Springer Series
in Statistics

Christopher G. Small

The Statistical Theory of Shape

Springer

Springer Series in Statistics

Andersen/Borgan/Gill/Keiding: Statistical Models Based on Counting Processes.
Andrews/Herzberg: Data: A Collection of Problems from Many Fields for the Student and Research Worker.
Anscombe: Computing in Statistical Science through APL.
Berger: Statistical Decision Theory and Bayesian Analysis, 2nd edition.
Bolfarine/Zacks: Prediction Theory for Finite Populations.
Brémaud: Point Processes and Queues: Martingale Dynamics.
Brockwell/Davis: Time Series: Theory and Methods, 2nd edition.
Daley/Vere-Jones: An Introduction to the Theory of Point Processes.
Dzhaparidze: Parameter Estimation and Hypothesis Testing in Spectral Analysis of Stationary Time Series.
Fahrmeir/Tutz: Multivariate Statistical Modelling Based on Generalized Linear Models.
Farrell: Multivariate Calculation.
Federer: Statistical Design and Analysis for Intercropping Experiments.
Fienberg/Hoaglin/Kruskal/Tanur (Eds.): A Statistical Model: Frederick Mosteller's Contributions to Statistics, Science and Public Policy.
Fisher/Sen: The Collected Works of Wassily Hoeffding.
Good: Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses.
Goodman/Kruskal: Measures of Association for Cross Classifications.
Grandell: Aspects of Risk Theory.
Haberman: Advanced Statistics, Volume I: Description of Populations.
Hall: The Bootstrap and Edgeworth Expansion.
Härdle: Smoothing Techniques: With Implementation in S.
Hartigan: Bayes Theory.
Heyer: Theory of Statistical Experiments.
Huet/Bouvier/Gruet/Jolivet: Statistical Tools for Nonlinear Regression: A Practical Guide with S-PLUS Examples.
Jolliffe: Principal Component Analysis.
Kolen/Brennan: Test Equating: Methods and Practices.
Kotz/Johnson (Eds.): Breakthroughs in Statistics Volume I.
Kotz/Johnson (Eds.): Breakthroughs in Statistics Volume II.
Kres: Statistical Tables for Multivariate Analysis.
Le Cam: Asymptotic Methods in Statistical Decision Theory.
Le Cam/Yang: Asymptotics in Statistics: Some Basic Concepts.
Longford: Models for Uncertainty in Educational Testing.
Manoukian: Modern Concepts and Theorems of Mathematical Statistics.
Miller, Jr.: Simultaneous Statistical Inference, 2nd edition.
Mosteller/Wallace: Applied Bayesian and Classical Inference: The Case of *The Federalist Papers.*

(continued after index)

Christopher G. Small

The Statistical Theory of Shape

With 46 Illustrations

Springer

Christopher G. Small
Department of Statistics
and Actuarial Science
University of Waterloo
Waterloo, Ontario
Canada N2L 3G1
smallmcl@watserv1.uwaterloo.ca

Library of Congress Cataloging-in-Publication Data
Small, Christopher G.
The statistical theory of shape / Christopher G. Small
p. cm. — (Springer series in statistics)
Includes bibliographical references and index.
ISBN 0-387-94729-9 (hard : alk. paper)
1. Shape theory (Topology)—Statistical methods. I. Title.
II. Series.
QA612.7.S58 1996
514—dc20 96-13587

Printed on acid-free paper.

Production managed by Bill Imbornoni; manufacturing supervised by Jeffrey Taub.
Camera-ready copy created from the author's LaTex files.
Printed and bound by Braun-Brumfield, Inc., Ann Arbor, MI.
Printed in the United States of America.

9 8 7 6 5 4 3 2 1

ISBN 0-387-94729-9 Springer-Verlag New York Berlin Heidelberg SPIN 10524357

Preface

In general terms, the shape of an object, data set, or image can be defined as the total of all information that is invariant under translations, rotations, and isotropic rescalings. Thus two objects can be said to have the same shape if they are similar in the sense of Euclidean geometry. For example, all equilateral triangles have the same shape, and so do all cubes. In applications, bodies rarely have exactly the same shape within measurement error. In such cases the variation in shape can often be the subject of statistical analysis.

The last decade has seen a considerable growth in interest in the statistical theory of shape. This has been the result of a synthesis of a number of different areas and a recognition that there is considerable common ground among these areas in their study of shape variation. Despite this synthesis of disciplines, there are several different schools of statistical shape analysis. One of these, the Kendall school of shape analysis, uses a variety of mathematical tools from differential geometry and probability, and is the subject of this book. The book does not assume a particularly strong background by the reader in these subjects, and so a brief introduction is provided to each of these topics. Anyone who is unfamiliar with this material is advised to consult a more complete reference. As the literature on these subjects is vast, the introductory sections can be used as a brief guide to the literature.

A few comments should be made about the numbering of figures and propositions. Figures are numbered in order within chapters. Thus Figure 2.3 is the third figure to be found in Chapter 3. Propositions, lemmas, corollaries, and definitions are numbered consecutively within each section. Thus Proposition 2.6.3 is the third result (whether proposition, lemma,

etc.) within Section 2.6.

Chapter 1 is the basic introductory chapter for the rest of the book. Many of the ideas that are developed in greater detail later are touched upon briefly in this first chapter. Chapter 2 is essentially a review of some basic tools from differential geometry and groups of transformations of Euclidean space. The reader who is familiar with these methods can skim this material for the notation that will be used throughout the rest of the book, and proceed to the next chapter. Chapter 3, which describes various ways of representing shapes on manifolds, is pivotal for all later material, and leads into Chapters 4 and 5, where a stochastic theory is developed on the shape manifolds. Chapter 6 has a collection of applications that are rather loosely bound together by the theme of this book.

This book would not have been written without the support of a number of people. Thanks are due to Martin Gilchrist at Springer, who approached me about writing a book on shape. Thanks must go to John Kimmel, also of Springer, whose timely and supportive responses to all my questions made the job of writing much easier. To Springer's production staff and my copyeditor, David Kramer, I offer my sincere thanks.

Whenever I had a problem in computing I turned to my colleague Michael Lewis, whose assistance was invaluable. Some of the better-looking pictures in this book are there through his help. Thanks also go to David Kendall, for his inspiration and valued support over the years. I first began to work on shape theory when I started my Ph.D. under David Kendall's supervision in 1978. What is good in this book is largely due to him. What is bad is my responsibility alone! Thanks also to my colleagues Huiling Le and Colin Goodall for their excellent advice on the subject, and to Fred Bookstein for his insights and energy. Zejiang Yang was also very helpful in catching a number of errors in the manuscript.

I could not conclude this checklist of indebtedness without acknowledging the support of my wife Kristin Lord, who put up with the long hours I spent working on the manuscript. Kristin was also instrumental in bringing the Mt. Tom dinosaur data set to my attention.

Christopher G. Small
University of Waterloo
June 1996

Contents

1
Introduction

1.1 Background of Shape Theory

In 1977, David Kendall published a brief note [87] in which he introduced a new representation of shapes as elements of complex projective spaces. The result stated in the paper was unusual: under an appropriate random clock, the shape of a set of independent particles diffusing according to a Brownian motion law could be regarded as a Brownian motion on complex projective space. Many statisticians, who knew little about complex projective spaces and who did not work on diffusion processes, did not see immediate applications to their own work. However, in a sequence of talks at conferences around the world, David Kendall continued to expound on his theory, with some applications to problems in archeology. Presented with great clarity and with excellent graphics, these talks gradually generated wider interest. It was not until 1984 that the full details of the theory were published [90]. At that point it became clear that Kendall's theory of shape was of great elegance and contained some interesting areas of research for both the probabilist and the statistician.

The full range of possible applications became much clearer when David Kendall was invited to be a discussant for a paper by Fred Bookstein [19] in the journal *Statistical Science*. Kendall and Bookstein, it turned out, had been thinking along the same lines, namely that shapes could be represented on manifolds. There were some intriguing differences. Whereas Kendall represented the shapes of triangles in the plane as points on a sphere, a space of positive curvature, a suggestion of Bookstein represented

the shapes of those triangles as points on a Poincaré half plane, a space of negative curvature. Perhaps more important were the different applications each researcher emphasized. Kendall's applications were in the archeological and astronomical sciences, and studied the shapes of random sets of points, such as are to be found in a Poisson scattering. Bookstein's applications were in the biological and medical sciences, and drew on the tradition of researchers such as D'Arcy Wentworth Thompson, Julian Huxley, and later researchers in allometry and multivariate morphometrics. For Bookstein and his colleagues, the points under consideration were biologically active sites on organisms called landmarks.

At present, we can speak of both Kendall and Bookstein schools of shape analysis. It is within this context that this book is written. The primary theme of the book will be the representation of shapes on differential manifolds, and the statistical consequences of this idea. The emphasis will be more toward the Kendall school, where the differential geometry of shape analysis is more developed. However, we shall frequently compare this with some of the work of Bookstein and others, insofar as this is relevant to our goal.

In tracing the history of methods that have produced this statistical theory of shape, it is quickly apparent that a great variety of past work is responsible for its development. It is difficult to imagine a time in history when people have not been fascinated by shapes. Our visual fine arts, such as painting and sculpture, have appeal across cultures and illustrate the universality of shapes or forms.

As D'Arcy Thompson pointed out in his pioneering book *On Growth and Form* [172], there is an important relationship between the form or shape of a biological structure and its function. Thus the study of shape is also the study of function. For example, the mathematical constraint that a body have minimum surface area for a given volume requires that it be roughly spherical in shape. This result is known in mathematics as an isoperimetric inequality, and can be used to explain why an organism that seeks to minimize its boundary with an external environment, for heat conservation or defense, will often have a simple spherical curvature. On the other hand, if the boundary of an organism is required to be permeable to allow oxygen or food to flow across it, then such a minimization of surface area would be inappropriate. One would expect the surface area in this case to be roughly proportional to the volume of the organism. However, an organism cannot grow while maintaining the same shape and continuing to have a constant ratio between its volume and its external surface area. In this case, growth usually involves a change of shape, possibly through the introduction of a highly convoluted boundary. The geometric structure of lung tissue is a case in point. Recent developments in the theory of fractal shapes have shown that the boundary of a three-dimensional structure need not scale upwards as the square of its length or diameter. In fact, a highly convoluted surface can be thought of as an approximation to a fractal.

Sometimes the relationship between shape and function is of a more contrived nature. For example, the amphorae used in antiquity had a variety of forms. The particular shapes of amphorae were guides to the nature of the contents. This relationship continues down to the present day: nobody need confuse the contents of a bottle of white wine with the contents of a bottle of whiskey, as shape tells all.

Much of statistical theory has been dedicated to the estimation of location and scale parameters. As the statistical theory of shape is concerned with aspects of the data that remain after location and scale information are discounted, statistical shape concepts have not been as prominent as the theory of inference for location and scale. In 1934, R. A. Fisher [57] introduced the concept of the *configuration* of a univariate sample. This concept is equivalent to the formal definition of shape for dimension one that we shall develop in this book. In 1939, E. J. G. Pitman [134] developed the theory of minimum variance equivariant estimation of location and scale parameters, and in so doing illustrated the importance of conditioning on invariant statistics in the construction of best equivariant estimators for location and scale. The idea of a shape statistic as a maximal invariant under location and scale transformation can be seen in this work, although the shape statistics play an ancillary role to the estimation of parameters associated with location and scale transformations.

The extension of the concept of invariance to multivariate data is straightforward. However, it is in the psychometric literature that statistical tools were first developed for the comparison of the shapes of data sets. The roots of Procrustes analysis can be traced to Mosier [123], and then through the work of Sibson [150, 151] and Gower [75]. In comparing the differences in shape between two data sets, Procrustes analysis proceeds by transforming one data set to try to match the other. The transformations allowed in a standard analysis include shifts in location, scale changes, and rotations. Together, these transformations are called *similarity transformations* or *shape-preserving transformations*. When a transformation of one of the data sets has been found to most nearly match the other, the sum of squared differences of the coordinates between them is called the *Procrustean distance* between the two data sets. We shall see that this concept is closely related to the natural measure on distance between shapes that we shall consider in Section 1.3.

