 | 0.234 | 0.196 |
| | 0.063 | 0.615 | 0.689 | 0.629 |
| | 0.094 | 0.967 | 0.980 | 0.963 |
| | 0.126 | 1 | 1 | 1 |
| 100 | 0.000 | 0.050 | 0.056 | 0.053 |
| | 0.003 | 0.051 | 0.069 | 0.053 |
| | 0.031 | 0.369 | 0.416 | 0.378 |
| | 0.063 | 0.929 | 0.957 | 0.934 |
| | 0.094 | 1 | 1 | 1 |
| | 0.126 | 1 | 1 | 1 |

Table 3.2: *Power of the tests for isotropic and highly concentrated distribution. The angular distance between the two true mean shape is $\phi$. The true eigenvalues of the populations are* $(0, 400, 400)$.

| Sample Size | $\phi$ | Bootstrap (3.11) | Goodall Test (1.32) | Hotelling Test (1.27) |
|---|---|---|---|---|
| 30 | 0.000 | 0.046 | 0.070 | 0.051 |
| | 0.003 | 0.050 | 0.070 | 0.054 |
| | 0.031 | 0.218 | 0.181 | 0.212 |
| | 0.063 | 0.770 | 0.653 | 0.756 |
| | 0.094 | 0.982 | 0.966 | 0.984 |
| | 0.126 | 1.000 | 1.000 | 1.000 |
| 50 | 0.000 | 0.036 | 0.057 | 0.043 |
| | 0.003 | 0.039 | 0.062 | 0.043 |
| | 0.031 | 0.340 | 0.271 | 0.350 |
| | 0.063 | 0.947 | 0.877 | 0.943 |
| | 0.094 | 1.000 | 1.000 | 1.000 |
| | 0.126 | 1.000 | 1.000 | 1.000 |
| 100 | 0.000 | 0.049 | 0.066 | 0.054 |
| | 0.003 | 0.051 | 0.070 | 0.055 |
| | 0.031 | 0.656 | 0.541 | 0.662 |
| | 0.063 | 1.000 | 1.000 | 1.000 |
| | 0.094 | 1.000 | 1.000 | 1.000 |
| | 0.126 | 1.000 | 1.000 | 1.000 |

Table 3.3: *Power of the tests for nonisotropic and highly concentrated distributions. Two highly concentrated complex Bingham distributions are simulated with parameter vector (0,50,100). The angular distance between the two mean shapes is* $\phi$.

| Sample Size | $\phi$ | Bootstrap (3.11) | Goodall Test (1.32) | Hotelling Test (1.27) |
|---|---|---|---|---|
| 30 | 0.000 | 0.048 | 0.071 | 0.058 |
| | 0.003 | 0.048 | 0.072 | 0.060 |
| | 0.006 | 0.051 | 0.077 | 0.066 |
| | 0.009 | 0.060 | 0.081 | 0.070 |
| | 0.019 | 0.112 | 0.106 | 0.123 |
| | 0.028 | 0.226 | 0.186 | 0.236 |
| 50 | 0.000 | 0.038 | 0.061 | 0.046 |
| | 0.003 | 0.044 | 0.064 | 0.049 |
| | 0.006 | 0.056 | 0.068 | 0.059 |
| | 0.009 | 0.070 | 0.076 | 0.081 |
| | 0.019 | 0.184 | 0.153 | 0.201 |
| | 0.028 | 0.376 | 0.291 | 0.380 |
| 100 | 0.000 | 0.045 | 0.066 | 0.054 |
| | 0.003 | 0.053 | 0.072 | 0.062 |
| | 0.006 | 0.070 | 0.080 | 0.072 |
| | 0.009 | 0.105 | 0.097 | 0.112 |
| | 0.019 | 0.354 | 0.261 | 0.369 |
| | 0.028 | 0.702 | 0.575 | 0.705 |

Table 3.4: *Power of the tests for highly concentrated populations with different dispersion structure. The parameters of the simulated complex Bingham distributions were* $0, 50, 100$ *and* $0, 100, 200$.

# Chapter 4

# Empirical Likelihood Methods in Shape Analysis

Empirical likelihood (EL) methods for building confidence regions and for testing hypotheses in the context of statistical shape analysis are studied in this chapter. The empirical likelihood method of Fisher et al. (1996) for building confidence regions for the mean direction and the mean axis can be used for building confidence regions for the mean shape. Those methods will be adapted to the shape context since the pre-shapes are complex unit vectors and not real unit vectors such as axes or directions. An extension of the method of Fisher et al. (1996) will be used to compare the means of several groups of objects.

The next sections are organized as follows. Before considering the shape context, the general idea and a literature review of EL are considered in §4.1. The formal definition and the main properties of empirical likelihood are given in §4.2. More details about EL are considered in §4.3, with a focus on inference for a univariate mean. In §4.4, the method presented by Fisher et al. (1996) for building a confidence region for the mean direction is reviewed. The

EL method for the mean shape is presented in §4.5. A method to produce a set of mutually orthogonal unit vectors, which is an important step of the EL algorithm, is presented in §4.6. In §4.7 the algorithm to calculate the EL is shown. Bootstrap calibration can be used to improve the accuracy of the EL methods, and §4.8 explains how to perform this task. A Monte Carlo simulation study of the EL methods is described in §4.9, and numerical results obtained in this study are discussed in §4.10. Graphical checking of the distribution of the EL statistic is considered in §4.11. The EL method is applied to a real data example in §4.12. The problem of using EL for hypothesis tests for several populations is addressed in §4.13. A method of EL hypothesis testing for statistical shape analysis is introduced in §4.14.

## 4.1 Main Ideas and Literature Review of Empirical Likelihood

Likelihood methods are very effective. They can be used to construct tests with good power properties, and they provide efficient estimators and small confidence regions.

However, nonparametric methods may be better than likelihood methods in some circumstances; especially when little is known about the underlying distribution. The main disadvantage of likelihood methods is that a family of distributions has to be assumed for the data. This problem can be avoided if nonparametric methods are used. In some real problems it may be hard to find a suitable parametric family of distributions. This often happens when the sample size is small but it can also happen in situations when the sample size is large.

Empirical likelihood (EL) is a type of nonparametric likelihood which can be used to obtain a nonparametric version of the theorem of Wilks (1938), which delivers an asymptotic chi-squared distribution of log likelihood ratios and therefore can be used for building confidence

regions and for testing hypotheses. More details about this theorem will be seen in $\S 4.2$.

There are many points to note about EL methods. EL does not assume a parametric family of distributions for the data. EL methods are generally very accurate when used with bootstrap calibration, a very powerful method. The shape of EL confidence regions are data-determined which does not happen with bootstrap confidence regions. EL automatically produces a pivotal statistic, and avoids the complications which can arise in constructing pivotal bootstrap statistic.

The first paper to introduce EL methods was published by Owen (1988). In that paper an EL method for the sample mean was presented. The method was based on a nonparametric analogue of Wilks (1938) theorem for parametric log-likelihood ratios. Owen (1988) presented a proof that the empirical log likelihood ratio has an asymptotic $\chi^2$ distribution under the null hypothesis, and he also compared his method to the bootstrap method in a simulation experiment using a $\chi^2$ distribution. He found out that the bootstrap-t was more accurate than the EL method in that particular setting.

In a second paper, Owen (1990) derives multivariate empirical likelihood regions for functions of several means. Multivariate means, covariance matrices and regression parameters are special cases of functions of means. For the multivariate mean, Owen (1990) illustrates in a numerical example that the shapes of empirical likelihood regions are determined by the data. He obtained a region different from an ellipse for the mean of a bivariate normal.

Owen (1991) introduced empirical likelihood methods for more complex regression models. He considered several models, including robust regression, heterocedastic regression and one-way anova.

Asymptotic properties of empirical likelihood methods and Bartlett correction have been

studied by several authors. The Bartlett correction is a scaling of the log likelihood ratio statistic which reduces the error of the asymptotic $\chi^2$ distribution under the null hypothesis from $O(n^{-1})$ to $O(n^{-2})$. So this scalar transformation can also be applied to the empirical likelihood to reduce the order of the error from $O(n^{-1})$ to $O(n^{-2})$.

DiCiccio et al. (1991) showed that the empirical likelihood is Bartlett correctable. It is a very good property of the EL method since it is the unique nonparametric method which is Bartlett correctable. They derived a general formula which can be used for parameters which can be expressed as functions of means, variance, covariance, correlation, skewness, kurtosis, mean ratio, mean difference and variance ratio.

DiCiccio et al. (1991) also showed that the bootstrap is not Bartlett correctable in any useful sense. Their arguments are based on the Edgeworth expansion. They showed that the Edgeworth expansions for bootstrap statistics have terms that cannot be removed by a simply scalar transformation like the Bartlett correction. Thus it is of considerable interest that EL is Bartlett correctable.

On the other hand, Jing and Wood (1996) showed that exponential empirical likelihood is not Bartlett correctable. They compare the relevant expansions of exponential empirical likelihood and empirical likelihood. They showed that a particular term of the expansion for the exponential empirical likelihood does not have the order $O(n^{-4})$, which is a necessary condition for it be Bartlett correctable.

The key reference for empirical likelihood and its applications is Owen (2001). Hall and La Scala (1990) give a very good review, introducing the ideas clearly. Owen also presents a list of related methods including the Bayesian bootstrap (Rubin, 1981), the nonparametric tilting

bootstrap (Efron, 1981), the survey sample estimator (Hartley and Rao, 1968) and the method of sieves (Grenander, 1981).

## 4.2 Definition and Properties of Empirical Likelihood

Owen's (1988) original idea was to use an empirical likelihood ratio to construct a confidence interval for the mean. To make this idea clear, the EL method for a functional will be reviewed. Recall that the concept of a functional is defined in $(2.1)$.

Let $m = \tilde{m}(F)$ be a population characteristic, such as the mean or the variance, of a population $F$, $\{u_1, \ldots, u_n\}$ a random sample from $F$ and $w = (w_1, \ldots, w_n)$ a vector of positive weights which sum to 1. Let $F_w$ denote the discrete probability distribution supported by the sample $\{u_1, \ldots, u_n, \}$ defined by $F_w(A) = \sum_{i=1}^n w_i I(u_i \in A)$, where $A$ is any set in the sample space and $I(.)$ is the indicator function.

The EL for $m$ is defined as

$$EL(m) = \max_{w_i \geq 0} \prod_{i=1}^{n} w_i \text{ subject to } \sum_{i=1}^{n} w_i = 1 \text{ and } \tilde{m}(F_w) = m. \tag{4.1}$$

There are some points to note about this definition.

i) If the constraint $\tilde{m}(F_w) = m$ is ignored, the EL is maximized when $w_i = 1/n$ for all $i$.

This result is easily shown using the Lagrange multiplier method. The Lagrange multiplier method to maximize a function $f(w)$ subject to the constraint $g(w) = 0$ has the following steps. One first calculates the value of $w = w_\lambda$ which solves

$$G(w) = \nabla f(w) - \lambda \nabla g(w) = 0, \tag{4.2}$$

and then obtains $\lambda$ to solve $g(w_\lambda) = 0$, where $\nabla$ is the gradient operator.

For the situation considered, ignoring the constraint $\tilde{m}(F_w) = m$, the functions used in the Lagrange multiplier method are

$$f(w) = \sum_{i=1}^{n} \log w_i \text{ and } g(w) = 1 - \sum_{i=1}^{n} w_i. \tag{4.3}$$

The equation $\nabla f(w) = \lambda \nabla g(w)$ becomes $(1/w_1, \ldots, 1/w_n) = -\lambda(1, \ldots, 1)$ which gives

$$w_i = 1/n$$

for all $i$.

However, the idea of EL is to find the set of $w_i's$ which maximizes their product subject to the constraints $\tilde{m}(F_w) = m$. This can be achieved by introducing additional Lagrange multipliers. In §4.2 and §4.3, we explain how to calculate EL for mean direction and mean shape, respectively.

ii) A major property of the EL method is that it admits a nonparametric version of Wilks's (1938) theorem.

Before discussing the Wilks's theorem for empirical likelihood, we review Wilks's theorem for parametric likelihood. Suppose that $\mathrm{u} = \{u_1, \ldots, u_n\}$ is random sample of $(q \times 1)$ random vectors, where each $u_i$ has pdf $f(u_i; \upsilon)$, where $\upsilon$ is a $(r \times 1)$ parameter vector. If a discrete variable was considered, $u_i$ would have a probability mass function (pmf). The following results are also valid for a pmf.

The likelihood function for $\upsilon$ in the case of an IID sample is defined by

$$L(\upsilon|\mathrm{u}) \equiv f(u_1, \ldots, u_n|\upsilon) = \prod_{i=1}^{n} f(u_i|\upsilon), \tag{4.4}$$

and the log-likelihood is given by

$$l(\upsilon|\mathrm{u}) = \sum_{i=1}^{n} \log f(u_i; \upsilon). \tag{4.5}$$

Suppose that parameter space is denoted by $\Upsilon$. Also denote by $\Upsilon_0 \subset \Upsilon$ a subset of strictly lower dimension than $\Upsilon$.