Another line of research that has contributed to the statistical theory of shape is to be found in the field of geometric probability and stochastic geometry. It is here that we see geometric objects, such as points, lines, and convex sets, as the basic data for the statistician. The set of outcomes of a random experiment can often be represented as a region in space whose volume, or p-dimensional content, can be ascertained. Within this region is to be found some subset E corresponding to an event. According to one definition, the probability $\mathcal{P}(E)$ of this event is the ratio of the p-dimensional content of this subset to that of the entire region. Such a

definition is problematic for certain applications, and leads to paradoxes such as that of Bertrand involving random lines. For this reason, the modern theory of geometric probability makes use of invariance of probabilities under Euclidean motions as a more fundamental notion for calculating the probability of geometric events. That is, a probability measure can be said to be geometric if it assigns equal probability that a random geometric object such as a point or line will hit congruent sets.

In 1980, David Kendall and his son Wilfrid Kendall [95] proposed the use of techniques from geometric probability to examine the hypothetical alignments of megalithic stones from Land's End in Cornwall. This data set of fifty-two sites at Land's End was originally investigated by Alfred Watkins [177], who advanced the theory that megalithic cultures had deliberately placed standing stones along straight lines known as *ley lines*. The folklore around the existence and interpretation of such lines is quite extensive despite the patchy evidence for the existence of ley lines. Kendall and Kendall [95] followed the approach of Simon Broadbent [33] by calculating the expected number of approximately collinear triplets of points if the megalithic sites had been positioned at random. As a triangle can be called approximately flat (or ϵ-blunt in their terminology) if its maximum internal angle is within tolerance ϵ of a straight angle, Kendall and Kendall were naturally drawn to the examination of the distribution of angles in a random triangle, and thereby to the concept of an induced marginal distribution on a space of triangle shapes. The paper by David Kendall [90] in 1984 was seminal for the development of the geometry and distribution theory of shape space. A key result of this paper was that the induced distribution of shape for a set of independent identically distributed bivariate normal points is uniform on the shape space when the covariance matrix is a multiple of the identity. The univariate version of this result also holds, although it is of much older vintage than the bivariate result. The work of Dryden and Mardia [53, 116, 117] generalized this work to the shapes of points from bivariate normal distributions having different means, and set the stage for the distribution theory to be tied in to the work on shape analysis developed in allometry, to which we now turn.

1.2 Principles of Allometry

Allometry can be defined as the study of the relationship between size and shape. If we take a set of measurements of distances between points on a body, then a *size variable* can be regarded as a summary of the overall scale of these measurements. For example, the arithmetic mean and the geometric mean of a set of distances are both size variables. Size variables are required to be *homogeneous* functions of the set of distances. This means that if all measurements are increased or decreased by a common scale

factor, then the size variable is itself increased or decreased by that same factor. If we standardize the distances by scaling them to have unit size variable, then the resulting ratios of dimensions are called *shape variables*.

Many of the key insights into the growth allometry of biological organisms were first outlined by Julian Huxley [85]. Allometry studies shape differences by taking ratios of dimensions of objects. As much of statistics is linear in nature, it is natural to take logarithms of the dimensions of objects and plot these logarithmic coordinates on a graph. Now, two objects of different size but common shape will have their dimensions in the same ratio. Therefore the shape statistics can be associated with differences between the logarithmic dimensions. For example, suppose we consider how an organism changes shape as it matures and grows with age. Let x_t and y_t be two recorded dimensions of the organism at age t, so that y_t/x_t is a partial description of the shape of the organism. Now, if all parts of the organism grow at a constant rate α as it matures, then growth will be exponential in nature, and we will have the formulas

$$x_t = x_0 \exp(\alpha t) \qquad y_t = y_0 \exp(\alpha t) \tag{1.1}$$

Thus

$$\log(y_t) - \log(x_t) \;=\; \log(y_0/x_0) \tag{1.2}$$

which does not involve the age t of the organism. So the logarithmic coordinates $(\log\, x_{t_j},\ \log\, y_{t_j})$, when plotted at different ages t_j, will all lie on a line of slope $+1$, which corresponds to constancy of shape. On the other hand, if these coordinates do not all lie on a line of slope $+1$, then we can deduce that there is some variation in shape between different ages. However, if x_t grows at a constant rate α and y_t grows at rate $\beta \neq \alpha$, then these logarithmic coordinates plotted at different ages will still lie on a straight line. In this case, the slope of the line will differ from unity.

This fact, namely that the logarithms of size variables lie on straight lines, is one of the basic empirical principles of allometry. This empirical principle has a theoretical foundation in a model that presupposes exponential growth at varying rates in different parts of an organism. In turn, this variation in the growth rate explains some of the variation in the shape of an organism as it matures.

It should be noted that the size variables need not be linear in nature in order that their logarithms lie on straight lines. We can extend from comparing distances of bodies or organisms to more general size variables such as surface area or volume, and we will still keep a linear functional relationship between their logarithmic coordinates if growth is exponential. The effect of using an area, say, rather than a length for y_t is to scale the slope by a factor of two in the plot of $\log(x_{t_j})$ and $\log(y_{t_j})$.

The analysis is seen to be statistical in nature when we reflect on the fact that measurement error and a slight unevenness of growth are to be expected under normal circumstances. Therefore, even when the model

assumptions are correct, we would not expect the points to lie on a perfect straight line. Statistical tools such as principal components analysis can be used to draw a line through the data. This is equivalent to fitting a bivariate normal distribution to the scatter plot of points $(\log x_{t_j}, \log y_{t_j})$ and finding the principal axis through the elliptical contours of the normal density.

At first sight, the extension from two size variables x_t and y_t to several would seem to be easy. While the linear statistical analysis of multivariate data through principal components is straightforward, the extension is problematic because the assumption of multivariate normality is quite stringent. In typical data sets, the set of size variables such as lengths have complicated nonlinear relationships among them. For example, if we were to record a set of 21 interpoint distances between 7 points on a two-dimensional image, we would only have 11 degrees of freedom among the 21 distances. The particular restrictions on these variables would be complicated and nonlinear, and would make modeling of their logarithms using normal assumptions difficult. It is at this point that the techniques of Procrustes analysis provide an avenue of escape from these difficulties. The problems that arise in taking ratios of size variables point us toward nonlinear mathematics and towards a theory of shape based upon configurations of points rather than ratios of size variables. This theory of shape will be the central topic of the book.

1.3 Defining and Comparing Shapes

When all information in a data set about its location, scale, and orientation is removed, the information that remains is called the *shape* of the data. Alternatively, we can say that two data sets have the same shape if a combination of a rigid motion and rescaling of one of the data sets will make it coincide with the other. In geometry, two figures that have the same shape are said to be similar. For example, two triangles will be similar provided their corresponding internal angles are equal.

To investigate the concept of shape more carefully, consider Figure 1.1, which shows three examples of side views of Iron Age brooches from a cemetery excavated at modern-day Münsingen, in Switzerland. As these brooches can be ordered chronologically from the layout of the cemetery, it is natural to consider how the shapes of the brooches developed over time. These three brooches represent only a fraction of the total data from the cemetery but will serve the purpose here of illustrating some basic principles of shape analysis.

Let us suppose for the moment that we are given these pictures as our primary data. How can we analyze the differences in shapes of the three brooches? A first step in such an analysis might be to construct a finite-

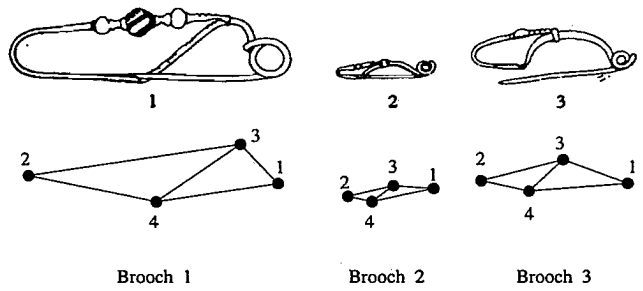

FIGURE 1.1. *Three Iron Age brooches. From each of the images, four landmarks are chosen at locations coinciding with features in the brooches. The landmarks correspond in a natural sense, so that landmarks in different images marking corresponding features are labeled in a similar fashion. The shape analysis proceeds by eliminating information in each of the configurations about location, scale and orientation. The brooches are adapted from Hodson, Sneath, and Doran,* Biometrika *53 (1966), p. 315, by kind permission of* Biometrika Trustees.

dimensional representation of some of the important geometric information from each picture. For example, we could construct a set of points $x_1, x_2, ..., x_n$ lying on each figure such that the locations of these points coincide with important features. On different bodies or figures, sites used for summarizing or encoding of geometric information are called *landmarks*. For example, on the human face, the positions of the eyes and other features can be used as landmarks to analyze the shape of the face. For our purposes, landmarks will be defined as points chosen from an image or object *to mark the location of important features and to give a partial geometric description of the image or object.*

Normally, we think of the features of a two-dimensional image as lying in a very high-dimensional space, or, in an idealized sense, in an infinite-dimensional space. If we keep this in mind, then we recognize that there is inevitably some loss of information in encoding pictures with a relatively small number of landmarks. Nevertheless, small numbers of landmarks can provide the basis for comparisons of important shape differences. Just as a small number of landmarks within a city might help us find our way around by identifying features of the city, so the landmarks chosen to summarize a figure can be regarded as identifying its important geometric features.

Let us consider how a set of four landmarks can be constructed for each of the three images. The centers of the coiled springs on the right of each figure represent corresponding points, and similarly, the leftmost points at which the curvature is sharpest also correspond. For each brooch, let x_1 and x_2 be these two points respectively. Additionally, let x_3 be the upper point on each brooch where the left piece bends back and fastens to form a

loop. Finally, we can choose the fourth landmark x_4 to be the lower bend on the loop. (This is the point of high curvature in the loop where the pin is secured.) This locates four landmarks for each figure.

More generally, n landmarks can be chosen so that the vector $(x_1, ..., x_n)$, which lies in $(\mathbf{R}^2)^n$, provides a $2n$-dimensional summary of some of the major geometric characteristics of the brooch, including location, orientation, scale, and shape information. To perform a shape analysis on these landmarks, we must determine the class of all functions of the vector $(x_1, ..., x_n)$ that measure its shape. This involves the elimination of information in $(x_1, ..., x_n)$ that describes the location, scale, or orientation of the landmarks. The location and scale statistics of a set of points are perhaps best known to statisticians because they can be described by standard statistical tools. For example, the location of a data set $(x_1, ..., x_n)$ can be described by its *sample mean*, or *centroid*, given by

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j \tag{1.3}$$

In addition, the size or scale of our configuration of landmarks can be described by a variety of statistics. Let us choose coordinates for each of the landmarks so that $x_j = (x_{j1}, x_{j2})$ for $j = 1, ..., n$, and $\bar{x} = (\bar{x}_1, \bar{x}_2)$. The column vectors of residuals about the means are

$$r_1 = \begin{pmatrix} x_{11} - \bar{x}_1 \\ x_{21} - \bar{x}_1 \\ \cdot \\ \cdot \\ \cdot \\ x_{n1} - \bar{x}_1 \end{pmatrix} \tag{1.4}$$

and

$$r_2 = \begin{pmatrix} x_{12} - \bar{x}_2 \\ x_{22} - \bar{x}_2 \\ \cdot \\ \cdot \\ \cdot \\ x_{n2} - \bar{x}_2 \end{pmatrix} \tag{1.5}$$

Then the matrix of squared residuals can be written as

$$\Gamma = \begin{pmatrix} r_1^T r_1 & r_1^T r_2 \\ r_2^T r_1 & r_2^T r_2 \end{pmatrix} \tag{1.6}$$

where $(\ .\ )^T$ denotes the transpose operation. The trace of Γ, given by

$$\mathrm{tr}(\Gamma) = r_1^T r_1 + r_2^T r_2 = \sum_{j=1}^{n} ||x_j - \bar{x}||^2 \tag{1.7}$$

is a natural measure of the size of the set of landmarks because it is independent of the orientation of the Cartesian coordinate system. The usual way to eliminate location and size information in data is by standardization, which is a combination of a location shift and a rescaling so that the data set has centroid $\bar{x}$ at the origin in $\mathbf{R}^2$ and the matrix Γ is standardized to have trace equal to one. For our example, the standardized data set becomes

$$\tau(x_1, ..., x_n) = \left(\frac{x_1 - \bar{x}}{\sqrt{\mathrm{tr}(\Gamma)}}, \ \cdots, \ \frac{x_n - \bar{x}}{\sqrt{\mathrm{tr}(\Gamma)}} \right) \tag{1.8}$$

A caveat must be mentioned here. In order for this representation to be meaningful, the landmarks $x_1, ..., x_n$ must not all be coincident. This presents no problem for our application to brooches. In general, a set of landmarks that are all coincident will be said to have indeterminate shape. Henceforth, we shall assume that this degeneracy does not arise. Note, however, that we do not exclude cases in which some but not all of the landmarks are coincident.

We shall refer to the vector τ defined in (1.8) as the *pre-shape* of the landmarks. While this terminology is not particularly descriptive, it does emphasize the order in which the reduction to shape progresses. The pre-shape τ lies in a constrained subset of the original Euclidean space $(\mathbf{R}^2)^n$. This subset can be represented by the intersection of the $(n-2)$-dimensional subspace

$$\mathbf{F}^{2n-2} = \{(x_1, ..., x_n) \in \mathbf{R}^{2n} : \sum_{j=1}^{n} x_j = 0\} \tag{1.9}$$

with the unit sphere

$$\mathbf{S}^{2n-1} = \{(x_1, ..., x_n) \in \mathbf{R}^{2n} : \sum_{j=1}^{n} ||x_j||^2 = 1\} \tag{1.10}$$

The intersection

$$\mathbf{S}_*^{2n-3} = \mathbf{F}^{2n-2} \cap \mathbf{S}^{2n-1} \tag{1.11}$$

is a $(2n-3)$-dimensional sphere within the ambient Euclidean space $\mathbf{R}^{2n}$. A subscripted star is included as a gentle reminder to ourselves that this $(2n-3)$-dimensional sphere is not the usual unit sphere embedded in $\mathbf{R}^{2n-2}$. We shall refer to this sphere as the *pre-shape space* or the *sphere of pre-shapes*.

At the next stage of our analysis, we must eliminate the information about the orientation of the data set, in order that the quantity which remains be a shape statistic. At first glance, the problem of defining and standardizing the orientation of the pre-shape of the data would seem to be similar to the problems of defining and standardizing the location and

scale. However, this is not the case. Some topological problems arise that cannot easily be removed.

By the orientation of a set of planar landmarks we intuitively understand the angle made by some axis through the landmarks with respect to some given axis, independent of the landmarks. For example, we could use the angle made by a ray from x_1 to x_2 as the description of the orientation of $(x_1, ..., x_n)$. While this will be quite satisfactory for the data that we are considering here, it will not suffice for orienting all configurations $(x_1, ..., x_n)$. Those sets of landmarks for which $x_1 = x_2$ cannot be oriented by such a definition. Of course, another definition can be used for these pre-shapes. However, we would obviously like to do better than this by finding a single definition that works for all samples.

Any angle can be represented as a point on $\mathbf{S}^1$, the unit circle about the origin in $\mathbf{R}^2$. So the orientation of $(x_1, ..., x_n)$ can be defined as a point $\Theta(x_1, ..., x_n) \in \mathbf{S}^1$. The process of standardizing the location and scale of $(x_1, ..., x_n)$ does not disturb its orientation. Therefore, we can also refer to the orientation $\Theta(\tau)$ of the pre-shape τ. It follows that the orientation of the pre-shape can be written as a function

$$\Theta : \mathbf{S}_*^{2n-3} \rightarrow \mathbf{S}^1 \tag{1.12}$$

from the sphere of pre-shapes into the unit circle of the plane. In addition, it is reasonable to suppose that an ideal orientation function would be a continuous function of its coordinates, so that Θ would be a continuous function on the sphere $\mathbf{S}_*^{2n-3}$. Now suppose that $\theta : \mathbf{R}^2 \rightarrow \mathbf{R}^2$ is a rotation of the plane about the origin. Under the rotation of the landmarks $x_j \rightarrow \theta(x_j)$, the corresponding pre-shapes transform as

$$\tau(x_1, x_2, ..., x_n) \rightarrow \tau[\theta(x_1), \theta(x_2), ..., \theta(x_n)] \tag{1.13}$$

This defines a mapping $\theta : \mathbf{S}_*^{2n-3} \rightarrow \mathbf{S}_*^{2n-3}$. Note that we abuse terminology slightly by using the symbol θ to refer to the rotation on $\mathbf{R}^2$ as well as the rotation on $\mathbf{S}_*^{2n-3}$. There is seen to be a simple correspondence between the two that makes the notation convenient. If Θ is an appropriate orientation function, then it should be compatible with the rotations of the plane, so that $\Theta[\theta(\tau)] = \theta[\Theta(\tau)]$ for any pre-shape $\tau \in \mathbf{S}_*^{2n-3}$. However, it is here that we get into trouble in attempting to define the function Θ. It can be shown that there does not exist a continuous function $\Theta : \mathbf{S}_*^{2n-3} \rightarrow \mathbf{S}^1$ that satisfies this property.

In order to see this, consider the following. The *orbit* of any pre-shape $\tau \in \mathbf{S}_*^{2n-3}$ will be the circle

$$\mathcal{O}(\tau) = \{\theta(\tau) : 0 \leq \theta < 2\pi\} \subset \mathbf{S}_*^{2n-3} \tag{1.14}$$

Therefore Θ, when restricted to the orbit $\mathcal{O}(\tau)$, would be a 1-1 correspondence between $\mathcal{O}(\tau)$ and $\mathbf{S}^1$. Let $\Theta^{-1} : \mathbf{S}^1 \rightarrow \mathcal{O}(\tau)$ be the inverse function. If Θ were continuous, then the function

$$\Theta^{-1}\Theta : \mathbf{S}_*^{2n-3} \rightarrow \mathcal{O}(\tau) \tag{1.15}$$

would be a retraction of $\mathbf{S}_*^{2n-3}$ onto the circle $\mathcal{O}(\tau)$. That is, $\Theta^{-1}\Theta$ would be a continuous function onto a subset of $\mathbf{S}_*^{2n-3}$ whose restriction to that subset would be the identity mapping. An argument in algebraic topology using fundamental homotopy groups, which we omit, shows that this is impossible. Thus we have the following:

Proposition 1.3.1. *For $n > 2$ there does not exist a continuous orientation function $\Theta : \mathbf{S}_*^{2n-3} \rightarrow \mathbf{S}^1$ that is compatible with rotations of the original coordinates $(x_1, ..., x_n)$ in the sense that $\Theta[\theta(\tau)] = \theta[\Theta(\tau)]$ for all $\tau \in \mathbf{S}_*^{2n-3}$.*

Proposition 1.3.1 tells us that continuous methods to standardize the orientation of pre-shapes will fail. That is, we cannot find a single definition that is continuous in the data and simultaneously orients all pre-shapes.

Our original purpose in standardizing the landmarks $(x_1, ..., x_n)$ with respect to location, scale, and orientation was to provide a set of coordinates for their shape. Proposition 1.3.1 does not exclude the possibility of our constructing coordinate systems that work for some shapes but not for others. In fact, we must distinguish between *representing shapes* and *constructing shape coordinates*. As we shall see in Chapter 3, shapes are naturally represented as points in a *shape manifold*. However, there will typically be no single coordinate system on that shape manifold that is non-degenerate and that provides coordinates for all points in the manifold. For example, on the Earth's surface, the coordinates of longitude and latitude work perfectly well except at the poles, where the longitude coordinate is redundant. Coordinates with latitude 90^oN and different longitudes refer to the same point, namely the north pole. The failure of a single coordinate system to work at all points on the sphere is simply a reflection of the fact that the sphere is not topologically equivalent to any subset of the plane.

Just as we do not identify the sphere with its coordinate system, so we should not identify shapes and shape representations with any particular coordinates used to construct them. As we shall see, the appropriate setting for representing shapes is as an *orbit space* Σ_2^n of a sphere $\mathbf{S}_*^{2n-3}$. By an orbit space of the sphere we mean a set Σ_2^n of equivalence classes, namely

$$\Sigma_2^n = \{\mathcal{O}(\tau) : \tau \in \mathbf{S}_*^{2n-3}\} \tag{1.16}$$

Two pre-shapes τ_1 and τ_2 will lie in the same equivalence class $\mathcal{O}(\tau)$ provided they have the same shape. If this is the case, there will exist a rotation θ such that $\theta(\tau_1) = \tau_2$.

However, this formal definition of Σ_2^n as a set of equivalence classes is of little value unless we can compare shapes and obtain some geometric intuition about Σ_2^n. To do this, we must define a *metric* on Σ_2^n. A metric is a mathematical generalization of the concept of Euclidean distance between points. Metrics have certain properties, which are listed in Problem 5 at the end of the chapter. If we think of Σ_2^n as a space, then its elements can be regarded as points in that space, for which we seek an appropriate definition of distance. An obvious way to do this is to use a metric between orbits on the pre-shape space $\mathbf{S}_*^{2n-3}$. As pre-shapes can be represented as points on this sphere, the distance between two pre-shapes is the geodesic, or great circle distance, between pre-shapes. On the earth, the great circle distance is the shortest distance one would have to travel to get from one place to another. This is quite easy to compute for spheres of any dimension. If τ_1 and τ_2 are two preshapes on $\mathbf{S}_*^{2n-3}$ then the great circle distance between τ_1 and τ_2 is given by

$$d(\tau_1, \tau_2) = \cos^{-1}(< \tau_1, \tau_2 >) \tag{1.17}$$

where $< \tau_1, \tau_2 >$ is the inner product between τ_1 and τ_2 as vectors in $\mathbf{R}^{2n}$. Note that the $\cos^{-1}$ function is defined so as to have range $[0, \pi]$. The induced metric on Σ_2^n is then defined as

$$d[\mathcal{O}(\tau_1), \mathcal{O}(\tau_2)] = \inf \{d[\theta_1(\tau_1), \theta_2(\tau_2)] : 0 \leq \theta_1, \theta_2 < 2\pi\} \tag{1.18}$$

where $\inf A$ is the infimum function over any set A of real numbers. In more informal language, we can say that the distance between two shapes, or orbits of $\mathbf{S}_*^{2n-3}$, is the minimum of the distances between all pairs of pre-shapes lying in the respective orbits. The reader should note that to perform the minimization, it is sufficient to fix θ_1 and minimize over all values of θ_2, or vice versa. Problem 5 at the end of the chapter asks the reader to show that formula (1.18) satisfies the properties of a metric. With this metric, the space Σ_2^n turns out be be a manifold. In fact, as we shall see in the next chapter, it is an example of a *complex projective space*. We shall leave the definition of these spaces to Section 2.3 and shall concentrate for the moment on calculating this metric on the shape space Σ_2^n.

To evaluate this metric on Σ_2^n we can make use of the algebraic properties of the complex plane. Suppose we consider the landmarks $x_1, \ldots, x_n$ to be elements of the complex plane by identification of the complex numbers $\mathbf{C}$ with $\mathbf{R}^2$. Then x_k can be regarded as a complex quantity by identifying the two coordinates of $x_k \in \mathbf{R}^2$ with the real and imaginary components of a complex number. Under this identification, the pre-shape coordinates

$$\tau_k = \frac{x_k - \bar{x}}{\sqrt{\sum_k |x_k - \bar{x}|^2}} \tag{1.19}$$

for $k = 1, 2, \ldots, n$ can also be regarded as complex quantities, being standardized versions of the original coordinates.