The maximum likelihood estimator (mle) $\hat{\upsilon}$ of $\upsilon$ under the hypothesis $\Upsilon$ is defined as

$$\hat{\upsilon} = arg \max_{\upsilon \in \Upsilon} L(\upsilon|\mathrm{u}),$$

where $\mathrm{u}$ is fixed. For testing the nested hypotheses,

$$H_0 : \nu \in \Upsilon_0 \text{ and } H_1 : \nu \in \Upsilon, \tag{4.6}$$

an asymptotic procedure can be used. Wilks (1938) proposed a theorem based on the large sample distribution of the likelihood ratio

$$\lambda(\mathrm{u}) = \frac{\sup_{\upsilon \in \Upsilon_0} L(\upsilon|\mathrm{u})}{\sup_{\upsilon \in \Upsilon} L(\upsilon|\mathrm{u})}, \tag{4.7}$$

where $L(\nu|\mathrm{u})$ is defined in $(4.4)$. For the particular case that $f$ is a pmf, intuitively the numerator of $\lambda(\mathrm{u})$ represents the maximum probability of $u$ when the parameters values are inside the set of values of the null hypothesis. The denominator is the maximum probability calculated under the more general alternative.

**Theorem 4.1.** *If $\Upsilon_0$ has $f_0$ free parameters and $\Upsilon$ has $f_1$ free parameters in the hypotheses $(4.6)$, then under mild regularly conditions, and assuming that the hypothesis $H_0$ holds, the likelihood ratio $(4.7)$ satisfies*

$$-2log(\lambda(u)) \rightarrow \chi^2_{f_1 - f_0}, \tag{4.8}$$

*when $n \rightarrow \infty$ (see Casella and Berger, 1990, p. 381).*

It is possible to present a nonparametric version of Wilks's theorem based on the empirical likelihood ratio. This EL ratio and its computation will be presented in §4.3 for the case of a scalar mean.

## 4.3 Empirical Likelihood for a Univariate Mean

The properties of EL and the EL ratio will be illustrated for the case of a univariate mean. The nonparametric version of the Wilks's theorem is presented (see Owen, 1988, p. 28). The algorithm for calculating the EL ratio and how to use this ratio to define hypothesis tests and confidence intervals is also reviewed (see Owen, 2001, p.p. 21-24).

Consider again a random sample $u_1, \ldots, u_n$ from a population with distribution function $F(u) = P(U \leq u)$. Suppose that $F$ itself is unknown with mean $\nu = E(u_i)$ and $var(u_i) < \infty$, where $E(.)$ and $var(.)$ denote the expectation and the variance, respectively. It should be noted that we can think of $\nu$ as being a functional of $F$, i. e., $\nu = \tilde{\nu}(F)$, where $\tilde{\nu}$ is the mean functional. When $F_w(.) = \sum_{i=1}^{n} w_i I_{(u_i \leq .)}$ then $\tilde{\nu}(F_w) = \sum_{i=1}^{n} w_i u_i$.

Suppose that one wants to test the hypotheses $H_0 : \nu = \nu_0$ and $H_1 : \nu$ unrestricted.

The EL ratio is given by

$$\frac{EL(\nu_0)}{EL(\widehat{\nu})},$$

where $EL(\nu_0)$ is the EL evaluated under $H_0$ and $EL(\widehat{\nu})$ is the maximised EL under $H_1$. Thus, using definition (4.1),

$$EL(\nu_0) = \max_{w_i \geq 0} \prod_{i=1}^{n} w_i \text{ subject to } \sum_{i=1}^{n} w_i = 1 \text{ and } \sum_{i=1}^{n} w_i u_i = \nu_0.$$

Also, from the discussion following $(4.3)$,

$$EL(\hat{\nu}) = n^{-n}.$$

For the parameter $\nu$ the profile EL is defined as

$$R(\nu) = \frac{EL(\nu)}{EL(\widehat{\nu})} = \max_{w_i \geq 0} \left\{ \prod_{i=1}^{n} (nw_i) \;\; | \;\; \sum_{i=1}^{n} w_i u_i = \nu, \text{and} \;\; \sum_{i=1}^{n} w_i = 1 \right\}. \tag{4.9}$$

To find the weights $w_i, \;\; i = 1, \ldots, n$, the Lagrange multipliers method, which was explained in §4.2, is used (see Owen, 2001, p. 22). The function $G$, defined in $(4.2)$, becomes

$$G(w) = \sum_{i=1}^{n} log(nw_i) - \lambda_1 \sum_{i=1}^{n} w_i(u_i - \nu) - \lambda_2 \left( \sum_{i=1}^{n} w_i - 1 \right),$$

where $\lambda_1$ and $\lambda_2$ are Lagrange multipliers.

The first step of the Langrange multiplier method is to differentiate the function $G$ and calculate the critical values of this function, where the critical values are the points where the derivate function is zero. The derivatives of the function $G$ are

$$\frac{\partial G}{\partial w_i} = \frac{1}{w_i} - \lambda_1(u_i - \nu) - \lambda_2,$$

for $i = 1, \ldots, n$.

Solving $\sum_{i=1}^{n} w_i \frac{\partial G}{\partial w_i} = 0$, we obtain $n - \lambda_2 = 0$ or $\lambda_2 = n$.

Thus, setting $\frac{\partial G}{\partial w_i} = 0$,

$$w_i = \frac{1}{n} \frac{1}{1 + \lambda_1(u_i - \nu)}, \tag{4.10}$$

where $\lambda_1$ solves

$$\sum_{i=1}^{n} w_i(u_i - \nu) = \frac{1}{n} \sum_{i=1}^{n} \frac{u_i - \nu}{1 + \lambda_1(u_i - \nu)} = 0,$$

and can be obtained numerically.

One of the important properties of EL is that it can be used to obtain a nonparametric version of Theorem 4.1. The details of this result in the case of a scalar mean are given in the following theorem.

**Theorem 4.2.** *Let $u_1, \ldots, u_n$ be independent identically distributed random variables with common distribution $F(u)$. Let $\nu_0 = E(u_i)$ and suppose that $0 < Var(u_i) < \infty$. Then $as\ n \to \infty$ the log empirical likelihood ratio satisfies*

$$-2logR(\nu_0) \xrightarrow{d} \chi^2_1. \tag{4.11}$$

**Proof.** Using expression $(4.10)$ for $w_i$, and viewing it as a function of $\lambda$, define

$$\begin{aligned} f(\lambda) &= \sum_{i=1}^{n} w_i(u_i - \nu) \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{u_i - \nu}{1 + \lambda(u_i - \nu)}. \end{aligned}$$

Using Taylor's expansion around the point $a$, we have

$$f(\lambda + a) = \sum_{k=0}^{\infty} \frac{\lambda^k f^{[k]}(a)}{k!} = f(a) + \frac{\lambda f^{[1]}(a)}{1!} + \frac{\lambda^2 f^{[2]}(a)}{2!} + \ldots, \tag{4.12}$$

where $f^{[k]}$ is the $k$-th derivative of $f$. Assuming $\lambda$ is small and therefore ignoring the terms after $k = 1$, the value of $\lambda$ is obtained from

$$\frac{1}{n}\sum_{i=1}^{n}(u_i - \nu) - \lambda\frac{1}{n}\sum_{i=1}^{n}(u_i - \nu)^2 = 0,$$

and it is given by

$$\lambda = \frac{\bar{u} - \nu}{S(\nu)}, \tag{4.13}$$

where $S(\nu) = \frac{1}{n}\sum_{i=1}^{n}(u_i - \nu)^2$.

To find the asymptotic distribution of the profile empirical likelihood ratio, it is necessary to apply the Taylor approximation to the expression of $R(\nu)$, which is given in $(4.9)$. Substituting $(4.10)$ in $(4.9)$, one can write

$$-2\log R(\nu) = 2\sum_{i=1}^{n}\log(1+\lambda(u_i-\nu)).$$

The second step is to apply Taylor's approximation to this function. Thus, one has to apply $(4.12)$ to

$$f(\lambda) = \sum_{i=1}^{n}\log(1+\lambda(u_i-\nu)).$$

Since

$$f^{[1]}(a) = \sum_{i=1}^{n}\frac{(u_i-\nu)}{1+a(u_i-\nu)}$$

and

$$f^{[2]}(a) = -\sum_{i=1}^{n}\frac{(u_i-\nu)^2}{(1+a(u_i-\nu))^2},$$

we have $f(0) = 0$, $f^{[1]}(0) = \sum_{i=1}^{n}(u_i-\nu)$ and $f^{[2]}(0) = \sum_{i=1}^{n}(u_i-\nu)^2$.

Thus $R(\nu)$ can be approximated by

$$R(\nu) \quad = \quad 2\lambda\sum_{i=1}^{n}(u_i-\nu) - \lambda^2\sum_{i=1}^{n}(u_i-\nu)^2 \tag{4.14}$$

and one can substitute the value of $\lambda$, given in $(4.13)$, in $(4.14)$. This gives

$$\begin{aligned} R(\nu) &= 2\frac{(\bar{u}-\nu)}{S(\nu)}\sum_{i=1}^{n}(u_i-\nu) - \frac{(\bar{u}-\nu)^2}{S(\nu)^2}\sum_{i=1}^{n}(u_i-\nu)^2 \\ &= 2n\frac{(\bar{u}-\nu)^2}{S(\nu)} - n\frac{(\bar{u}-\nu)^2}{S(\nu)} \\ &= n\frac{(\bar{u}-\nu)^2}{S(\nu)} \\ &= \left(\frac{\sqrt{n}(\bar{u}-\nu)}{\sqrt{S(\nu)}}\right)^2. \end{aligned}$$

By the law of large numbers, $S(\nu_0) \xrightarrow{p} var(u_i)$ under $H_0$ and by the central limit theorem,

$$\left( \frac{\sqrt{n}(\bar{u} - \nu)}{\sqrt{S(\nu)}} \right) \xrightarrow{d} N(0,1),$$

and therefore

$$R(\nu) = \left( \frac{\sqrt{n}(\bar{u} - \nu)}{\sqrt{S(\nu)}} \right)^2 \xrightarrow{d} \chi^2_1.$$

The proof of this result for the case that $\nu$ is a vector, which is broadly similar, is given by Owen (2001, pp. 219-222).

## 4.4 Empirical Likelihood Regions for The Mean Direction

The EL method of Fisher et. al. (1996) for directional data is now reviewed since it is closely related to the method for shape data that will be explained in §4.5.

Let $m$ be a unit vector in $\mathbb{R}^3$, so $m$ is a point on the sphere $\mathbb{S}^3 = \{m \in \mathbb{R}^3 : ||m|| = 1\}$. Any vector $m \in \mathbb{S}^3$ can be written as

$$m = (\cos(\theta), \sin(\theta)\cos(\phi), \sin(\theta)\sin(\phi))^T, \tag{4.15}$$

where $0 \leq \theta \leq \pi$ and $0 \leq \phi \leq 2\pi$.

Let $\mathrm{x} = \{x_1, \ldots, x_n\}$ be a random sample of unit $3-$vectors from a population $F$. In the case of a mean direction, the empirical likelihood $EL(m)$ at a candidate mean direction $m$ is defined as follows:

$$EL(m) = \max_{w_i \geq 0} \prod_{i=1}^{n} w_i \text{ subject to } \sum_{i=1}^{n} w_i = 1 \text{ and } \tilde{m}(w) \equiv \sum_{i=1}^{n} w_i x_i / || \sum_{i=1}^{n} w_i x_i || = m.$$

Since the sphere $\mathbb{S}^3$ is not a Euclidean space, the constraint $\sum_{i=1}^{n} w_i x_i / || \sum_{i=1}^{n} w_i x_i || = m$ cannot be used directly in the Lagrange multiplier method. The reason is that $m$ is constrained to lie in the unit sphere. The constraint needs to be represented in a suitable form for the Lagrange multiplier method to be applicable.

Let us first consider the case of a mean direction in $\mathbb{R}^3$. The suitable constraints are given by

$$m_1^T m = m_2^T m = m_1^T m_2 = 0, \tag{4.16}$$

where the unit vectors $m_1 = m_1(m)$ and $m_2 = m_2(m)$ are chosen to be mutually orthogonal and orthogonal to $m$.

It should be noted that $m$ in $(4.15)$ can represent any direction. However, the question is how to represent $m_1$ and $m_2$ as a function of $m$ in a such way that $(4.16)$ is true if $m$ is written in the form $(4.15)$. If $m$ is written in the form $(4.15)$, then $m_1$ and $m_2$ can be written

$$m_1^T = (\cos\theta\cos\phi, \cos\theta\sin\phi, -\sin\theta), \tag{4.17}$$

and

$$m_2^T = (-\sin\phi, \cos\phi, 0). \tag{4.18}$$

Note that this is just one of an infinite number of possible choices.

The function to be maximised by the Lagrange multiplier method is given by

$$G(w) = \sum_{i=1}^{n} log(w_i) + \lambda_0(1 - \sum_{i=1}^{n} w_i) - \lambda_1 m_1^T \sum_{i=1}^{n} w_i x_i - \lambda_2 m_2^T \sum_{i=1}^{n} w_i x_i,$$

where $\lambda_0$, $\lambda_1$ and $\lambda_2$ are Lagrange multipliers corresponding to the three constraints. The first step is to calculate the partial derivatives

$$\frac{\partial G}{\partial w_i} = \frac{1}{w_i} - \lambda_0 - \lambda_1 m_1^T x_i - \lambda_2 m_2^T x_i, \tag{4.19}$$

and set them to zero. After that, we may solve $\sum_{i=1}^{n} w_i(\partial G/\partial w_i) = 0$ to obtain $\lambda_0$.

Since

$$m = \sum_{i=1}^{n} \frac{w_i x_i}{||\sum_{i=1}^{n} w_i x_i||} \text{ and } \sum_{i=1}^{n} w_i = 1,$$

we have

$$\begin{aligned}
\sum_{i=1}^{n} w_i \frac{\partial G}{\partial w_i} &= n - \lambda_0 - \lambda_1 m_1^T \sum_{i=1}^{n} w_i x_i - \lambda_2 m_2^T \sum_{i=1}^{n} w_i x_i, \\
&= n - \lambda_0 - \lambda_1 ||\sum_{i=1}^{n} w_i x_i|| m_1^T m - \lambda_2 ||\sum_{i=1}^{n} w_i x_i|| m_2^T m \\
&= n - \lambda_0,
\end{aligned}$$

so the solution is $\lambda_0 = n$.

Replacing $\lambda_0$ by $n$ in (4.19) and setting $\partial G/\partial w_i = 0$, the weights are given by

$$w_i = \frac{1}{n(1 + \lambda_1 m_1^T x_i + \lambda_2 m_2^T x_i)}, \tag{4.20}$$

redefining $\lambda_1$ and $\lambda_2$ by $\lambda_1/n$ and $\lambda_2/n$, respectively, for notational convenience.

In order to obtain a confidence region, consider a coordinate system such that the sample mean direction $\widehat{m} = \frac{\sum_{i=1}^{n} x_i}{||\sum_{i=1}^{n} x_i||}$ is given by

$$\widehat{m}^T = (0, 0, 1).$$

Let $\log R(m)$ be the empirical log-likelihood ratio which is given by

$$\log R(m) = \log\{EL(m)/EL(\widehat{m})\}.$$

The EL confidence region with confidence coefficient $\alpha$ for the mean direction is given by

$$R_\alpha = \{m : \log R(m) \leqslant \rho_\alpha\}, \tag{4.21}$$

where $\rho_\alpha$ is chosen to satisfy $P(\log R(m_0) \leqslant \rho_\alpha) = 1 - \alpha$, under the null hypothesis

$$H : m = m_0,$$

where $m_0$ is the true value.

## 4.5 Empirical Likelihood Regions for the Mean Shape

This section describes how to adapt the EL method for axial datasets of Fisher et al. (1996) to shape datasets. The empirical likelihood confidence region for the mean shape is calculated similarly to the mean axis. The steps are similar but the constraints will be different.

Consider a random sample of preshapes $z = \{z_1, \ldots, z_n\}$, as described in § 2.1. Here, the relevant constraint is that $m$ is an eigenvector of the matrix $S(w) = \sum_{i=1}^{n} w_i z_i z_i^\star$ corresponding to the largest eigenvalue, where the $w_i$ are non-negative weights which sum to 1 and are to be determined.

If $\mathbf{Re}(a)$ and $\mathbf{Im}(a)$ represent the real and imaginary part of a complex vector $a$, then the constraints are given by

$$\mathbf{Re}\{m_j^\star S(w) m\} = 0,\ \mathbf{Im}\{m_j^\star S(w) m\} = 0,\ j = 1, \ldots, k-2.$$

It should be noted that the number of $m_j$ vectors is $k-2$ because $k-1$ is the dimension of a Helmertized vector and one of those $k-1$ vectors is the mean shape $m$. So there are only $k-2$ vectors remaining.

Define

$$\gamma_j(w) = m_j^\star S(w) m \tag{4.22}$$

and

$$\delta_{ij} = m_j^\star z_i z_i^\star m = \frac{\partial \gamma_j(w)}{\partial w_i}. \tag{4.23}$$

Using the definitions $(4.22)$ and $(4.23)$, the profile empirical likelihood function for the mean shape is given by

$$EL(\mu) = \max_{w_i \geq 0} \prod_{i=1}^{n} w_i \text{ subject to } \sum_{i=1}^{n} w_i = 1 \tag{4.24}$$

$$\text{and } \sum_{j=1}^{k-2} \{\lambda_j^{(R)} \mathbf{Re}(\gamma_j(w)) + \lambda_j^{(I)} \mathbf{Im}(\gamma_j(w))\} = 0. \tag{4.25}$$

Thus the function to be maximised by the Lagrange multiplier method is given by

$$G(w) = \sum_{i=1}^{n} \log w_i + \lambda_0 (1 - \sum_{i=1}^{n} w_i) + \sum_{j=1}^{k-2} \{\lambda_j^{(R)} \mathbf{Re}(\gamma_j(w)) + \lambda_j^{(I)} \mathbf{Im}(\gamma_j(w))\}.$$

The partial derivatives of $G$ are given by

$$\frac{\partial G(m)}{\partial w_i} = \frac{1}{w_i} - \lambda_0 + \sum_{j=1}^{k-2} \{\lambda_j^{(R)} \mathbf{Re}(\delta_{ij}) + \lambda_j^{(I)} \mathbf{Im}(\delta_{ij})\}. \tag{4.26}$$

Multiplying by $w_i$ and summing, it is seen that

$$\sum_{i=1}^{n} w_i \frac{\partial G}{\partial w_i} = n - \lambda_0 \sum_{i=1}^{n} w_i + \sum_{j=1}^{k-2} \{\lambda_j^{(R)} \sum_{i=1}^{n} w_i \mathbf{Re}(\delta_{ij}) + \lambda_j^{(I)} \sum_{i=1}^{n} w_i \mathbf{Im}(\delta_{ij})\}.$$

At the optimum $w$, $\sum_{i=1}^{n} w_i = 1$, $\delta_j(w) = 0$, and $\frac{\partial G}{\partial w_i} = 0$, so that $n - \lambda_0 = 0$, i.e. $\lambda_0 = n$. The optimum weights can now be calculated from $(4.26)$ :

$$w_i = \frac{1}{n\Big(1 + \sum_{j=1}^{k-2} \Big\{\lambda_j^{(R)} \mathbf{Re}(\delta_{ij}) + \lambda_j^{(I)} \mathbf{Im}(\delta_{ij})\Big\}\Big)} \quad i = 1, \ldots, n.$$

where $\lambda_j^{(R)}$ and $\lambda_j^{(I)}$ have been redefined as $\lambda_j^{(R)}/n$ and $\lambda_j^{(I)}/n$. Substituting for $w_i$ in the constraints, it is seen that $\lambda_j^{(R)}$ and $\lambda_j^{(I)}$ must satisfy

$$\sum_{i=1}^{n} \Big\{1 + \sum_{j=1}^{k-2} \Big(\lambda_j^{(R)} \mathbf{Re}(\delta_{ij}) + \lambda_j^{(I)} \mathbf{Im}(\delta_{ij})\Big)\Big\}^{-1} \delta_{im} = 0, \quad m = 1, 2. \tag{4.27}$$

The analogue of Theorem $4.2$ in the case of the mean shape is given by the following result.

**Theorem 4.3.** *If the population mean shape $[m_0]$ is well-defined and the population distribution has a density with respect to the uniform distribution on $CS^{k-1}$, then*

$$-2log\{EL(m_0)/EL(\widehat{m})\} \xrightarrow{d} \chi^2_{2k-4}, \tag{4.28}$$

*where $m_0$ is any pre-shape corresponding to $[m_0]$.*

The idea of the proof is similar to that given in the scalar mean case. See Owen (2001) for the proof of the vector mean case.

## 4.6 Explicit Calculation of a set of Orthogonal Unit Vectors

When applying the empirical likelihood approach to directional data or shape data consisting of, respectively, unit vectors in $\mathbb{R}^k$ or $\mathbb{C}^k$, it is necessary to perform the following task repeatedly: given a unit vector $m$, determine a set of mutually orthogonal unit vectors $m_1, \ldots, m_{k-1}$ which are orthogonal to $m$. This can be done conveniently using the following results.

**Lemma 4.1.** *(The real case.) Suppose that $c \in \mathbb{R}$, where $c > -1$, and $b \in \mathbb{R}^k$ are such that $m = \binom{b}{c}$ is a unit vector in $\mathbb{R}^k$, i. e. $c^2 + ||b||^2 = 1$. Define the $(k-1) \times k$ matrix $A = [A_1 : A_2]$ by*

$$A_1 = I_{k-1} - (1+c)^{-1} bb^T, \; A_2 = -b,$$

*where, by implication, $A_1$ is $(k-1) \times (k-1)$ and $A_2$ is $(k-1) \times 1$. Then $(i)$ $Am = 0_{k-1}$ and $(ii)$ $AA^T = I_{k-1}$.*

**Lemma 4.2.** *(The complex case.) Suppose $c \in \mathbb{C}$, $c \neq 0$, $b \in \mathbb{C}^{k-1}$, and $m = \binom{b}{c}$ is a complex unit vector in $\mathbb{C}^k$ , so that $c^\star c + ||b||^2 = 1$. Define the $(k-1) \times k$ matrix $A = [A_1 : A_2]$ by*

$$A_1 = \frac{c}{|c|} I_{k-1} - \frac{c}{|c|}(1+|c|)^{-1} bb^\star, \; A_2 = -b$$

*where $A_1$ is $(k-1) \times (k-1)$ and $A_2$ is $(k-1) \times 1$. Then $(i)$ $Am = 0_{k-1}$ and $(ii) AA^\star = I_{k-1}$.*

**Comment**. Given $m$, we may choose $m_1, \ldots, m_{k-1}$ as follows: in the real case, as the columns of $A^T$; and in the complex case, as the columns of $A^\star$.

The proofs of the lemmas are very similar. Only Lemma 4.2 is proved here since it is the one which is relevant for the method of the next section.

**Proof of Lemma 4.2**

(i) From the definitions,

$$\begin{aligned} Am &= [A_1 : A_2]\begin{pmatrix} b \\ c \end{pmatrix} \\ &= A_1 b + A_2 c \\ &= \frac{c}{|c|} b - \frac{c}{|c|}(1+|c|)^{-1}(b^\star b)b - cb \\ &= \frac{c}{|c|} b + (c - \frac{c}{|c|})b - cb \\ &= 0_{k-1}, \end{aligned}$$

since $b^\star b = 1 - |c|^2$.

(ii) Note that

$$\big[A_1 : A_2\big]\big[A_1 : A_2\big]^\star = A_1 A_1^\star + A_2 A_2{}^\star,$$

and

$$\begin{aligned} A_1 A_1^\star &= \Big(\frac{c}{|c|} I_{k-1} + \Big(c - \frac{c}{|c|}\Big)\frac{bb^\star}{||b||^2}\Big)\Big(\frac{c}{|c|} I_{k-1} + \Big(c - \frac{c}{|c|}\Big)\frac{bb^\star}{||b||^2}\Big)^\star \\ &= \frac{c^\star c}{|c|^2} I_{k-1} + \Big[\Big(c - \frac{c}{|c|}\Big)\Big(c - \frac{c}{|c|}\Big)^\star + \frac{c}{|c|}\Big(c - \frac{c}{|c|}\Big)^\star + \frac{c^\star}{|c|}\Big(c - \frac{c}{|c|}\Big)\Big]\frac{bb^\star}{||b||^2} \\ &= I_{k-1} - (1 - c^\star c)\frac{bb^\star}{||b||^2} \\ &= I_{k-1} - bb^\star, \end{aligned}$$

since $||b||^2 = b^\star b = 1 - c^\star c$.

Therefore, since $A_2 A_2^\star = bb^\star$, we have

$$A_1 A_1^\star + A_2 A_2^\star = I_{k-1} - bb^\star + bb^\star = I_{k-1}$$

as required.

## 4.7 Algorithm

The algorithm to find EL estimates is reviewed in this section. The estimating equations given in (4.27) are in closed form. So the set of Lagrange multipliers $\lambda_j^{(R)}$ and $\lambda_j^{(I)}$ should be evaluated numerically. A algorithm which uses Owen's algorithm for multivariate vectors is introduced. This involves separating the real and imaginary parts of the pre-shape vectors and applying Owen's algorithm to those two parts in the way explained below.