Let σ_1 and σ_2 be two shapes in Σ_2^n, and let us choose representative pre-shapes τ_1 and τ_2 so that $\sigma_j = \mathcal{O}(\tau_j)$ for $j = 1, 2$. Write

$$\tau_j = (\tau_{j1}, \tau_{j2}, \ldots, \tau_{jn}) \tag{1.20}$$

where τ_{jk} is the kth complex standardized coordinate of τ_j. Furthermore, let τ_{jk}^* be the complex conjugate of τ_{jk}. We will go into the details of the mathematics in Example 2.3.16 of the next chapter. For the moment, we shall note that the minimum in formula (1.18) can be found algebraically to be

$$d(\sigma_1, \sigma_2) = \cos^{-1}\left(|\sum_{k=1}^{n} \tau_{1k}\tau_{2k}^*|\right) \tag{1.21}$$

This is called the *Procrustean distance* or the *Procrustean metric* from σ_1 to σ_2. As the argument of the $\cos^{-1}$ function is always nonnegative, we note the curious fact that the maximum Procrustean distance between shapes in Σ_2^n is $\pi/2$. The reader should also note that the right hand side of this identity does not depend upon the orientation of the pre-shapes τ_1 and τ_2. A rotation of these pre-shapes corresponds to multiplying each τ_{jk} by an element of the unit circle in the complex plane. This factors out of the summation and has modulus one.

Let us apply this formula to the shape differences of the landmarks of Figure 1.1. An inspection of this figure would suggest that the landmarks of the second and third brooches are closer in shape to each other than they are to the landmarks of the first brooch. It remains to be seen whether the shape analysis from landmarks supports this conclusion. In each of the three images, we have $n = 4$ landmarks. Let τ_1, τ_2, and τ_3 be the pre-shapes of the respective configurations of landmarks shown in Figure 1.1, as defined by formula (1.8). Additionally, let σ_1, σ_2, and σ_3 be the respective shapes of these sets of landmarks. Then from formula (1.21), we get $d(\sigma_1, \sigma_2) = 0.380$, $d(\sigma_1, \sigma_3) = 0.308$, and $d(\sigma_2, \sigma_3) = 0.132$. As would be expected, the smallest shape difference is between the second and third brooches. The first brooch can be distinguished from the other two by the fact that its loop is fastened at the top much further to the right than the others. In terms of landmarks, we can see that x_3 and x_4 are shifted closer to x_1 in the first brooch than is the case for the second and third brooches. The landmark analysis also suggests that the third brooch is slightly closer in shape to the first brooch than is the second.

The three brooches that we have considered here for the sake of example are part of a larger set of brooches. In Section 3.7 we shall conduct a shape analysis of the complete set of images. Of course, such conclusions are dependent upon the choice of landmarks on the brooches. Four landmarks are too few to draw more than crude comparisons between the shapes of the brooches. In Chapter 6 we shall consider methods to study the shape variation between the brooches in finer detail.

While the shape metric provides a geometric structure to Σ_2^n, we are still left with a considerable difficulty in interpreting and visualizing this space. In Chapter 3, we shall construct some concrete representations of shape spaces. Moreover, before we leap upon such a choice for the geometry of shape space, it is worth bearing in mind that this choice of metric is closely connected to the concept of a metric between pre-shapes. However, the great circle distance between pre-shapes on the sphere $\mathbf{S}_*^{2n-3}$ is a consequence of the standardization technique, namely the rescaling of the original centered landmarks so that $\mathrm{tr}(\Gamma) = 1$, where $\mathrm{tr}(\,.\,)$ is the trace function defined in formula (1.7). The conclusions drawn from a shape analysis based upon a metric geometry of shape space will depend in part upon the choice of size variable used to compare shapes. In Chapter 3, we will examine various geometries of shape space and will find some simple representations for special cases.

1.4 A Few More Examples

1.4.1 A Simple Example in One Dimension

Throughout this and subsequent chapters, we shall be primarily concerned with the representations of shapes of landmarks in dimension two or above. However, before proceeding to that material, it is useful to consider what happens with landmarks that lie along a line.

First and foremost, we should note that one dimensional configurations of landmarks cannot be rotated. Therefore, the pre-shapes of such configurations of landmarks can be identified with their shapes, there being nothing more to remove upon reduction to the pre-shape. This makes the representation of shapes in dimension one a very easy thing to do. Pre-shapes lie naturally on a sphere. We have seen this, in particular, for landmarks in the plane. However, it remains true for landmarks in any dimension. If we have $n \geq 3$ landmarks along a line, then the pre-shape

$$\tau = \left(\frac{x_1 - \bar{x}}{\sqrt{\sum (x_j - \bar{x})^2}}, \ \cdots, \ \frac{x_n - \bar{x}}{\sqrt{\sum (x_j - \bar{x})^2}} \right) \tag{1.22}$$

of these n landmarks will lie in a sphere

$$\mathbf{S}_*^{n-2} = \{(x_1, ..., x_n) : \sum x_j = 0, \ \sum x_j^2 = 1\} \tag{1.23}$$

of dimension $n-2$. A sphere of dimension one is, of course, a circle.

Even three landmarks along a line can sometimes be used to make basic shape comparisons. Consider for example Figure 1.2, which shows the profiles of four skulls. Also plotted over each of the skulls is a set of three landmarks, chosen according to a landmark selection method proposed by

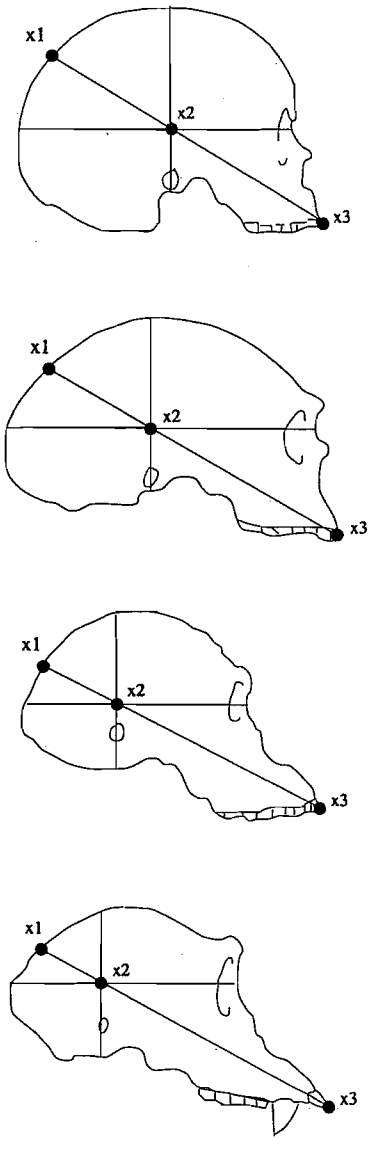

FIGURE 1.2. *Side view of skulls. From top to bottom: modern human, Neanderthal, australopithecine, chimpanzee. The skull profiles are redrawn from Figure 3.53 of [131].*

Michael Lewis of the University of Waterloo. Upon examination, the four skulls are seen to vary particularly in the ratio of the size of the cranium to the size of the jaw. In the human skull this ratio is the largest, while it is smallest for the chimpanzee. The landmarks x_1, x_2, and x_3 capture some of this variation because the cranium-to-jaw ratio is proportional to the ratio of the distances from x_1 to x_2 and from x_2 to x_3.

As x_1, x_2, and x_3 lie along a line in each picture, we can put some coordinates along each line and consider (x_1, x_2, x_3) to be a vector in $\mathbf{R}^3$. The pre-shape τ of such a vector will then be an element of the unit circle $\mathbf{S}^1$. Figure 1.3 shows the pre-shapes of these four configurations of three points plotted on a circle. The reader may be surprised by the small amount of arc length enclosed within the range of the four pre-shape points in Figure 1.3. This is quite typical of landmarks chosen on biological organisms. Usually, the amount of variation of landmark coordinates between images is small compared to the distances between the landmarks within an image.

A small arc of a circle can be approximated by a line segment. So it is tempting to approximate the positions of pre-shapes on the circle in Figure 1.3 by a similar configuration along a straight line. Such an approximation is called a *tangent approximation*, and works quite well for many biological data sets. More generally, however, configurations of points on a circle cannot be approximated by a configuration of points along a line without major distortion of the interpoint distances. Similarly, a configuration of points on a shape space such as Σ_2^n cannot be approximated by a multivariate configuration in $\mathbf{R}^{2n-4}$ without distorting the interpoint Procrustean distances. So it is fortunate when such a tangent approximation is possible, because it permits the researcher to apply the large collection of multivariate statistical techniques designed for data in Euclidean space. In general, the tangent approximation cannot always be used. Therefore, we must turn to the methods of differential geometry to represent shapes.

1.4.2 Dinosaur Trackways From Mt. Tom, Massachusetts

The statistical theory of shape is particularly concerned with the study of *random* shapes, and shape comparisons in the presence of random variation in shape. Why should a theory of shape incorporate stochastic assumptions? Let us consider two examples in this and the following section.

Consider Figure 1.4, which shows the footprints of dinosaurs of the Late Triassic/Early Jurassic period at the Mt. Tom site north of Holyoke, Massachusetts. This data set is described by Ostrom [130]. One of the interesting features of this data set is the presence of multiple tracks that are sufficiently separated to permit the examination of

- variation of tracks along the path of a single dinosaur;
- variation of tracks between dinosaurs of the same species;

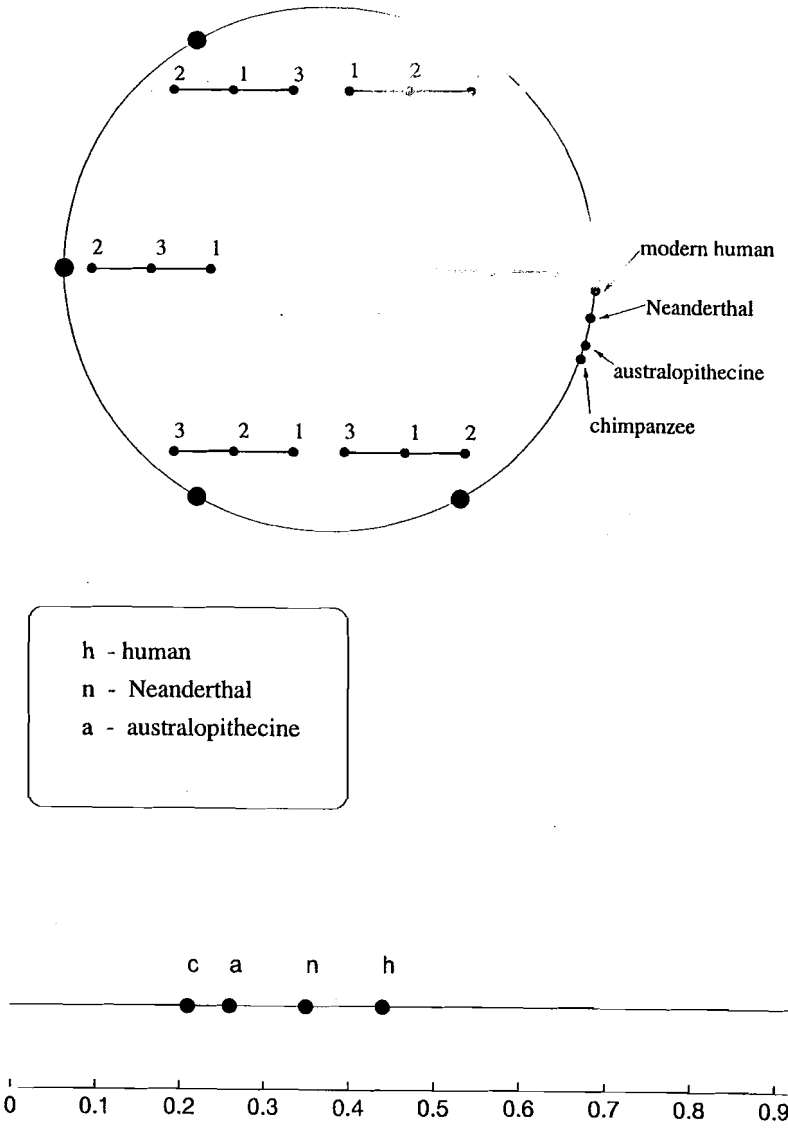

FIGURE 1.3. *Pre-shapes of the four skulls plotted on a circle (above), and with a tangent approximation (below). Also marked on the circle are the six pre-shapes of configurations of equally spaced points for reference purposes.*