Let $L_{EL}(m)$ be the empirical likelihood $(4.24)$ and $(4.25)$ evaluated at $m$. To calculate $L_{EL}(m)$ the following steps should be performed.

**Algorithm 4.1.** *Calculating the Empirical Likelihood*

*Step 1* - Given $m$, find a set of mutually orthogonal unit vectors $m_1, \ldots, m_{k-2}$ also orthogonal to $m$, using Lemma 4.2.

*Step 2* - Calculate a vector $\delta_i$ with components

$$\delta_i^T = [Re(\delta_{i1}), Im(\delta_{i1}), \ldots, Re(\delta_{ik-2}), Im(\delta_{ik-2})],$$

where $i = 1, \ldots, n$ and $\delta_{ij} = m_j^\star z_i z_i^\star m$.

*Step 3* - The vector $\delta_1, \ldots, \delta_n$ can be used in an empirical likelihood procedure for a $2(k-2)$ real vector to find $\widehat{\lambda}$ which maximizes

$$\sum_{i=1}^{n} \log(1 + \lambda^T \delta_i).$$

Thus Step 3 uses Owen's S-Plus function which is explained in Appendix D.

*Step 4* - The weights are given by

$$w_i = \frac{1}{n(1 + \widehat{\lambda}^T \delta_i)}.$$

*Step 5* - The loglikelihood ratio is

$$W_{EL}(m) = -2 \sum_{i=1}^{n} log(n w_i). \tag{4.29}$$

After calculating the empirical log-likelihood, the confidence region can be defined by

$$\mathrm{R}_\alpha = \{m : W_{EL}(m) \leq l_\alpha\}, \tag{4.30}$$

where $\alpha$ is the chosen confidence level. Asymptotically, $L_{EL}(m)$ has a $\chi^2_{2k-4}$ distribution by Theorem 4.3, hence the constant $l_\alpha$ is approximately given by

$$P(\chi^2_{2k-4} \leq l_\alpha) = 1 - \alpha.$$

Bartlett correction was mentioned in the introduction. It is a scalar transformation that, when applied to a log likelihood ratio statistic, reduces the order of the error under the null hypothesis from $O(n^{-1})$ to $O(n^{-2})$. Bartlett correction or bootstrap calibration can be used to improve the coverage probability of $R_\alpha$. Thus corrected values for $l_\alpha$ would be used. The bootstrap calibrated version of $l_\alpha$ is presented in §4.8.

## 4.8 Bootstrap Calibration

The combination of EL and bootstrap methods delivers very accurate results. The empirical loglikehood ratio statistic has an asymptotic $\chi^2$ distribution (see Hall and La Scala (1990)) under the null hypothesis. Bootstrap calibration using this statistic will reduce coverage error from $O(n^{-1})$ to $O(n^{-2})$; see Fisher et al. (1996).

The bootstrap algorithm in this case can be described as follows:

*A***lgorithm 4.2.** *Bootstrap Calibration of the Empirical Likelihood*

*Step 1* - Generate $B$ resamples $\mathrm{z}^{(b)}$, randomly with replacement, from the original sample $\mathrm{z} = \{z_1, \ldots, z_n\}$.

*Step 2* - For each bootstrap sample $\mathrm{z}^{(b)}$, calculate the EL at a point $\widehat{\mu}$, using algorithm $4.1$ of $\S 4.7$ with some minor changes. In the Step 1 of this algorithm, the EL is now evaluated at the sample mean shape $\widehat{\mu}$. Thus $(4.29)$ is used to calculate $L^{(b)}$ for the resample $\mathrm{z}^{(b)}$ as $L^{(b)} = L_{EL}^{(b)}(\widehat{\mu})$. The values of $L^{(b)}$ are stored in a $B \times 1$ vector $L_B$.

*Step 3* - Let $l_\alpha^B$ be the bootstrap version of $l_\alpha$. $l_\alpha^B$ can be calculated from the ordered values

$$L_B[1] \leq L_B[2], \ldots, L_B[B-1] \leq L_B[B].$$

For instance, if $B = 100$ and the nominal level of the confidence region is $\alpha = 0.10$, then $l_\alpha^B = L_B[90]$.

*Step 4* - The empirical likelihood region with bootstrap calibration is given by

$$\mathrm{R}_\alpha^B = \{m : W_{EL}(m) \leq l_\alpha^B\}. \tag{4.31}$$

## 4.9 Monte Carlo Simulation Study

A simulation experiment was performed to examine the coverage accuracy of EL confidence regions. Two types of EL confidence regions were compared: those obtained using the limiting $\chi^2$ distribution under the null hypothesis, and those obtained using the bootstrap calibration.

As discussed in the previous section, bootstrap calibration uses the bootstrap to calculate a percentile which is then used to determine an empirical likelihood confidence region.

Using the notation of §2.6.1, the Monte Carlo simulation is performed generating $n_M$ Monte Carlo samples and $B$ bootstrap samples for each Monte Carlo sample. The output of this experiment is $n_M$ confidence regions obtained using the two methods above.

Let $\widehat{\mu}_i$ be the sample mean shape of the $i$th Monte Carlo sample. Also let $C_{Tab}$ and $C_{BC}$ denote the estimated coverage probability of the confidence regions $(4.30)$ and $(4.31)$, defined by

$$\hat{C}_{Tab}(EL) = \#\{i : W_{EL}(\widehat{\mu}_i) \leqslant l_\alpha, i = 1, \ldots, n_M\}/n_M,$$

where $W_{EL}(\widehat{\mu}_i)$ is $W_{EL}(\mu)$ for the $i$th Monte Carlo sample, and $l_\alpha$ is obtained from $\chi^2_{2k-4}$ tables, and

$$\hat{C}_{BC}(EL_B) = \#\{i : L_{EL}(\widehat{\mu}_i) \leqslant l^B_{(\alpha,i)}, i = 1, \ldots, n_M\}/n_M,$$

where $l^B_{(\alpha,i)}$ is $l^B_\alpha$ for the $i$th Monte Carlo sample, obtained by bootstrap calibration.

## 4.10 Simulation Results

In Table 4.1 the results of a simulation experiment involving the complex Bingham distribution are presented. This experiment is similar to the one described in §2.9 but in this case the "tabular" EL and EL with bootstrap calibration methods are used. The Hotelling and Goodall tests were also considered. The number of Monte Carlo samples is 1000 in each cell of the table. For each Monte Carlo sample, 200 bootstrap samples were used. The nominal coverage of the confidence region is $0.90$. The results show that for highly concentrated distributions the estimated coverage probabilities of the 4 methods are close to the nominal value $0.90$. On the other hand, for distributions with low concentration about the mean shape and sample size 100, the estimated coverage probabilities of the Hoteling and Goodall methods are very far from $0.90$, while the estimated coverage probability of the EL (tabular) and EL (bootstrap), is still very close to $0.90$. For the low concentrated distributions, if the sample size is 30, the estimated coverage probability of the EL method is far from $0.90$. Thus the bootstrap calibration improves the EL method in this case. Generally, when the sample size is small, say 30, the $\chi^2$ distribution is not a good approximation for the distribution of the EL.

Another experiment, using the complex Bingham distribution, is presented in Table 4.2. This table considers the same statistics as in Table 4.1. The number of Monte Carlo and bootstrap samples are 1000 and 200, respectively. The sample size is 30. The nominal coverage of the coverage accuracy is $0.90$. The first values of the parameters of the complex Bingham distribution define very highly concentrated distributions. The last values of the parameters define very low concentrated distributions. For highly concentrated distributions, the observed

| Parameters - Eigenvalues of the Complex Bingham | Sample Size | EL $(4.25)$ | EL Bootstrap $(Algorithm\ 4.2)$ | Hotelling Test $(2.12)$ | Goodall Test $(1.25)$ |
|---|---|---|---|---|---|
| 0 0 800 | 30 | 0.794 | 0.885 | 0.900 | 0.890 |
| | 50 | 0.856 | 0.892 | 0.900 | 0.892 |
| | 100 | 0.893 | 0.901 | 0.906 | 0.901 |
| 0 50 850 | 30 | 0.795 | 0.884 | 0.893 | 0.828 |
| | 50 | 0.856 | 0.893 | 0.893 | 0.823 |
| | 100 | 0.893 | 0.904 | 0.899 | 0.858 |
| 0 0 1 | 30 | 0.840 | 0.890 | 0.023 | 0.015 |
| | 50 | 0.887 | 0.904 | 0.013 | 0.008 |
| | 100 | 0.888 | 0.900 | 0.008 | 0.011 |
| 0 1 2 | 30 | 0.845 | 0.891 | 0.057 | 0.049 |
| | 50 | 0.887 | 0.909 | 0.036 | 0.032 |
| | 100 | 0.903 | 0.908 | 0.020 | 0.024 |

Table 4.1: *Coverage probabilities of the tabular EL, EL with bootstrap calibration, Hotelling and Goodall confidence regions for the mean shape. An algorithm to generate a complex Bingham was used in 4 special cases: eigenvalues 0, 0 and 800, which is a highly concentrated complex Watson distribution; eigenvalues 0, 450 and 800, which represents a highly concentrated Bingham distribution; eigenvalues 0, 0 and 1, which is a low concentrated complex Watson distribution and eigenvalues 0, 1 and 2, which is a low concentrated complex Watson distribution.*

coverage accuracy of EL (Bootstrap), Hotelling and Goodall are similar, but EL (Tabular) is less accurate. For very low concentrated distributions, the observed coverage probabilities of the Goodall and Hotelling tests are very far from the nominal value $0.90$, while EL (Bootstrap) retains accuracy very well.

| Parameters - Eigenvalues of the Complex Bingham | EL $(4.25)$ | EL(Bootstrap) $(Algorithm\ 4.2)$ | Hotelling Test $(2.12)$ | Goodall Test $(1.25)$ |
|---|---|---|---|---|
| 0 0 200 | 0.795 | 0.885 | 0.897 | 0.882 |
| 0 0 30 | 0.800 | 0.886 | 0.866 | 0.856 |
| 0 0 25 | 0.803 | 0.887 | 0.851 | 0.850 |
| 0 0 20 | 0.802 | 0.886 | 0.837 | 0.845 |
| 0 0 15 | 0.803 | 0.890 | 0.817 | 0.810 |
| 0 0 10 | 0.813 | 0.889 | 0.772 | 0.742 |
| 0 0 8 | 0.820 | 0.891 | 0.72 | 0.696 |
| 0 0 7 | 0.822 | 0.892 | 0.669 | 0.649 |
| 0 0 5 | 0.833 | 0.900 | 0.533 | 0.485 |
| 0 0 4 | 0.859 | 0.915 | 0.433 | 0.378 |
| 0 0 3 | 0.866 | 0.899 | 0.290 | 0.247 |
| 0 0 2 | 0.839 | 0.890 | 0.125 | 0.097 |
| 0 0 1 | 0.840 | 0.890 | 0.023 | 0.015 |

Table 4.2: *Coverage probabilities for the Confidence Region for the Mean Shape for the sample size 30. The parameters of the complex Watson distribution varies from a very highly concentrated case* $(0,\ 0,\ 200)$ *to a very low concentrated case* $(0,\ 0,\ 1)$. *The nominal value for the coverage probability is 0.90*

## 4.11 Graphical Representation of the EL's Asymptotic Distribution

Theory suggests that the histogram of the log EL ratio should have a shape similar to that of a $\chi^2$ distribution due to Wilks theorem. Since the histogram of realizations of a random variable gives an approximate graphical representation of its pdf, the histogram of the log EL ratio should be similar to the density of a $\chi^2$ distribution with $2k-4$ degrees of freedom.

In Figure 4.1, a graphical representation of the EL indicates that its asymptotic distribution broadly agrees with the theoretical considerations. The EL variable is obtained from 400 samples of size 100 from a highly concentrated complex Watson distribution with parameters 0, 0 and 800. The histogram of the EL of Figure 4.1 has a shape broadly similar to that of a $\chi^2$ distribution with $2k-4=4$ degrees of freedom.

In Figure 4.2, a graphical representation of EL, calculated for bootstrap samples, indicates that this statistic has also an appropriate asymptotic distribution. The bootstrap samples were obtained according to the following scheme. A Monte Carlo sample of size 100 was generated from a complex Watson distribution with parameters 0, 0 and 800, the same parameters of the previous simulation experiment. For this Monte Carlo Sample, 400 bootstrap samples were selected. The histogram of the EL for the 400 bootstrap samples suggests that the distribution of this statistic is roughly $\chi^2$ with $4$ degrees of freedom.

## 4.12 Analysing Real Data

The empirical likelihood method is applied to the neural spines of T2 mouse vertebra. This data set was considered in §2.8.2. The number of bootstrap samples was 200.

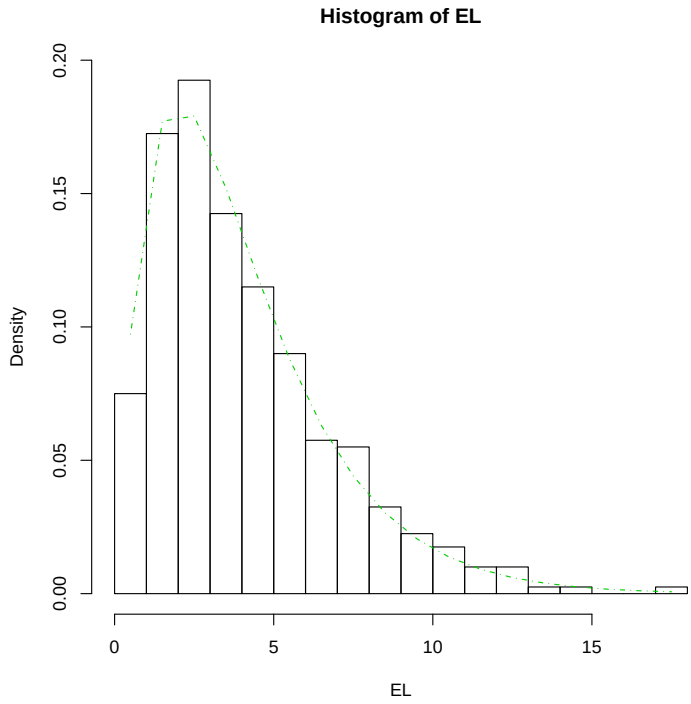


Figure 4.1: Histogram of the EL. The EL is calculated for 400 samples from a very highly concentrated complex Watson distribution with parameters (0,0,800). The line is the density of the chi-square with 4 degrees of freedom.

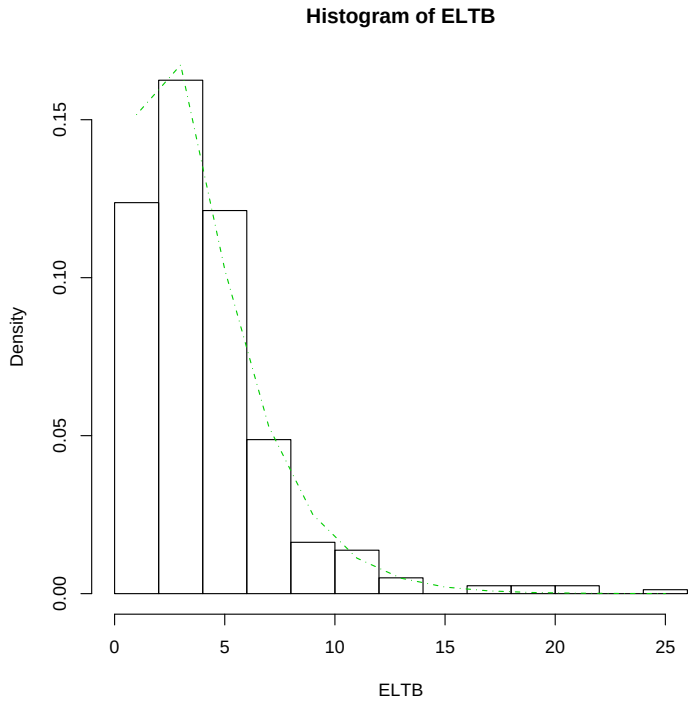


Figure 4.2: Histogram of the EL for the 400 bootstrap samples. The bootstrap samples were obtained from a Monte Carlo sample of a very highly concentrated complex Watson distribution with parameters (0,0,800). The line represents the density of the corresponding chi-square.

On the left side of Figure $4.3$ the mean shapes of the bootstrap samples are shown. Those correspond to the EL bootstrap samples that are inside the bootstrap calibrated convex hull. Thus this set of points is a good representation of the EL bootstrap calibrated confidence region on the landmark space.

On the right side of Figure $4.3$ the NA confidence regions, which were defined in $(1.21)$, are shown.

From the simulation results, it was seen that EL (Bootstrap) confidence regions have better coverage probability than NA confidence regions. On the other hand, Figure $4.3$ shows that EL bootstrap calibrated confidence regions are bigger than the NA confidence regions, which were defined in $(1.21)$. Since the data set considered has low concentration, this real example illustrates that this difference should be noted, and EL (Bootstrap) methods are more appropriate for low concentrated data sets than EL (Tabular), Goodall and Hotelling methods.

## 4.13 Empirical Likelihood Tests for Several Samples

Contrary to the bootstrap methods, EL confidence regions and hypothesis tests are very closely related. Bootstrap confidence regions and hypothesis tests are treated separately in the literature. However, EL was originally developed to be a nonparametric version of the Wilks's theorem (see comments above Theorem $4.2$), and the EL ratio is used for both confidence regions and hypothesis tests. Once a confidence region has been calculated, hypothesis tests can be derived naturally.

The case of hypothesis tests for several samples will be considered. The situation is that of the one way analysis of variance (ANOVA) (see Owen, 2001, pp. 87-90).

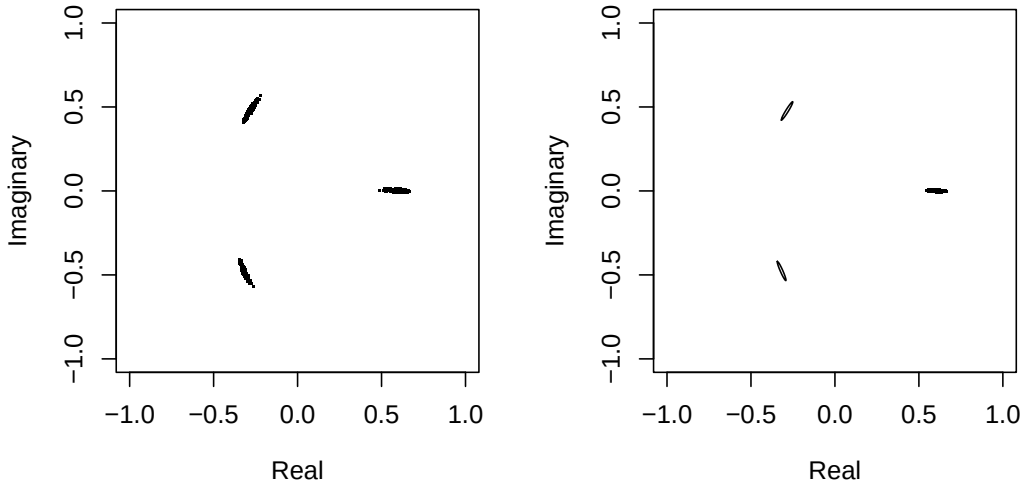


Figure 4.3: The graph on the left has the bootstrap EL means shape inside the bootstrap calibrated EL Confidence Region. A rule using bootstrap calibration is defined to decide if the mean shape of a bootstrap sample is inside or outside the confidence region. This rule is used to choose the samples that appear on the graph. The graph on the right presents the normal approximated confidence regions. Those regions are built by using the principal components in the tangent space. Those principal components are projected back to the landmark space to deliver this graphical representation.

Consider $p$ groups $u^{[1]}, \ldots, u^{[p]}$. Suppose $u^{[j]} = \{u_{ij} : i = 1, \ldots, n_j\}$, where $u$ has distribution $F(\nu^{[j]}, \psi^{[j]})$, where $\nu^{[j]}$ is a unknown location parameter and $\psi^{[j]}$ is an unknown scale parameter. Thus the groups can have different dispersion structure. In the experimental design literature, each group corresponds to the levels of a factor.

Consider the following hypotheses

$$H_0 : \nu = \nu^{[1]} = \ldots = \nu^{[p]} \text{ versus } H_1 : \nu^{[1]}, \nu^{[2]}, \ldots, \nu^{[p]} \text{ unrestricted.}$$

The anova statistic is given by

$$F = \frac{\frac{1}{p-1} \sum_{j=1}^{p} n_j (\bar{u}_{.j} - \bar{u}_{..})^2}{\frac{1}{n-p} \sum_{j=1}^{p} \sum_{i=1}^{n_j} (u_{ij} - \bar{u}_{.j})^2},$$

where $\bar{u}_{.j} = \frac{1}{n_j} \sum_{i=1}^{n_j} u_{ij}$ and $\bar{u}_{..} = \sum_{j=1}^{p} \sum_{i=1}^{n_j} u_{ij}$. If the variances $v^{[j]}$ are equal and the observations $u_{ij}$ are normally distributed, the statistic $F$ has an $F_{p-1,n-p}$ distribution. EL provides an interesting nonparametric alternative to the classical one-way anova. It does not need the assumptions that the observations are normally distributed and that the variances of the different groups are the same. So the EL method can be applied in other cases where the normality assumption is not suitable.

## 4.14 Empirical Likelihood Hypothesis Tests in Shape Analysis

This section introduces EL methods to test hypotheses in shape analysis. The approach we describe is a natural extension of the EL method for building confidence regions. We focus on $p$-sample problems where there is interest in testing for a common mean shape in each of $p$ populations.

Consider $y^{[j]} \equiv \{Y_{ij}, 1 \leq i \leq n_j\}$ as a random sample of configurations from $p$ populations of objects $\Pi^{[j]}$, where $1 \leq j \leq p$. Let $z^{[j]} = \{z_{ij} : i = 1, \ldots, n_j\}$ be the pre-shapes of $y^{[j]}$. Let the matrices $S^{[1]}, \ldots, S^{[p]}$ be the product matrices of the groups $1, \ldots, p$, given by

$$S^{[j]}(w) = \sum_{i=1}^{n_j} w_{ij} z_{ij} z_{ij}^{\star}, \quad \sum_{i=1}^{n_j} w_{ij} = 1.$$

We now define the EL ratio in the case of several samples.

Even though $EL^{[j]}$ is similar to $(4.25)$, this function can be more precisely defined. Again the constraint is that $m$ is an eigenvector of the matrix $S^{[j]}(w)$ corresponding to the largest eigenvalue, where $w_i$ are non-negative weights to be determined.

The constraints are given by

$$\mathbf{Re}\{m_l^{\star[j]} S^{[j]}(w) m\} = 0, \ \mathbf{Im}\{m_l^{\star[j]} S^{[j]}(w) m\} = 0, \ l = 1, \ldots, k-2, \ \ j = 1, \ldots, p.$$

Define

$$\gamma_l(w) = m_l^{\star} S^{[j]}(w) m \tag{4.32}$$

and

$$\delta_{il}^{[j]} = m_l^{\star} z_{ij} z_{ij}^{\star} m = \frac{\partial \gamma_l^{[j]}(w)}{\partial w_i}. \tag{4.33}$$

Using the definitions $(4.32)$ and $(4.33)$, the profile empirical likelihood ratio function for the mean shape is given by

$$EL^{[j]}(m) = \max_{w_{ij} \geq 0} \left\{ \prod_{i=1}^{n_j} n w_{ij} \;\; \middle| \;\; \sum_{i=1}^{n_j} w_{ij} = 1 \text{ and } C2 = 0 \right\}, \tag{4.34}$$

where $C2 = \sum_{l=1}^{k-2} \{ \lambda_l^{(R)[j]} \mathbf{Re}(\gamma_l^{[j]}(w)) + \lambda_l^{(I)[j]} \mathbf{Im}(\gamma_l^{[j]}(w)) \}$.

An EL method is presented to test the hypothesis

$$H_0 : m^{[1]} = m^{[2]} =, \ldots, m^{[p]} = \mu \text{ versus } H_1 : m^{[1]}, m^{[2]}, \ldots, m^{[p]} \text{ unrestricted} \tag{4.35}$$

The main computational challenge is to maximize the EL under $H_0$. In other words, to maximize the function

$$\prod_{j=1}^{p} EL^{[j]}(m) \tag{4.36}$$

over $m$, where $EL^{[j]}(m)$ is defined in $(4.34)$. A numerical procedure from the computer program $R$ was used to calculate $(4.36)$. This procedure is called $BFGS$. The $BFGS$ procedure is a quasi-newton method that finds the optimum value for a parameter vector of a given function. The details about this procedure are given by Nocedal and Wright (1999).

The following theorem parallels Theorem 3.1.