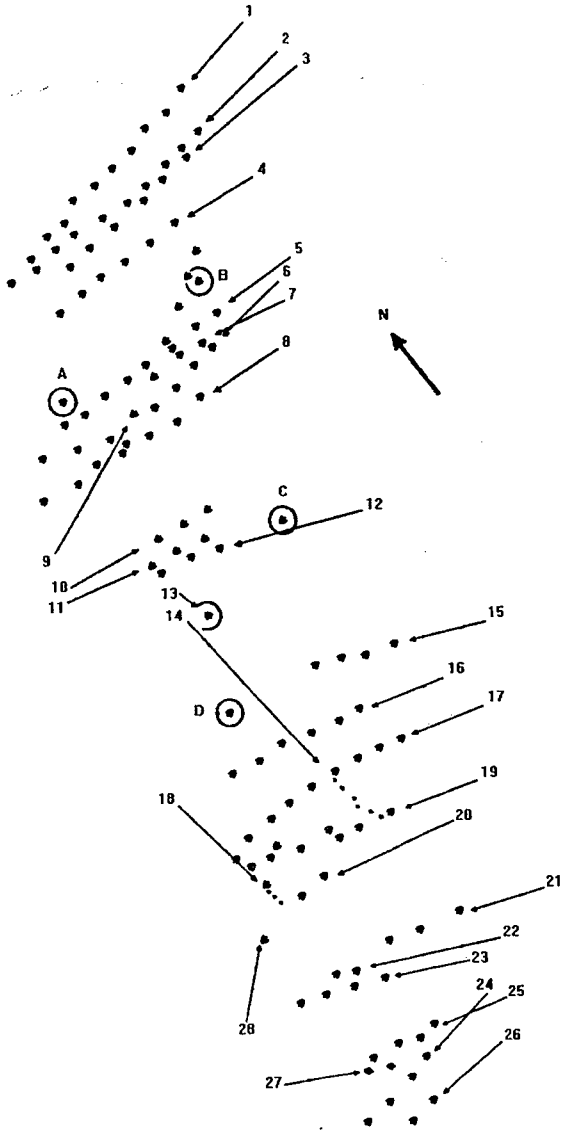

FIGURE 1.4. *Dinosaur footprints at the Mt. Tom site near Holyoke, Massachusetts. Footprints can be grouped in partly overlapping trackways corresponding to three species of dinosaurs.*

and, to a certain extent,

- variation in tracks between species.

Multiple comparisons between species and individuals are possible when footprints can be clearly delineated as belonging to different dinosaurs or different species.

For example, Ostrom was struck by the tendency of most of the dinosaur tracks to go in roughly the same direction. He considered the evidence from this site and others for the possible gregarious behavior of dinosaurs. The question of whether dinosaurs had any tendency to congregate in packs or herds is an interesting problem within a much larger issue. Experts have long recognized that dinosaurs had a combination of reptilian and avian features. In modern animals, gregarious behavior is most commonly found in birds rather than reptiles. So any evidence for such behavior would support a more avian interpretation of dinosaurs. In examining the site, Ostrom found indications of twenty-eight trackways made of three distinct types of footprints: large broad footprints identified as made by *Eubrontes*, intermediate size prints resembling those made by *Anchisauripus*, and small prints identified as made by *Grallator*. Each of the twenty-eight trackways was assigned an overall direction, and these directions were examined within and between species. The trackway directions were classified into two types: those tracks pointing in a roughly westerly direction ranging through an angle of about $30°$ and sundry directions far removed from the westerly trackways.

The fact that the majority of the trackways point in a westerly direction is suggestive of herding behavior. However, we must be cautious with this conclusion. We cannot automatically conclude that the directionality is due to herding because we do not know about the presence of other external agencies that might have forced the dinosaurs in this direction. A more reliable indicator is any possible relationship between species (as determined by footprint classification) and behavior (as determined by track direction). Ignoring trackway 13, which consists of a single print pointing south and whose identification as *Eubrontes* is suspect, we can classify the trackways using a 2×2 table as follows.

Track	West	Other
Eubrontes	19	3
Other	1	4

A simple method for detecting the presence of gregarious behavior from this table is to test for independence between species, listed vertically, and direction, listed horizontally. So the null hypothesis that gregarious behavior is absent can be modeled by the hypothesis of independence of rows and

Michael Lewis of the University of Waterloo. Upon examination, the four skulls are seen to vary particularly in the ratio of the size of the cranium to the size of the jaw. In the human skull this ratio is the largest, while it is smallest for the chimpanzee. The landmarks x_1, x_2, and x_3 capture some of this variation because the cranium-to-jaw ratio is proportional to the ratio of the distances from x_1 to x_2 and from x_2 to x_3.

As x_1, x_2, and x_3 lie along a line in each picture, we can put some coordinates along each line and consider (x_1, x_2, x_3) to be a vector in $\mathbf{R}^3$. The pre-shape τ of such a vector will then be an element of the unit circle $\mathbf{S}^1$. Figure 1.3 shows the pre-shapes of these four configurations of three points plotted on a circle. The reader may be surprised by the small amount of arc length enclosed within the range of the four pre-shape points in Figure 1.3. This is quite typical of landmarks chosen on biological organisms. Usually, the amount of variation of landmark coordinates between images is small compared to the distances between the landmarks within an image.

A small arc of a circle can be approximated by a line segment. So it is tempting to approximate the positions of pre-shapes on the circle in Figure 1.3 by a similar configuration along a straight line. Such an approximation is called a *tangent approximation*, and works quite well for many biological data sets. More generally, however, configurations of points on a circle cannot be approximated by a configuration of points along a line without major distortion of the interpoint distances. Similarly, a configuration of points on a shape space such as Σ_2^n cannot be approximated by a multivariate configuration in $\mathbf{R}^{2n-4}$ without distorting the interpoint Procrustean distances. So it is fortunate when such a tangent approximation is possible, because it permits the researcher to apply the large collection of multivariate statistical techniques designed for data in Euclidean space. In general, the tangent approximation cannot always be used. Therefore, we must turn to the methods of differential geometry to represent shapes.

1.4.2 Dinosaur Trackways From Mt. Tom, Massachusetts

The statistical theory of shape is particularly concerned with the study of *random* shapes, and shape comparisons in the presence of random variation in shape. Why should a theory of shape incorporate stochastic assumptions? Let us consider two examples in this and the following section.

Consider Figure 1.4, which shows the footprints of dinosaurs of the Late Triassic/Early Jurassic period at the Mt. Tom site north of Holyoke, Massachusetts. This data set is described by Ostrom [130]. One of the interesting features of this data set is the presence of multiple tracks that are sufficiently separated to permit the examination of

- variation of tracks along the path of a single dinosaur;
- variation of tracks between dinosaurs of the same species;

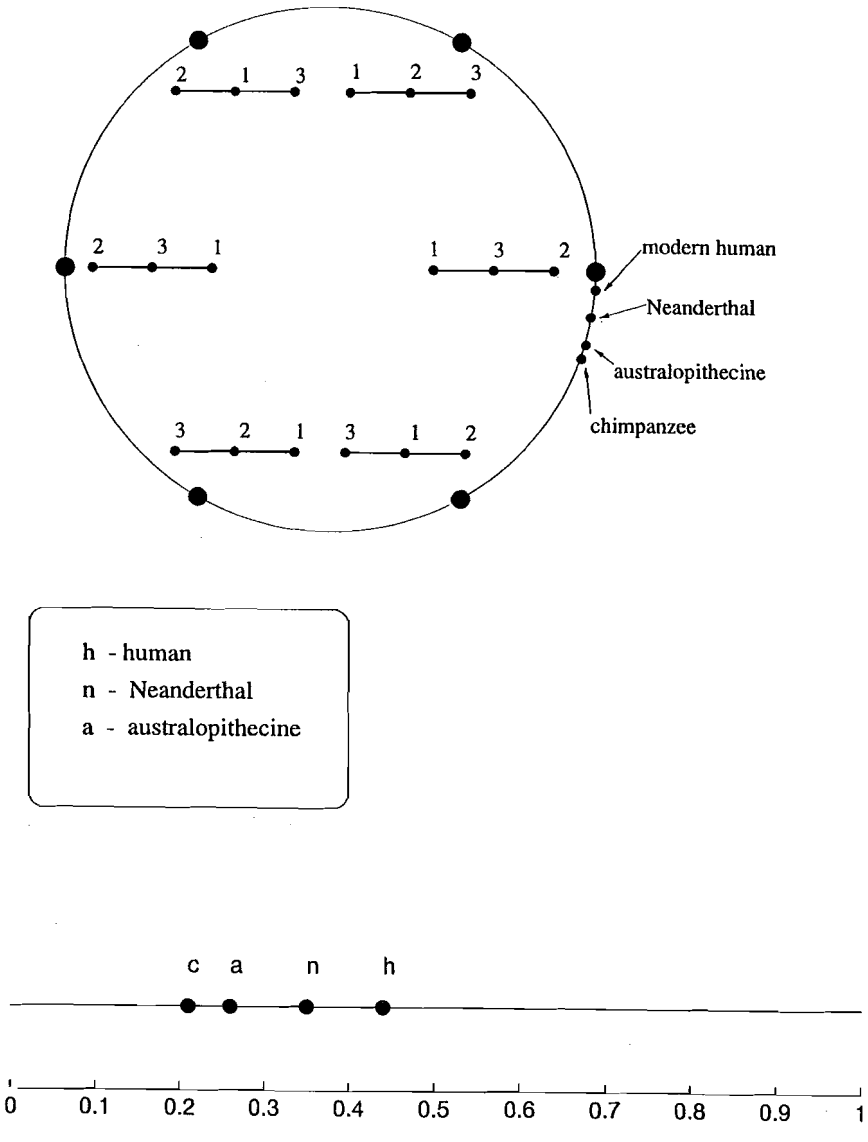

FIGURE 1.3. *Pre-shapes of the four skulls plotted on a circle (above), and with a tangent approximation (below). Also marked on the circle are the six pre-shapes of configurations of equally spaced points for reference purposes.*

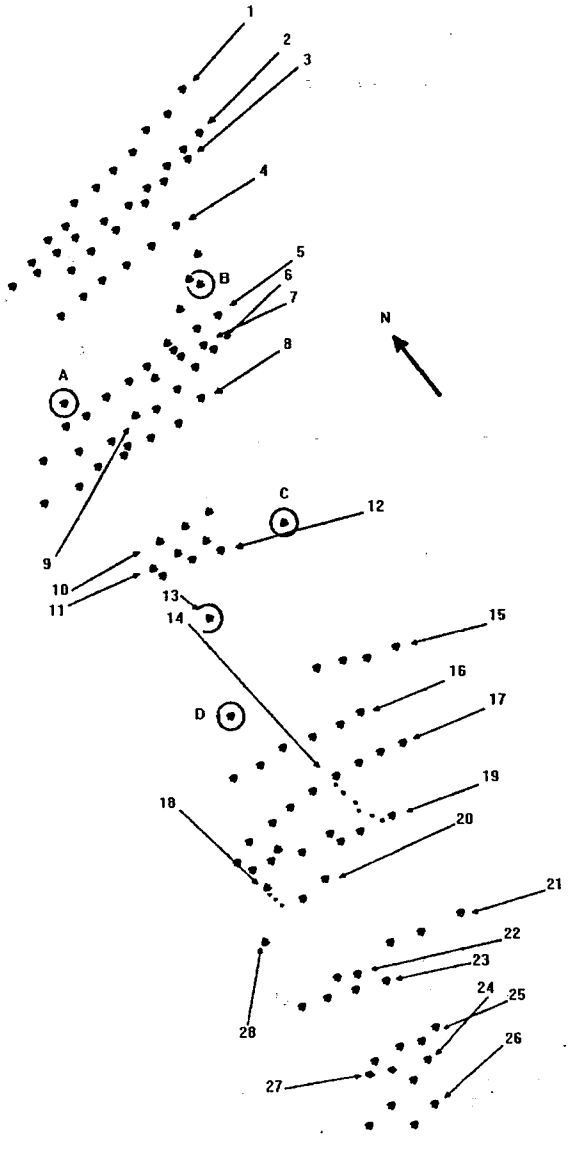

FIGURE 1.4. *Dinosaur footprints at the Mt. Tom site near Holyoke, Massachusetts. Footprints can be grouped in partly overlapping trackways corresponding to three species of dinosaurs.*

and, to a certain extent,

- variation in tracks between species.

Multiple comparisons between species and individuals are possible when footprints can be clearly delineated as belonging to different dinosaurs or different species.

For example, Ostrom was struck by the tendency of most of the dinosaur tracks to go in roughly the same direction. He considered the evidence from this site and others for the possible gregarious behavior of dinosaurs. The question of whether dinosaurs had any tendency to congregate in packs or herds is an interesting problem within a much larger issue. Experts have long recognized that dinosaurs had a combination of reptilian and avian features. In modern animals, gregarious behavior is most commonly found in birds rather than reptiles. So any evidence for such behavior would support a more avian interpretation of dinosaurs. In examining the site, Ostrom found indications of twenty-eight trackways made of three distinct types of footprints: large broad footprints identified as made by *Eubrontes*, intermediate size prints resembling those made by *Anchisauripus*, and small prints identified as made by *Grallator*. Each of the twenty-eight trackways was assigned an overall direction, and these directions were examined within and between species. The trackway directions were classified into two types: those tracks pointing in a roughly westerly direction ranging through an angle of about $30°$ and sundry directions far removed from the westerly trackways.

The fact that the majority of the trackways point in a westerly direction is suggestive of herding behavior. However, we must be cautious with this conclusion. We cannot automatically conclude that the directionality is due to herding because we do not know about the presence of other external agencies that might have forced the dinosaurs in this direction. A more reliable indicator is any possible relationship between species (as determined by footprint classification) and behavior (as determined by track direction). Ignoring trackway 13, which consists of a single print pointing south and whose identification as *Eubrontes* is suspect, we can classify the trackways using a 2×2 table as follows.

Track	West	Other
Eubrontes	19	3
Other	1	4

A simple method for detecting the presence of gregarious behavior from this table is to test for independence between species, listed vertically, and direction, listed horizontally. So the null hypothesis that gregarious behavior is absent can be modeled by the hypothesis of independence of rows and

columns. A test for independence on this 2×2 table is quite significant, and in favor of the hypothesis that there is gregarious behavior. However, we must be cautious in our conclusions because other factors could affect the relationship between track direction and species other than the herding hypothesis.

More generally, we might seek to model dinosaur movements across the area so as to make inferences about differences between individuals within species and between species. Quite a large number of footprints of *Eubrontes* are available. In track 1, for example, the footprints are clearly defined as belonging to a single *Eubrontes*, and can be interpreted in order as a sequence of successive footprints. Can we use this and similar tracks to model dinosaur motion? We can model a sequence of consecutive footprints as generated by some appropriate random mechanism and then attempt to make inferences by decomposing the geometric configuration of footprints in a trackway into orientation, size, and shape information. We have already performed a rough analysis of the orientations in considering herding behavior. In Chapter 6, we shall consider how size information, available through stride length, can be used to estimate the speed with which the dinosaurs crossed the site. Finally, we shall perform a shape analysis on the trackways and in particular shall investigate how the shape of the triangle formed by three successive footprints is correlated with size variables such as stride length. The unifying approach to such data sets will be to decompose the geometric information into its orientation, size, and shape components, and to consider the variation in these components and their relation to each other.

1.4.3 Late Bronze Age Post Mold Configurations in England

Consider the configuration of post molds from two Late Bronze Age sites at Aldermaston Wharf and at South Lodge camp in Wiltshire, England. See Figures 1.5 and 1.6. In archeological excavations, clear evidence is often found for the existence of wooden buildings at the site through the configurations of supporting posts of the structure. While these posts are no longer present at the site, the positions of many of them can be determined from the presence of round discolorations of the soil beneath the surface. These discolorations, or post molds, are often found in a roughly regular geometric pattern that indicates the presence of a wall. However, complications can arise in interpreting the post mold evidence. Destructive processes such as erosion can prevent post molds from being detected. Sometimes a building at a particular location was demolished and a succession of other buildings erected at the same place. In these cases, the superimposed post mold patterns can be very difficult to disentangle. From Figures 1.5 and 1.6 we see such problems. It is known that typical buildings of the time were circular structures called roundhouses. Neighboring posts were usually 1.6 to 2.2 meters apart, and possibly up to three meters apart. In Figure 1.6, the

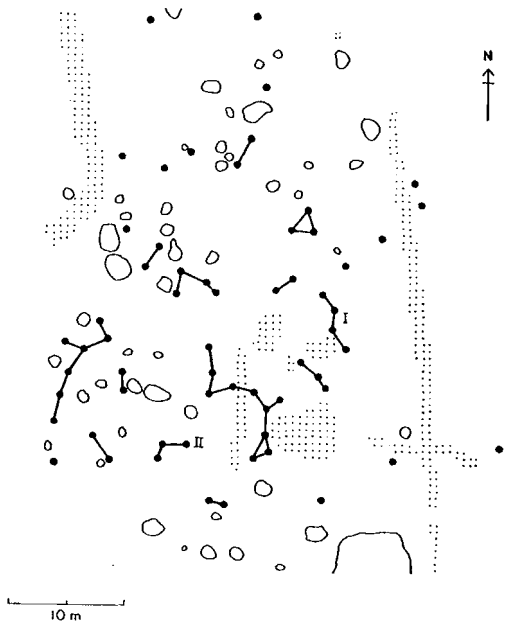

FIGURE 1.5. *Post mold configuration at Aldermaston Wharf showing links between neighboring post molds. Later features are marked as shaded regions. Irregular unshaded regions are pits at the site. This figure is adapted from [32] by kind permission of The Museum Applied Science Center at the University of Pennsylvania.*

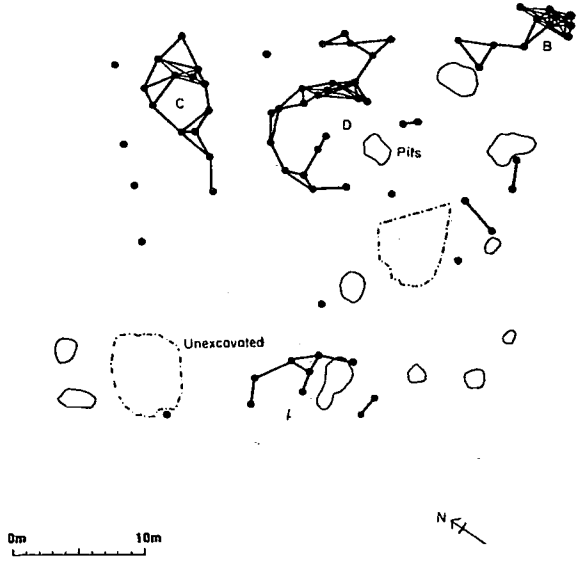

FIGURE 1.6. *Post mold configuration at South Lodge Camp showing links between neighboring post molds. A large, highly regular circular configuration of post molds can be seen on the east side of the site. A smaller circle adjacent to it is also visible. This figure is adapted from [32] by kind permission of The Museum Applied Science Center at the University of Pennsylvania.*

post molds whose interpoint distances are less than three meters have been linked by a line segment. Four main clusters of points, labeled A, B, C, and D can be seen. Strong visual evidence for the existence of a roundhouse can be seen in cluster D of the outline plan of South Lodge Camp. The clearly circular arrangement of posts would be difficult to explain as a coincidence from a purely random mechanism. On the other hand, the evidence from cluster C is more ambiguous. Here there is also some indication of a roundhouse. However, in this case it is more difficult to determine whether the circular pattern is too regular to arise simply by chance. Finally, in cluster A there is a very slight indication of a roundhouse. But here we would have to admit that any evidence of a circle could quite possibly be coincidental. There is no clear confirmation that a circular building was present here, although there is a suggestion of circularity in the positions of the post molds.

At Aldermaston Wharf, the evidence for circular buildings is provided by the positions of post molds clustered visually as Structure I and Structure II in Figure 1.5. Of these two, Structure II is the better formed, and has six post molds that can be placed on a rough circle. Structure I looks very irregular. Again, there are six post molds that can be interpreted as circular. Neither structure is as compelling as Cluster D from South Lodge Camp.

How should we assess the patterns at these two sites, and how can we determine whether such configurations are likely by chance in a random scattering? A method for fitting circles that is particularly amenable to analysis of this kind has been provided by Cogbill [44]. He proposed that circular configurations of posts can be detected by running an annulus across the window in which the posts are plotted. If the inner and outer radii of the annulus are close, the thin annulus will cover few points in any given position. However, by chance, at certain positions a larger number of points will be covered. Such configurations of posts can be examined for the possibility that they form the circular boundary of a roundhouse. For example, the six points of Structure II at Aldermaston Wharf can be completely contained in an annulus whose inner radius is 3.66 meters and whose outer radius is 3.95 meters. Is such a fit likely by chance? We could define chance configurations as those arising in a random uniform scattering of equally many points over a similar region. In such a scattering, what is the expected number of circles that will be found of six points covered by an annulus of inner and outer radii 3.66 and 3.95 meters respectively? Early work by Mack [111] provides a powerful tool for answering this question. In Chapter 6, we shall see that we would expect to discover a circular arrangement of this tolerance simply by chance if the posts were randomly scattered across the region of excavation. Such a calculation casts doubt upon the strength of the archeological interpretation at Aldermaston. A similar analysis of Cluster D at South Lodge Camp is more reassuring for archeological interpretation. In this case, a set of eight points can be fit with

columns. A test for independence on this 2×2 table is quite significant, and in favor of the hypothesis that there is gregarious behavior. However, we must be cautious in our conclusions because other factors could affect the relationship between track direction and species other than the herding hypothesis.

More generally, we might seek to model dinosaur movements across the area so as to make inferences about differences between individuals within species and between species. Quite a large number of footprints of *Eubrontes* are available. In track 1, for example, the footprints are clearly defined as belonging to a single *Eubrontes*, and can be interpreted in order as a sequence of successive footprints. Can we use this and similar tracks to model dinosaur motion? We can model a sequence of consecutive footprints as generated by some appropriate random mechanism and then attempt to make inferences by decomposing the geometric configuration of footprints in a trackway into orientation, size, and shape information. We have already performed a rough analysis of the orientations in considering herding behavior. In Chapter 6, we shall consider how size information, available through stride length, can be used to estimate the speed with which the dinosaurs crossed the site. Finally, we shall perform a shape analysis on the trackways and in particular shall investigate how the shape of the triangle formed by three successive footprints is correlated with size variables such as stride length. The unifying approach to such data sets will be to decompose the geometric information into its orientation, size, and shape components, and to consider the variation in these components and their relation to each other.