**Theorem 4.4.** *Considers the hypotheses* $(4.35)$, *in the case where there are* $k$ *landmarks. Then provided that each population satisfies the conditions of Theorem 2.1, the test statistic*

$$2log[\max_{H_1} EL / \max_{H_0} EL]$$

*has an asymptotic* $\chi^2_{(p-1)(2k-4)}$ *distribution under* $H_0$.

| Bingham eigenvalues | n | EL (Tabular) | EL (Bootstrap) | Hotelling Test $(1.32)$ | Goodall Test $(1.27)$ |
|---|---|---|---|---|---|
| 0, 6, 8 | 30 | 0.06 | 0.05 | 0.18 | 0.33 |
| | 50 | 0.11 | 0.08 | 0.12 | 0.25 |
| | 100 | 0.15 | 0.14 | 0.13 | 0.29 |
| 0, 4, 6 | 30 | 0.10 | 0.08 | 0.21 | 0.44 |
| | 50 | 0.14 | 0.13 | 0.19 | 0.42 |
| | 100 | 0.12 | 0.11 | 0.14 | 0.42 |
| 0, 2, 4 | 30 | 0.10 | 0.06 | 0.18 | 0.69 |
| | 50 | 0.08 | 0.04 | 0.18 | 0.72 |
| | 100 | 0.08 | 0.07 | 0.18 | 0.75 |
| 0, 1, 2 | 30 | 0.04 | 0.03 | 0.25 | 0.92 |
| | 50 | 0.06 | 0.04 | 0.20 | 0.95 |
| | 100 | 0.09 | 0.04 | 0.22 | 0.95 |

Table 4.3: *Observed significance level of the tests for populations with low concentration and heterogeneous variance structure. This experiment is similar to the one of Table 3.1 but the number of Monte Carlo samples is only 100. The vector of eigenvalues of the first and second populations are* $(0,\ 1,\ 2)$ *and* $(0,\ 15,\ 30)$, *respectively. The nominal significance level is* $0.05$.

## 4.15 Simulation Experiment

This section presents a simulation study to compare the EL test with Hotelling and Goodall two sample tests. The computation of the EL method is very computationally intensive. For example, we estimated that, for 1000 Monte Carlo samples and 200 bootstrap resamples for each Monte Carlo sample, our program would take at least 10 months to finish. Thus this simulation experiment was done with 100 Monte Carlo samples and 200 bootstrap samples for each Monte Carlo sample.

## 4.16 A Real-data Example

The EL method for hypothesis tests is applied to the schizophrenic dataset (see Dryden and Mardia, 1998, p. 11). This example has 14 schizophrenic and 14 normal patients. The number of landmarks placed in each object is 13.

The EL method cannot be applied in this example if the total number of landmarks is considered. The algorithm cannot find the estimates of the parameters for this case. Thus to apply the EL method in this example, only four landmarks are considered. The labels of those landmarks are 1, 2, 4 and 13. The observed significance level of the tests based on EL (Tabular), EL (Bootstrap), Goodall and Hotelling tests were 0.4633, 0.7462, 0.0002 and 0.0147, respectively. The performance of the EL methods differs from that of the Goodall and Hotelling tests. This example is very challenging for EL method since the sample size is small.

# Chapter 5

# Conclusions and Directions for Further Research

The aim of this thesis was to show how to apply computer intensive methods such as bootstrap and empirical likelihood methods in statistical shape analysis.

The final conclusions about using bootstrap and empirical likelihood methods in shape analysis are presented in this chapter. Also, some suggestions for further work are given. The chapter is organized as follows: §5.1 gives a comparison of the two approaches, considering methodological and numerical aspects. Some comments about directions for further research are given in §5.2 .

## 5.1 Comparing the Two Methods

Since two distinct approachs are considered, the reader might wonder which one is the most appropriate for statistical shape analysis. The conclusion is not simply that one is definitely better than the other. It depends on the objectives of the reader and also the computational resource available since the computing time is a very relevant point. In addition to the aims of

the reader, one should bear in mind that there is a huge diference between developing a new method and using a method which already exists in a particular problem.

The structure of this chapter is as follows. The positive and negative points of each method are summarized first; and then the two methods are compared. In these comparisons we attempt to clarify the diferent perspectives between someone who is developing and someone who is just using the method. The simulation results from the previous chapters are also used to compare the two methods.

### 5.1.1 Bootstrap Methods

Bootstrap methods are often easy to implement once a suitable statistic has been identified. The user who wishes to apply the bootstrap method of this thesis for a real dataset just needs to implement the steps of the bootstrap Algorithm 2.2, in the one sample case, or Algorithm 3.1, in the multisample case.

However, if it is necessary to develop a new bootstrap method, the derivation of the theoretical basis can be very hard work; see the proofs given in §2.7 and §3.4. The difficulties are more pronounced when it is necessary to find an asymptotically pivotal statistic. The proof that a statistic is asymptotically pivotal can be a very laborious task.

### 5.1.2 Empirical Likelihood Methods

The computational effort with empirical likelihood can be very intensive. A numerical optimization procedure is one of the steps of the EL Algorithm 4.1, for example; and this step can be very intensive. Also, when the EL method is used with bootstrap calibration, which is the case in Algorithm 4.2, the numerical optimization step is done for each bootstrap sample which

involves a substantial computational effort. This is particularly noticable in the case of hypothesis tests. For example, the processing time of the program which applies bootstrap calibration using the function $(4.36)$, is about $10$ months if the number of Monte Carlo samples is 1000, the number of bootstrap resamples is 200 and the sample size is 100.

For someone who wants to develop a new nonparametric method for a particular problem, empirical likelihood seems to be attractive since it uses a statistic that is automatically pivotal under very mild conditions, an advantage not shared by the bootstrap. However, EL needs bootstrap calibration if good coverage accuracy of confidence regions is to be achieved. This involves a big computational effort.

Owen's algorithm (see appendix D) makes it easier to implement EL methods in some circunstances, including the shape context considered here. Since this algorithm is numerically very stable it helps researchers in the field of empirical likelihood.

### 5.1.3 Simulation Results

In this section some numerical comparisons between bootstrap and EL methods are presented. The tables of this section are obtained from combining columns from tables in previous chapters. At this stage is not necessary to compare the bootstrap and empirical likelihood methods to Hotelling and Goodall tests since these comparisons were already done previously.

The coverage probabilities of the EL and bootstrap confidence regions are displayed in Table 5.1, which is obtained from Table 2.2 and Table 4.1. Thus all the conditions of the experiment are the same as in those tables: 1000 Monte Carlo samples were used and 200 bootstrap resamples were draw from each sample. The third and fourth columns are from the

| Eigenvalues of the Complex Bingham | Sample Size | EL $(4.25)$ | EL Bootstrap $Algorithm\ 4.2$ | Pivotal Bootstrap (2.15) | Hotelling's $T^2$ Bootstrap (2.17) |
|---|---|---|---|---|---|
| 0 0 800 | 30 | 0.794 | 0.885 | 0.899 | 0.909 |
| | 50 | 0.856 | 0.892 | 0.898 | 0.894 |
| | 100 | 0.893 | 0.901 | 0.903 | 0.901 |
| 0 50 850 | 30 | 0.795 | 0.884 | 0.899 | 0.909 |
| | 50 | 0.856 | 0.893 | 0.898 | 0.894 |
| | 100 | 0.893 | 0.904 | 0.903 | 0.901 |
| 0 0 1 | 30 | 0.840 | 0.890 | 0.822 | 0.719 |
| | 50 | 0.887 | 0.904 | 0.864 | 0.745 |
| | 100 | 0.888 | 0.900 | 0.871 | 0.823 |
| 0 1 2 | 30 | 0.845 | 0.891 | 0.863 | 0.769 |
| | 50 | 0.887 | 0.909 | 0.870 | 0.811 |
| | 100 | 0.903 | 0.908 | 0.891 | 0.857 |

Table 5.1: *Coverage probabilities for the Confidence Region for the Mean Shape of the EL and bootstrap methods. Four different special cases of the complex Bingham distribution are considered. The third and fourth columns are from the Table 4.1 and the last two columns are from Table 2.2. The results here are based on 1000 Monte Carlo samples and 200 bootstrap resamples for each Monte Carlo sample.*

Table 4.1 and the last two columns are from Table 2.2. Since the nominal level is $0.90$, the EL with bootstrap calibration is the most accurate method and the EL (Tabular) is the least accurate. Also, the asymptotically pivotal bootstrap is more accurate than the Hotelling's $T^2$ bootstrap. For example, when the parameters of the complex Bingham are 0, 1 and 2 and the sample size is 30, the coverage probability of the EL with bootstrap calibration is 0.891 and this is closer 0.90 than the 3 other methods.

More coverage probabilities of the EL and bootstrap confidence regions are displayed in

| Eigenvalues of the Complex Bingham | EL $(4.25)$ | EL Bootstrap $Algorithm\ 4.2$ | Pivotal Bootstrap (2.15) | Hotelling's $T^2$ Bootstrap (2.17) | Modified $T$ Test (2.12) |
|---|---|---|---|---|---|
| 0 0 200 | 0.795 | 0.885 | 0.899 | 0.909 | 0.857 |
| 0 0 30 | 0.800 | 0.886 | 0.901 | 0.904 | 0.857 |
| 0 0 25 | 0.803 | 0.887 | 0.902 | 0.903 | 0.858 |
| 0 0 20 | 0.802 | 0.886 | 0.903 | 0.903 | 0.859 |
| 0 0 15 | 0.803 | 0.890 | 0.897 | 0.899 | 0.859 |
| 0 0 10 | 0.813 | 0.889 | 0.901 | 0.893 | 0.860 |
| 0 0 8 | 0.820 | 0.891 | 0.898 | 0.882 | 0.857 |
| 0 0 7 | 0.822 | 0.892 | 0.901 | 0.888 | 0.857 |
| 0 0 5 | 0.833 | 0.900 | 0.897 | 0.891 | 0.846 |
| 0 0 4 | 0.859 | 0.915 | 0.901 | 0.896 | 0.854 |
| 0 0 3 | 0.866 | 0.899 | 0.897 | 0.879 | 0.844 |
| 0 0 2 | 0.839 | 0.890 | 0.880 | 0.831 | 0.782 |
| 0 0 1 | 0.840 | 0.890 | 0.821 | 0.719 | 0.672 |

Table 5.2: *Coverage probabilities for the Confidence Region for the Mean Shape for the sample size 30 of the EL and bootstrap methods. In this case, 1000 Monte Carlo samples and 200 bootstrap samples are generated from the complex Watson distribution. The second and the third columns are from Table 4.2 and the last three columns are from Table 2.3.*

| Eigenvalues of the Complex Bingham | $n$ | Bootstrap $(3.11)$ | EL $(4.25)$ | EL bootstrap $Algorithm\ 4.2$ |
|---|---|---|---|---|
| I:0, 1, 2 | 30 | 0.071 | 0.04 | 0.03 |
| II:0, 15, 30 | 50 | 0.066 | 0.06 | 0.04 |
| | 100 | 0.057 | 0.09 | 0.04 |
| I:0, 2, 4 | 30 | 0.063 | 0.10 | 0.06 |
| II:0, 30, 60 | 50 | 0.052 | 0.08 | 0.04 |
| | 100 | 0.037 | 0.08 | 0.07 |
| I:0, 4, 6 | 30 | 0.045 | 0.10 | 0.08 |
| II:0, 60, 90 | 50 | 0.049 | 0.14 | 0.13 |
| | 100 | 0.051 | 0.12 | 0.11 |
| I:0, 6, 8 | 30 | 0.041 | 0.06 | 0.05 |
| II:0, 90, 120 | 50 | 0.057 | 0.11 | 0.08 |
| | 100 | 0.053 | 0.15 | 0.14 |

Table 5.3: *Observed significance level of the tests for populations with low concentration and heterogeneous variance structure. The vector of eigenvalues of the first and second populations are* $(0,\ 1,\ 2)$ *and* $(0,\ 15,\ 30)$*, respectively. The nominal significance level is* $0.05$*. The first column of results come from Table 3.1 and the last two columns are from Table 4.3.*

Table 5.2. The conditions of the experiment are the same as in Tables 4.2 and 2.3: 1000 Monte Carlo samples are generated from complex Watson distributions, where those distribution varies from low concentration, with eigenvalues 0, 0 and 1, to a high concentration with eigenvalues 0, 0 and 200. The number of bootstrap resamples per Monte Carlo sample is 200. The results show that the estimated coverage probability of the EL method with bootstrap calibration is the closest to the nominal value 0.90, specially for low concentrated distributions.

The observed significance levels of the bootstrap test and the empirical likelihood tests

are displayed in Table 5.3. One should bear in mind that the bootstrap results are based on 1000 Monte Carlo samples and 200 bootstrap resamples; and the EL results are based on 100 Monte Carlo samples and 200 bootstrap resamples, and therefore the results are not directly comparable. However, the table at least shows that EL and EL bootstrap methods give similar results to the asymptotically pivotal bootstrap method. For example, when the eigenvalues of the complex Bingham distribution are 0, 1 and 2, which is a low concentration case, the observed significance values of the EL (Tabular) and EL with bootstrap calibration are very close to the nominal value 0.05, especially for the sample sizes 30 and 50. It shows the EL (Tabular) and EL with bootstrap calibration tests are very competitive in relation to the asymptotically pivotal bootstrap.

## 5.2 Further Work

This section presents several possible directions for future work in statistical shape analysis.