1.4.3 Late Bronze Age Post Mold Configurations in England

Consider the configuration of post molds from two Late Bronze Age sites at Aldermaston Wharf and at South Lodge camp in Wiltshire, England. See Figures 1.5 and 1.6. In archeological excavations, clear evidence is often found for the existence of wooden buildings at the site through the configurations of supporting posts of the structure. While these posts are no longer present at the site, the positions of many of them can be determined from the presence of round discolorations of the soil beneath the surface. These discolorations, or post molds, are often found in a roughly regular geometric pattern that indicates the presence of a wall. However, complications can arise in interpreting the post mold evidence. Destructive processes such as erosion can prevent post molds from being detected. Sometimes a building at a particular location was demolished and a succession of other buildings erected at the same place. In these cases, the superimposed post mold patterns can be very difficult to disentangle. From Figures 1.5 and 1.6 we see such problems. It is known that typical buildings of the time were circular structures called roundhouses. Neighboring posts were usually 1.6 to 2.2 meters apart, and possibly up to three meters apart. In Figure 1.6, the

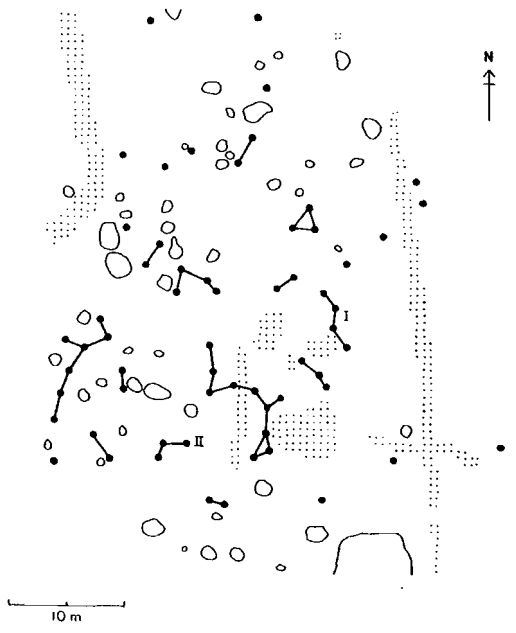

FIGURE 1.5. *Post mold configuration at Aldermaston Wharf showing links between neighboring post molds. Later features are marked as shaded regions. Irregular unshaded regions are pits at the site. This figure is adapted from [32] by kind permission of The Museum Applied Science Center at the University of Pennsylvania.*

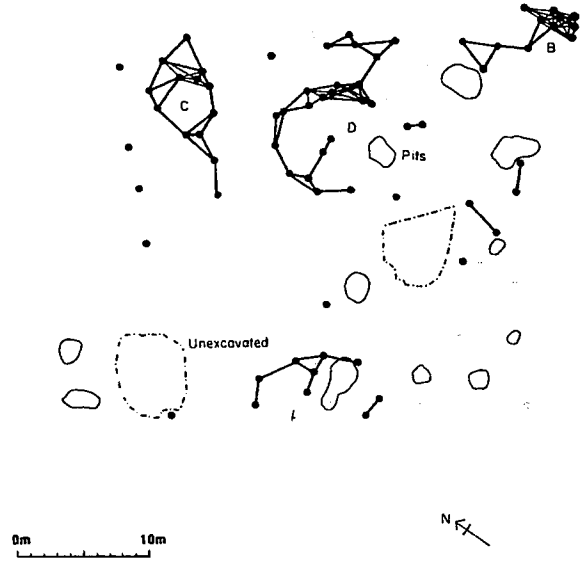

FIGURE 1.6. *Post mold configuration at South Lodge Camp showing links between neighboring post molds. A large, highly regular circular configuration of post molds can be seen on the east side of the site. A smaller circle adjacent to it is also visible. This figure is adapted from [32] by kind permission of The Museum Applied Science Center at the University of Pennsylvania.*

post molds whose interpoint distances are less than three meters have been linked by a line segment. Four main clusters of points, labeled A, B, C, and D can be seen. Strong visual evidence for the existence of a roundhouse can be seen in cluster D of the outline plan of South Lodge Camp. The clearly circular arrangement of posts would be difficult to explain as a coincidence from a purely random mechanism. On the other hand, the evidence from cluster C is more ambiguous. Here there is also some indication of a roundhouse. However, in this case it is more difficult to determine whether the circular pattern is too regular to arise simply by chance. Finally, in cluster A there is a very slight indication of a roundhouse. But here we would have to admit that any evidence of a circle could quite possibly be coincidental. There is no clear confirmation that a circular building was present here, although there is a suggestion of circularity in the positions of the post molds.

At Aldermaston Wharf, the evidence for circular buildings is provided by the positions of post molds clustered visually as Structure I and Structure II in Figure 1.5. Of these two, Structure II is the better formed, and has six post molds that can be placed on a rough circle. Structure I looks very irregular. Again, there are six post molds that can be interpreted as circular. Neither structure is as compelling as Cluster D from South Lodge Camp.

How should we assess the patterns at these two sites, and how can we determine whether such configurations are likely by chance in a random scattering? A method for fitting circles that is particularly amenable to analysis of this kind has been provided by Cogbill [44]. He proposed that circular configurations of posts can be detected by running an annulus across the window in which the posts are plotted. If the inner and outer radii of the annulus are close, the thin annulus will cover few points in any given position. However, by chance, at certain positions a larger number of points will be covered. Such configurations of posts can be examined for the possibility that they form the circular boundary of a roundhouse. For example, the six points of Structure II at Aldermaston Wharf can be completely contained in an annulus whose inner radius is 3.66 meters and whose outer radius is 3.95 meters. Is such a fit likely by chance? We could define chance configurations as those arising in a random uniform scattering of equally many points over a similar region. In such a scattering, what is the expected number of circles that will be found of six points covered by an annulus of inner and outer radii 3.66 and 3.95 meters respectively? Early work by Mack [111] provides a powerful tool for answering this question. In Chapter 6, we shall see that we would expect to discover a circular arrangement of this tolerance simply by chance if the posts were randomly scattered across the region of excavation. Such a calculation casts doubt upon the strength of the archeological interpretation at Aldermaston. A similar analysis of Cluster D at South Lodge Camp is more reassuring for archeological interpretation. In this case, a set of eight points can be fit with

an annulus with inner radius 3.95 meters and outer radius 4.21 meters. As we shall see, we expect such circular arrangements in a comparable random scattering less than one time in six. Even this looks rather high in view of the precision of the circle of points in Cluster D. However, the circular fit does not take into account the even spacing of posts, which is also unlikely in a random scattering.

1.5 The Problem of Homology

In the biological sciences, sites or landmarks on different organisms are said to be *homologous* if they share a common structure and evolutionary origin. For example, the eyes of a chimpanzee are homologous to the eyes of a human despite the shape differences between the head of a chimpanzee and the head of a human. More generally, outside the biological sciences, sites on different bodies or images are said to be homologous if they naturally correspond due to a common structure. We considered an example of this in Section 1.3, where we chose four landmarks on each of three Iron Ages brooches so that correspondingly labeled landmarks were homologous between images. Homologous landmarks are not always obvious, and may depend upon insight or expert opinion for their construction.

As an illustration of the problems associated with constructing satisfactory homologies between images, let us consider the work of Thompson [172], who devised a method for examining shape differences between biological organisms called the *method of coordinates*. The reader can find an example of Thompson's method by looking at Figure 1.7. In this figure, we see four lateral views of the skulls that we considered in Section 1.4.1 and Figure 1.2. Thompson proposed the placement of a rectangular grid over one of the images, say the modern human skull at the top. Now, each of the intersection points of the grid corresponds to a feature of some kind in the skull. (The detection of such features requires more detailed information than is available in Figure 1.7.) Suppose that for each feature at every intersection point in the top grid we are able to find the corresponding (homologous) feature in the other skulls. A horizontal or vertical line of the Cartesian grid on the top image is mapped to a curvilinear line in each of the other images by connecting sites in the other images that are homologous to sites on the same horizontal or vertical line of the top image. The resulting coordinate system superimposed on the second image is typically curvilinear in nature. The degree to which the curvilinear coordinate systems depart from a Cartesian frame is a measure of the shape differences between the images.

By looking at the curvilinear coordinate systems of Figure 1.7, we can make some detailed observations about the shape variation among the four skulls. In particular, by looking at the upper left and lower right corners,

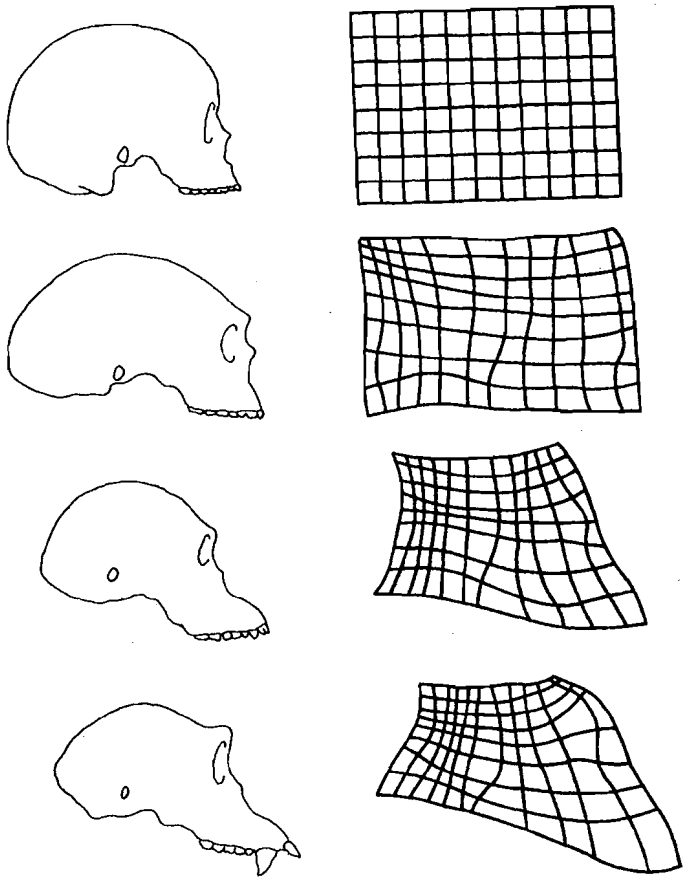

FIGURE 1.7. *Side view of skulls. From top to bottom: modern human, Neanderthal, australopithecine, chimpanzee. To the right of each skull is a coordinate grid determined with Thompson's method of coordinates, with the modern human skull as the base image. Reproduced from Figure 3.53 of [131] by kind permission of Hong Kong University Press.*

we can see what was observed in Section 1.4.1, namely that an important source of variation is to be found in the change in the relative sizes of jaw and cranium. With far more coordinates available for comparison, we are able to make a much more detailed examination of these differences.

However, Thompson's method of coordinates has several problems. The first of these is the problem of how to draw a smooth line appropriately through a set of points. This is essentially an interpolation, or fitting problem. The second problem is that mentioned above, namely of finding a correspondence, or homology, between landmarks on different images. A final problem is to decide how to summarize the information available about the differences in shapes among the images from such complicated grids of curvilinear coordinates. We will consider these problems again in Chapter 3.

1.6 Notes

The theory of shape owes much to D'Arcy Thompson [172] for its inspiration. His work has long been regarded as a model for the fusion of scientific, mathematical, and literary skills. Although his analyses of biological growth and form are now dated, his exposition of the theory of biological shape is unparalleled for its clarity. The reader who has not encountered his work is strongly encouraged to do so.

For a comprehensive discussion of the theory and methods of morphometrics, the reader is referred to [139]. Brief surveys of allometric methods are to be found in [81] and [125]. A variety of applications is readily available in the literature, including [7], [10], [13], [24], [45], [63], [85], [107], and [126], to name a few.

The mathematical theory of shape that has been introduced in this chapter can be found in Kendall [90]. This paper was seminal for the development of this particular school of shape theory, which can be called the Kendall school or perhaps the Procrustean school of shape analysis. Much of the material in the following chapters relies on the Kendall school of shape and takes advantage of its comprehensive methodology for the analysis of finite point sets in arbitrary dimensions. In particular, the definition and metric of Σ_p^n, the space of shapes of n points in p dimensions, is due to Kendall. For extensions of Kendall's work to more general multivariate normal models, see [53].

The Bookstein school of shape analysis uses a different geometric structure on shape spaces that will be discussed in Chapter 3. As mentioned earlier, we have used the word *landmark* in a more general sense than Bookstein as a point chosen from a body that helps summarize its geometric features. Bookstein has recommended the use of landmarks for the analysis of biological features and constrains the choice of landmarks to

prominent features of the organism or biological structure. For the analysis of more general shapes outside the biological sciences, the choice of natural sites for landmarks remains a desirable goal, but is very restrictive for shape description. Therefore, we choose a generalized interpretation of landmark data. A synthesis of the Kendall, or Procrustean, school of shape with the use of landmarks can be found in the survey paper of Goodall [66].

An alternative approach to the selection of landmarks can be found in [60].