One direction is to use other methods for the problems considered in this thesis, e.g., building confidence regions and testing hypotheses. A second direction is to use computer intensive methods, like the bootstrap and empirical likelihood, in other problems of shape analysis. Some details about both directions will be given.

### 5.2.1 A Bayesian Method

Only classical computer intensive methods have been used in this thesis. It would be of interest to develop Bayesian methods for tackling the problem of comparing the mean shapes of several groups of objects. A Bayesian approach to problems in shape analysis is given by Dryden and

Mardia (1998, p. 149), who consider the case that a random sample of pre-shapes $z_1, \ldots, z_n$ has a complex Watson distribution with mode $\mu$ and known concentration parameter $\kappa$. They use a complex Bingham distribution for the prior $f(\mu)$. In mathematical terms,

$$\begin{aligned} f(\mu|z_1, \ldots, z_n) &\propto f(z_1, \ldots, z_n|\mu) f(\mu) \\ &\propto \exp\left(\kappa \sum_{i=1}^{n} z_i^\star \mu \mu^\star z_i\right) \exp\left(\mu^\star A \mu\right) \\ &\propto \exp\left(\kappa \sum_{i=1}^{n} z_i^\star \mu \mu^\star z_i + \mu^\star A \mu\right) \\ &\propto \exp\{\mu^\star (kS + A)\mu\}, \end{aligned}$$

where $S = \sum_{i=1}^{n} z_i z_i^\star$ is the product matrix.

So the posterior distribution is also a complex Bingham and since the prior and posterior are in the same family of distributions the prior is called conjugate.

This Bayesian model is restrictive since the Bingham distribution is assumed for the prior and the Watson distribution is assumed for the data. A possible research topic would be to consider other models for shape datasets. Since these possibilities are analytically very complex, it would be necessary to use a computer intensive method called Markov chain Monte Carlo (MCMC) to implement the Bayesian approach.

### 5.2.2 Size-and-Shape

In §1.3, the definitions of complex configuration, Helmertized configuration, pre-shape and shape were given. Recalling from that section that shape is the remaining information when location, scale and rotation are removed, it is possible to consider another way of doing shape analysis. In this way the information about scale is retained. This type of analysis is called size-

and-shape analysis. Mathematically, a size-and-shape study is performed using the pre-shapes $(1.4)$ with the scale information retained. This is given by

$$w = Hz^0,$$

which is defined in $(1.3)$.

The bootstrap and empirical likelihood methods of this thesis can be applied in a size-and-shape study. Figure 5.1 shows a bootstrap confidence region obtained by applying the Algorithm 2.2 to the Helmertized configurations $w_1, \ldots, w_n$ of the dataset of example 2.1, which is T2 mouse vertebra. These numerical results look reasonable and illustrate the feasibility of applying the methods of this thesis to the analysis of size-and-shape.

### 5.2.3 Shape Variation

Shape variation is studied by using principal components on the tangent space. This topic was seen in §1.6. The study of shape variation uses the sample covariance matrix on the tangent space $S_v$ which was given in $(1.18)$.

The idea for studying shape variation is to apply the principal components method to the matrix $S_v$ and then to project the two first principal components to the landmark space. Thus the shape variation is represented by

$$\widehat{\mu} + c\sqrt{\phi_1}u_1 \text{ and } \widehat{\mu} + c\sqrt{\phi_2}u_2,$$

which were defined in $(1.21)$.

A topic for further research is to use the principal components from bootstrap resamples and

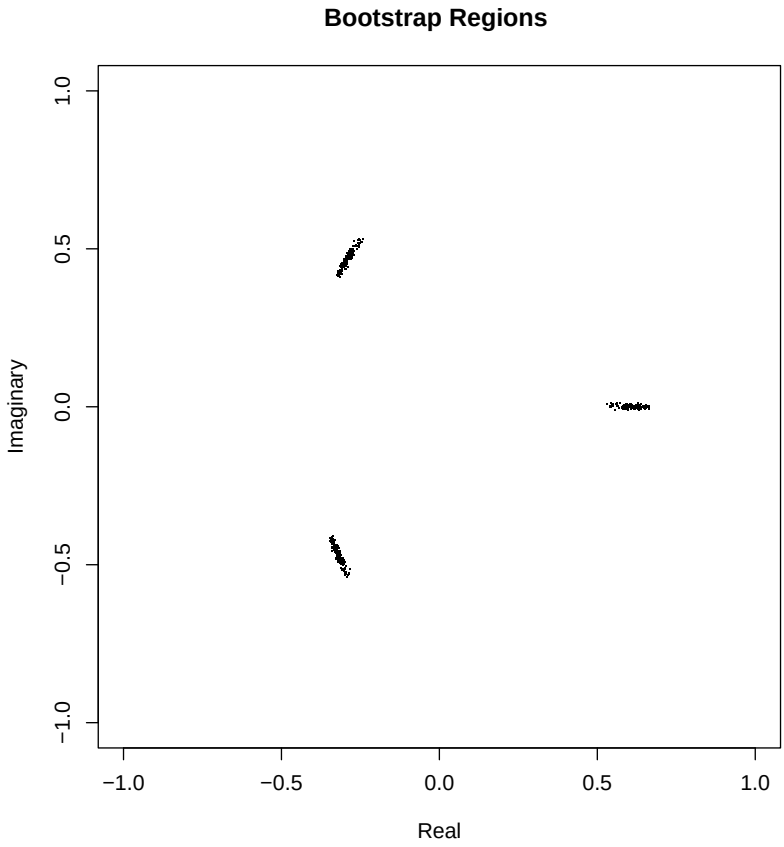


Figure 5.1: Bootstrap Confidence Regions for a Size-and-Shape Case

to try to find ways of improving the coverage of bootstrap confidence regions in the landmark space.

# Appendix A

# Matrix Results

Some basic results which are used in the thesis will be reviewed in this appendix.

Consider $L : V \rightarrow W$, where $L$ is a linear transformation and $V$ and $W$ are two complex vectors spaces of dimensions $d_v$ and $d_w$. The kernel of $L$ is defined by

$$ker(L) = \{v \in V : Lv = 0\}. \tag{A.1}$$

Also consider that the orthogonal complement of the subspace $V$ is defined as

$$V^{\perp} = \{u : u^{\star}v = 0\}. \tag{A.2}$$

The spectral decomposition theorem for complex Hermitian matrices is now stated. It plays a very important role in what follows. Some other basic properties of complex numbers, matrices and vectors are needed as well (see e.g. Fraleigh and Beauregard, 1995, pp. 454-486).

Let $c = a + bi$, where $a$ and $b$ are real numbers and $i = \sqrt{-1}$. The number $\bar{c} = a - bi$ is said to be the complex conjugate of $c$.

If $C = [C_{jk}]$ is a $p \times p$ complex matrix, the conjugate transpose of $C$ is given by $C^{\star} = [\bar{c}_{kj}]$.

A complex matrix $C$ is said to be Hermitian if it is equal to its conjugate transpose, i.e., $C^\star = C$. It should be noted that the eigevalues of a Hermitian matrix are real even though the eigenvectors are complex vectors (see Kent, 1994).

**Theorem A.1.** *(Spectral Decomposition theorem for Hemitian Matrices)*

*Let $C$ be a $p \times p$ Hermitian matrix. Then we may write*

$$C = \sum_{j=1}^{s} \xi_j P_j,$$

*where $\xi_1 < \ldots < \xi_s$ are the distinct eigenvalues of $C$ and $\xi_j$ has multiplicity $r_j$, where $\sum_{j=1}^{s} r_j = p$; and the $P_j(p \times p)$ are Hemitian projective matrices ($P_j^\star = P_j$ and $P_j^2 = P_j$).*

Another useful concept is the generalized inverse of a real symmetric matrix $R$ $(a \times a)$. If the symmetric matrix $R$ has rank $p \leq a$, the Moore-Penrose generalized inverse of $R$ is given by

$$R^+ = \sum_{j=1}^{p} \kappa_j^{-1} \gamma_j \gamma_j^T, \tag{A.3}$$

where the $\kappa_j' s$ are the non-zero eigenvalues of $R$ and the $\gamma_j$'s are their corresponding eigevectors (see Dryden and Mardia, 1998, p. 152).

Minimizing the quadratic form $a^\star C a$, where $C$ is a Hermitian matrix and $a$ is a complex unit vector, is a relevant topic for the Chapter 3, when bootstrap hypothesis tests are considered. It is also relevant for Procrustes fit. This result for real symmetric matrices is given by Mardia et. al. (1979, p. 479) and for Hermitian matrices see Mirsky (1955, p. 388).

*L***emma A.1.** *Let $C$ be a $(p \times p)$ Hermitian matrix with eigenvalues $\epsilon_1 \leqslant \epsilon_2 \ldots \leqslant \epsilon_p$. Then* $\min_{a:a^\star a=1} a^\star C a = \epsilon_1$.

**Proof.**

From Theorem A.1 we may write $C = TET^\star$, where $T = [\tau_1, \ldots, \tau_p]$, $E = diag\,[\epsilon_1, \ldots, \epsilon_p]$, and $\tau_1, \ldots, \tau_p$ are unit eigenvectors of $C$. Then for any complex unit vector $a$,

$$\begin{aligned} a^\star Ca &= a^\star TET^\star a \\ &= y^\star Ey \\ &= \sum_{i=1}^{p} \epsilon_i |y_i|^2, \end{aligned}$$

where $y = (y_1, \ldots, y_p)^T = T^\star a$.

Thus to minimize $a^\star Ca$ consider

$$\begin{aligned} a^\star Ca &= \sum_{i=1}^{p} \epsilon_i |y_i|^2 \\ &\geq \epsilon_1 \sum_{i=1}^{p} |y_i|^2 \\ &\geq \epsilon_1, \end{aligned}$$

since $\sum_{i=1}^{p} |y_i|^2 = 1$.

Thus the minimum is attained when $a = \tau_1$.

Power series of matrices and convergent matrix sequences are also relevant topics; see Mirsky (1955) for further background. A power series for a complex square matrix $A$ is defined by

$$\sum_{m=0}^{\infty} c_m A^m,$$

where $c_m$ is a scalar and by definition $A^0 = I_p$, the identity matrix. A matrix power series $\sum_{m=0}^{\infty} c_m A^m$ is said to be absolutely convergent if

$$\sum_{m=0}^{\infty} c_m ||A||^m < \infty$$

where $||.||$ is a suitable matrix norm, or distance.

On the space of $p \times p$ complex matrices the Euclidean matrix distance is defined by

$$||A|| = \{tr(A^\star A)\}^{1/2},$$

where $tr(.)$ denotes the trace of a matrix. If $\{A_m\}_{m\geq 1}$ is a sequence of complex matrices, we say that $A_m \to A$ if $||A_m - A|| \to 0$ as $m \to \infty$.

For any square complex matrix $A$, we define the exponential $\exp(A)$ by

$$\exp A = \sum_{s=0}^{\infty} \frac{1}{s!} A^s.$$

Note that $\exp A$ is convergent for any matrix $A$.

# Appendix B

# Order Notation

Suppose that $a_n$ and $b_n$ are sequences of real numbers. The notation

$$a_n = O(b_n) \tag{B.1}$$

means that

$$\limsup_{n\to\infty} \frac{|a_n|}{|b_n|} < \infty.$$

For example, if $a_n = \mu + n\sigma^2$, where $\mu$ and $\sigma^2$ are constants, then $a_n = O(n)$ since

$$\limsup_{n\to\infty} \frac{|\mu + n\sigma^2|}{|n|} = \sigma^2.$$

In this thesis, only the order notation $O(.)$ for sequences of real variables will be used. The order notation is used to represent the accuracy of confidence regions and hypothesis tests.

# Appendix C

# The Factor 2 in $(2.12)$

The claim that a 2 is required in $(2.12)$ will follow from $(1.14)$ if it can be shown that

$$(z-\mu)^{\star}\Sigma^{-1}(z-\mu) = ((x-\mu_1)^T,(y-\mu_2)^T)\begin{pmatrix}\Sigma_1 & -\Sigma_2\\ \Sigma_2 & \Sigma_1\end{pmatrix}^{-1}\begin{pmatrix}x-\mu_1\\ y-\mu_2\end{pmatrix},$$

where $\Sigma = \Sigma_1 + i\Sigma_2,\ \Sigma_1^T = \Sigma_1,\ \Sigma_2^T = -\Sigma_2$ and $\mu = \mu_1 + i\mu_2$.

Write $\Sigma^{-1} = \Sigma^1 + i\Sigma^2$. Then the identity $(\Sigma_1 + i\Sigma_2)(\Sigma^1 + i\Sigma^2) = I_{k-1}$

implies that

$$\Sigma_1\Sigma^1 - \Sigma_2\Sigma^2 = I_{k-1} \text{ and } \Sigma_1\Sigma^2 + \Sigma_2\Sigma^1 = O_{k-1}. \tag{C.1}$$

Moreover, (C.1) implies that

$$\begin{pmatrix}\Sigma_1 & -\Sigma_2\\ \Sigma_2 & \Sigma_1\end{pmatrix}^{-1} = \begin{pmatrix}\Sigma^1 & -\Sigma^2\\ \Sigma^2 & \Sigma^1\end{pmatrix}.$$

Because

$$\begin{pmatrix} \Sigma_1 & -\Sigma_2 \\ \Sigma_2 & \Sigma_1 \end{pmatrix}\begin{pmatrix} \Sigma^1 & -\Sigma^2 \\ \Sigma^2 & \Sigma^1 \end{pmatrix} = \begin{pmatrix} \Sigma_1\Sigma^1 - \Sigma_2\Sigma^2 & -\Sigma_1\Sigma^2 - \Sigma_2\Sigma_1 \\ -\Sigma_2\Sigma^1 + \Sigma_1\Sigma^2 & \Sigma_1\Sigma^1 - \Sigma_2\Sigma^2 \end{pmatrix}$$

which implies that

$$\begin{pmatrix}\Sigma_1 - \Sigma_2\\ \Sigma_2 & \Sigma_1\end{pmatrix}\begin{pmatrix}\Sigma^1 - \Sigma^2\\ \Sigma^2 & \Sigma^1\end{pmatrix} \quad = \quad I_{2k-2}.$$

Therefore,

$$\begin{aligned}
&(z-\mu)^\star\Sigma^{-1}(z-\mu)\\
&= (x-\mu_1-i(y-\mu_2))^T(\Sigma^1+i\Sigma^2)(x-\mu_1+i(y-\mu_2))\\
&= (x-\mu_1)^T\Sigma^1(x-\mu_1)+(y-\mu_2)^T\Sigma^1(y-\mu_2)\\
&\qquad -(x-\mu_1)^T\Sigma^2(y-\mu_2)+(y-\mu_2)^T\Sigma^2(x-\mu_1)\\
&= ((x-\mu_1)^T(y-\mu_2)^T)\begin{pmatrix}\Sigma^1 & -\Sigma^2\\ \Sigma^2 & \Sigma^1\end{pmatrix}\begin{pmatrix}x-\mu_1\\ y-\mu_2\end{pmatrix}\\
&= ((x-\mu_1)^T(y-\mu_2)^T)\begin{pmatrix}\Sigma_1 & -\Sigma_2\\ \Sigma_2 & \Sigma_1\end{pmatrix}^{-1}\begin{pmatrix}x-\mu_1\\ y-\mu_2\end{pmatrix}
\end{aligned}$$

as required.

# Appendix D

# Owen's Empirical Likelihood Program for a Vector Mean

The EL ratio for a vector mean is defined as

$$R(\mu) = \max_{w_i \geq 0} \left\{ \prod_{i=1}^{n} (nw_i) \mid \sum_{i=1}^{n} w_i u_i = \nu \text{ and } \sum_{i=1}^{n} w_i = 1 \right\}. \tag{D.1}$$

Adopting a Lagrange multiplier approach, we consider

$$G = \sum_{i=1}^{n} \log(nw_i) - n\lambda^T \left( \sum_{i=1}^{n} w_i (u_i - \nu) \right) + \gamma \left( \sum_{i=1}^{n} w_i - 1 \right),$$

where $\lambda \in \mathbb{R}^d$ are the multipliers to be determined. The steps to find the maximum of $G$ are the same as the scalar case, see §4.3. We find that $\gamma = -n$ and $\lambda \in \mathbb{R}^d$. The weights are estimated as

$$w_i = \frac{1}{n} \frac{1}{1 + \lambda^T (u_i - \mu)}.$$

Also as in the univariate case, replacing $w_i$ in the first constraint

$$\sum_{i=1}^{n} w_i(u_i - \mu) = 0_d,$$

gives

$$\frac{1}{n}\sum_{i=1}^{n} \frac{u_i - \mu}{1 + \lambda^T(u_i - \mu)} = 0_d. \tag{D.2}$$

So $\lambda$ must satisfy the $d$ equalities of $(D.2)$.

There is another way to solve this problem using convex duality. Convex duality in this context results in a maximization over $n$ variables with $d + 1$ constraints becoming a minimization over $d$ variables. The number of variables would be $d + 1$ but the multiplier $\gamma$ is already known.

The convex dual of $(D.2)$ is given by

$$\log R(\mu) = \log \prod_{i=1}^{n} n w_i = -\sum_{i=1}^{n} \log(1 + \lambda^T(u_i - \mu)) \equiv L(\lambda).$$

The system $(D.2)$ is equivalent to

$$\nabla L(\lambda) = 0,$$

where $\nabla$ is the gradient of $L(.)$.

It should be noted that $L(\lambda)$ has $n$ inequality constraints because the cases of $w_i \leq 0$ should be excluded. So those $n$ inequality constraints are

$$1 + \lambda^T(u_i - \mu) \geq 0, \;\; i = 1, \ldots, n. \tag{D.3}$$

Thus the maximization problem, defined by $(D.1)$, is equivalent to minimizing $L(\lambda)$. This is the convex duality formulation for this particular problem.

It is possible to discard the constraints $(D.3)$. It is done by defining a pseudo-logarithm function. This function, when used in $L(\lambda)$, delivers

$$L_\star = -\sum_{i=1}^{n} \log_\star(1 + \lambda^T(u_i - \mu)).$$

So to minimize this new function is not necessary to impose any constraint.

The algorithm of Owen uses the formulation above to find the empirical likelihood for a parameter vector. The explanation above was to clarify the main points of the algorithm. The technical details can be found at the website

http://www-stat.stanford.edu/owen/empirical/el.S.

# Bibliography

Beran, R. J. (1987). Prepivoting to Reduce Level Error of Confidence Sets. *Biometrika* **74**:457-468.

Beran, R. J. (1988). Prepivoting Test Statistics: a Bootstrap View of Asymptotic Refinements. *Journal of the American Statistical Association* **83**:687-697.

Bhattacharya, R. and Pantragenaru, V. (2003). Large Sample Theory of Intrinsic and Extrinsic Sample Means on Manifolds. Tentatively accepted by *Annals of Statistics*.

Bickel, P. J. and Freedman, D. A. (1981). Some Asymptotic Theory for the Bootstrap, *The Annals of Statistics*, **9**:1196-1217.

Bookstein, F. L.(1986). Size and Shape Spaces for Landmark Data in Two Dimensions (with discussion), *Statistical Science*, **1**:181-242.

Bookstein, F. L. (1994). Can Biometrical Shape be a Homologous Character? In Homology:the Hierarchical Basis of Comparative Biology, pages 197-227. Academic Press, New York.

Bratley, P., Fox, B. L. and Schrage, L. E. (1983). *A Guide to Simulation*, Springer-Verlag, New York.

Casella, G. and Berger, R. L. (1990). *Statistical Inference*, Duxbury Press, Belmont, California.

Chernick, M. C. (1999). *Bootstrap Methods*, John Wiley and Sons, New York.

Davidson, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Applications*, Cambridge University Press, Cambridge.

DiCiccio, T. J., Hall, P. and Romano, J. P. (1991). Empirical Likelihood is Bartlett-Correctable, *The Annals of Statistics*, **19**, 1053–1061.

Dryden, I. L. and Mardia, K. V. (1998). *Statistical Shape Analysis*, Wiley and Sons, Chichester.

Ducharme, G. R., Jhun, M., Romano, J. P. and Truong, K. (1985). Bootstrap Confidence Cones for Directional Data, *Biometrika*, **72**, 637–645.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife., *The Annals of Statistics* **7**, 1-26.

Er, F. (1998). *Robust Methods in Statistical Shape Analysis*, PhD dissertation, Department of Statistics, University of Leeds.

Fisher, N. I. and Hall, P. (1989). Bootstrap Confidence Regions for Directional Data, *Journal of the American Statistical Association*, **84**, 996–1002.

Fisher, N. I. and Hall, P. (1990). On bootstrap Hypothesis Testing, *Australian Journal Statistics*, **32**, 177-190.

Fisher, N. I., Hall, P., Jing, B. and Wood, A. T. A. (1996). Improved Pivotal Methods for Constructing Confidence Regions with Directional Data, *Journal of the American Statistical Association*, **91**, 1062–1069.

Fraleigh, J. B. and Beauregard, R. A. (1995). *Linear Algebra*, Wiley, New York.

Goodall, C. R. (1991). Procrustes Methods in the Statistical Analysis of Shape (with discussion), *Journal of the Royal Statistical Society, Series B*, **53**, 285–339.

Grenander, U. (1981). *Abstract inference*, Wiley, New York.

Hall, P. (1986). On the Bootstrap and Confidence Intervals, *Journal of the American Statistical Association*, **91**, 1062–1069.

Hall, P. (1988a). Theoretical Comparison of Bootstrap Confidence Intervals, *The Annals of Statistics*, **16**, 927-953.

Hall, P. (1988b). On the Bootstrap and Symmetric Confidence Intervals, *Journal Royal Statistical Society, series B* **50**, 35-45.

Hall, P. (1990). On bootstrap Hypothesis Testing, *Australia Journal Statistics*, B **32**, 177-190.

Hall, P. (1992). *The bootstrap and Edgeworth expansion*, Springer-Verlag, New York.

Hall, P. and LaScala, B. (1991). Methodology and Algorithms of Empirical Likelihood, *International Statistical Review*, **58**, 109-127.

Hall, P. and Wilson, S. R. (1991). Two Guidelines for Bootstrap Hypothesis Testing, *Biometrics*, **47**, 757-762.

Hartley, H. O. & Rao, J. N. K. (1968). A New Estimation Theory for Sample Surveys, *Biometrika*, **55**, 547–557.

Hinkley, D. V. (1988). Bootstrap Methods (with discussion). *Journal os the Royal Statistical Society, series B*, **50**, 321-337.

Jing, B.-Y. and Wood, A. T. A. (1996). Exponential Empirical Likelihood is not Bartlett Correctable. *The Annals of Statistics*, **24**, 365–369.

Kendall, D. G. (1977). The Diffusion of Shape, *Advances in Applied Probability*, **9**, 428–430.

Kendall, D. G. (1984). Shape Manifolds, Procrustean Metric and Complex Projective Spaces, *Bulletin of the London Mathematical Society*, **16**, 81–121.

Kendall, D. G., Barden, D., Carne, T. K. and Le, H. (1999). *Shape and Shape Theory*, John Wiley & Sons, London.

Kent, J. T. (1992). *New Directions in Shape Analysis*, The Art of Statistical Science, 115–127, Edited by K. V. Mardia, John Wiley & Sons Ltd, London.

Kent, J. T. (1994). The Complex Bingham Distribution and Shape Analysis, *Journal of the Royal Statistical Society, Series B* **56**, 285–299.

Kent, J. T. (1997). Data Analysis for Shapes and Images, *Journal of Statistical Inference and Planning* **57**, 181–193.

Le, H.-L. and Kendall, D. G. (1993). The Rimannian Structure of Euclidean Shape Spaces: a Novel Environment for Statistics., *Annals of Statistics*, **21**, 1225–1271.

Liu, R. Y. and Sing, K. (1987). On a Partial Correction by the Bootstrap, *Ann. Statist.* **15**, 1713–1718.

Mardia, K. V. and Dryden, I. L. (1999). The Complex Watson Distribution and Shape Analysis, *Journal of the Royal Statistical Society , Series B* **61**, 913-926.

Mardia, K. V. and Jupp, P. E. (2000). *Directional Statistics*, Academic Press, London.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivariate Analysis*, Academic Press, London.

Mardia, K. V. and Walder, A. N. (1994). Shape Analysis of Paired Landmark Data. *Biometrika*, **81**:185-196.

Mirsky, L. (1955). *An introduction to Linear Algebra.* Oxford University press, London.

Nocedal, J., Wright, S. J. (1999). *Numerical Optimization.* Springer.

Owen, A. B. (1988). Empirical Likelihood Ratio Confidence Intervals for a Single Functional, *Biometrika* **75**, 237-249.

Owen, A. B. (1990). Empirical Likelihood Ratio Confidence Regions, *The Annals of Statistics* **18**, 90-120.

Owen, A. B. (1991). Empirical Likelihood for Linear Models, *The Annals of Statistics* **19**, 1725-1747.

Owen, A. B. (2001). *Empirical Likelihood*, Chapman & Hall, New York.

Rao, C. R. (1972). *Linear Statistical Inference and Its Applications 2nd Edition*, Academic Press, London.

Rubin, D. B. (1981). The Bayesian bootstrap, *The Annals of Statistics*, **9**, 130-134.

Small, C. G. (1996). *The Statistical Theory of Shape*, Springer-Verlag, New York.

Sing, K. (1981). On the Asymptotic Accuracy of Efron's Bootstrap, *Annals of Statistics* **9**, 1187-1195.

Wilks, S. S. (1938). The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses, *The Annals of Mathematical Statistics* **9**, 60–2.