1.7 Problems

1. A researcher proposes to define the shape of a triangle as a vector $(\alpha_1, \alpha_2, \alpha_3)$ of three internal angles. Discuss the advantages and disadvantages of encoding shape information in this way.

2. Two triangles are congruent if their corresponding sides are of equal length. A researcher proposes to encode the size and shape information about a triangle as a vector $(d_{12}, d_{13}, d_{23}) \in \mathbf{R}^3$ where d_{jk} is the length of the side joining the jth and kth vertices. A size variable $w(d_{12}, d_{13}, d_{23})$ is a nonnegative function that is homogeneous, in the sense that

$$w(td_{12},\ td_{13},\ td_{23}) \ = \ t\, w(d_{12},\ d_{13},\ d_{23}) \tag{1.24}$$

for all $t \geq 0$. Give two distinct examples of size variables and show how shape coordinates for a triangle can be constructed by standardizing the d_{jk} with respect to size.

3. The next two problems involve the concept of a random shape statistic. In this problem, the shape statistic in question is the maximum internal angle of a random triangle. In the next problem, the statistic is an indicator of the event that a random quadrilateral is convex. The reader who is not familiar with the probability theory used in these questions can safely pass over these problems until we return to probability theory in Chapter 4.

Three random planar points are independent, with a common absolutely continuous distribution. Let M be the maximum internal angle of the triangle whose vertices are the three points. Show that

$$\mathcal{P}(M \geq 120^\circ) \ \geq \ 1/20 \tag{1.25}$$

(Hint: consider six such random points.)

4. Four random planar points are independent, with a common absolutely continuous distribution. Show that with probability greater than or equal

an annulus with inner radius 3.95 meters and outer radius 4.21 meters. As we shall see, we expect such circular arrangements in a comparable random scattering less than one time in six. Even this looks rather high in view of the precision of the circle of points in Cluster D. However, the circular fit does not take into account the even spacing of posts, which is also unlikely in a random scattering.

1.5 The Problem of Homology

In the biological sciences, sites or landmarks on different organisms are said to be *homologous* if they share a common structure and evolutionary origin. For example, the eyes of a chimpanzee are homologous to the eyes of a human despite the shape differences between the head of a chimpanzee and the head of a human. More generally, outside the biological sciences, sites on different bodies or images are said to be homologous if they naturally correspond due to a common structure. We considered an example of this in Section 1.3, where we chose four landmarks on each of three Iron Ages brooches so that correspondingly labeled landmarks were homologous between images. Homologous landmarks are not always obvious, and may depend upon insight or expert opinion for their construction.

As an illustration of the problems associated with constructing satisfactory homologies between images, let us consider the work of Thompson [172], who devised a method for examining shape differences between biological organisms called the *method of coordinates*. The reader can find an example of Thompson's method by looking at Figure 1.7. In this figure, we see four lateral views of the skulls that we considered in Section 1.4.1 and Figure 1.2. Thompson proposed the placement of a rectangular grid over one of the images, say the modern human skull at the top. Now, each of the intersection points of the grid corresponds to a feature of some kind in the skull. (The detection of such features requires more detailed information than is available in Figure 1.7.) Suppose that for each feature at every intersection point in the top grid we are able to find the corresponding (homologous) feature in the other skulls. A horizontal or vertical line of the Cartesian grid on the top image is mapped to a curvilinear line in each of the other images by connecting sites in the other images that are homologous to sites on the same horizontal or vertical line of the top image. The resulting coordinate system superimposed on the second image is typically curvilinear in nature. The degree to which the curvilinear coordinate systems depart from a Cartesian frame is a measure of the shape differences between the images.

By looking at the curvilinear coordinate systems of Figure 1.7, we can make some detailed observations about the shape variation among the four skulls. In particular, by looking at the upper left and lower right corners,

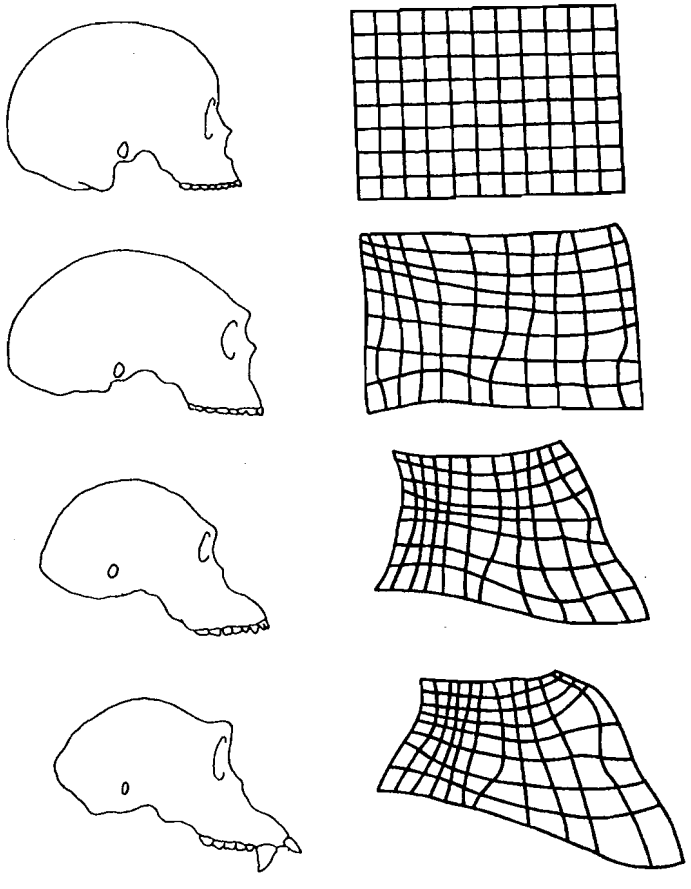

FIGURE 1.7. *Side view of skulls. From top to bottom: modern human, Neanderthal, australopithecine, chimpanzee. To the right of each skull is a coordinate grid determined with Thompson's method of coordinates, with the modern human skull as the base image. Reproduced from Figure 3.53 of [131] by kind permission of Hong Kong University Press.*

we can see what was observed in Section 1.4.1, namely that an important source of variation is to be found in the change in the relative sizes of jaw and cranium. With far more coordinates available for comparison, we are able to make a much more detailed examination of these differences.

However, Thompson's method of coordinates has several problems. The first of these is the problem of how to draw a smooth line appropriately through a set of points. This is essentially an interpolation, or fitting problem. The second problem is that mentioned above, namely of finding a correspondence, or homology, between landmarks on different images. A final problem is to decide how to summarize the information available about the differences in shapes among the images from such complicated grids of curvilinear coordinates. We will consider these problems again in Chapter 3.

1.6 Notes

The theory of shape owes much to D'Arcy Thompson [172] for its inspiration. His work has long been regarded as a model for the fusion of scientific, mathematical, and literary skills. Although his analyses of biological growth and form are now dated, his exposition of the theory of biological shape is unparalleled for its clarity. The reader who has not encountered his work is strongly encouraged to do so.

For a comprehensive discussion of the theory and methods of morphometrics, the reader is referred to [139]. Brief surveys of allometric methods are to be found in [81] and [125]. A variety of applications is readily available in the literature, including [7], [10], [13], [24], [45], [63], [85], [107], and [126], to name a few.

The mathematical theory of shape that has been introduced in this chapter can be found in Kendall [90]. This paper was seminal for the development of this particular school of shape theory, which can be called the Kendall school or perhaps the Procrustean school of shape analysis. Much of the material in the following chapters relies on the Kendall school of shape and takes advantage of its comprehensive methodology for the analysis of finite point sets in arbitrary dimensions. In particular, the definition and metric of Σ_p^n, the space of shapes of n points in p dimensions, is due to Kendall. For extensions of Kendall's work to more general multivariate normal models, see [53].

The Bookstein school of shape analysis uses a different geometric structure on shape spaces that will be discussed in Chapter 3. As mentioned earlier, we have used the word *landmark* in a more general sense than Bookstein as a point chosen from a body that helps summarize its geometric features. Bookstein has recommended the use of landmarks for the analysis of biological features and constrains the choice of landmarks to

prominent features of the organism or biological structure. For the analysis of more general shapes outside the biological sciences, the choice of natural sites for landmarks remains a desirable goal, but is very restrictive for shape description. Therefore, we choose a generalized interpretation of landmark data. A synthesis of the Kendall, or Procrustean, school of shape with the use of landmarks can be found in the survey paper of Goodall [66].

An alternative approach to the selection of landmarks can be found in [60].

1.7 Problems

1. A researcher proposes to define the shape of a triangle as a vector $(\alpha_1, \alpha_2, \alpha_3)$ of three internal angles. Discuss the advantages and disadvantages of encoding shape information in this way.

2. Two triangles are congruent if their corresponding sides are of equal length. A researcher proposes to encode the size and shape information about a triangle as a vector $(d_{12}, d_{13}, d_{23}) \in \mathbf{R}^3$ where d_{jk} is the length of the side joining the jth and kth vertices. A size variable $w(d_{12}, d_{13}, d_{23})$ is a nonnegative function that is homogeneous, in the sense that

$$w(td_{12},\ td_{13},\ td_{23}) \ = \ t\, w(d_{12},\ d_{13},\ d_{23}) \tag{1.24}$$

for all $t \geq 0$. Give two distinct examples of size variables and show how shape coordinates for a triangle can be constructed by standardizing the d_{jk} with respect to size.

3. The next two problems involve the concept of a random shape statistic. In this problem, the shape statistic in question is the maximum internal angle of a random triangle. In the next problem, the statistic is an indicator of the event that a random quadrilateral is convex. The reader who is not familiar with the probability theory used in these questions can safely pass over these problems until we return to probability theory in Chapter 4.

Three random planar points are independent, with a common absolutely continuous distribution. Let M be the maximum internal angle of the triangle whose vertices are the three points. Show that

$$\mathcal{P}(M \geq 120^\circ) \ \geq \ 1/20 \tag{1.25}$$

(Hint: consider six such random points.)

4. Four random planar points are independent, with a common absolutely continuous distribution. Show that with probability greater than or equal

to $1/5$ one of the four points will lie in the triangle formed by the other three.

5. In formula (1.21) we encountered the *Procrustean metric*. A *metric* $d(x, y)$ between points x, y of a set is a nonnegative real valued function satisfying
(i) $d(x, y) = 0$ if and only if $x = y$;
(ii) $d(x, y) = d(y, x)$ for all x and y;
(iii) $d(x, z) \leq d(x, y) + d(y, z)$ for all x, y, and z.
Show that the Procrustean metric d defined in Section 1.3 satisfies these properties on the set Σ_2^n.

2

Background Concepts and Definitions

2.1 Transformations on Euclidean Space

In this section, we shall begin with some preliminary definitions relevant to shape analysis.

2.1.1 Properties of Sets

Let $\mathbf{R}^p$ be the usual p-dimensional Euclidean space. A subset $A \subset \mathbf{R}^p$ is said to be *open* if for every $x \in A$, there is some $\epsilon > 0$ such that $y \in A$ whenever $\|x - y\| < \epsilon$. A subset A is said to be *closed* if its complement A^c in $\mathbf{R}^p$ is open. By the *interior* A° of any subset A we mean the largest open subset of A, possibly the empty set. The interior of A is found as the union of all open subsets of A.

A subset $A \subset \mathbf{R}^p$ is said to be *convex* if for every $x, y \in A$, the line segment with endpoints at x and y lies entirely in A. The *convex hull* of any given $A \subset \mathbf{R}^p$ is the smallest convex set that contains A. The convex hull of A is found as the intersection of all convex sets that contain the set A.

2.1.2 Affine Transformations

Let $\Lambda = (\Lambda_{jk})$ be a $q \times p$ matrix. By a *linear transformation* from $\mathbf{R}^p$ to $\mathbf{R}^q$ we shall mean a mapping of the form $x \to \Lambda x$, where x is a $p \times 1$ column vector. Linear transformations are special cases of *affine*

transformations, which have the general form $x \to \Lambda x + a$, where a is any $p \times 1$ column vector.

Suppose that $x_1, ..., x_{p+1}$ are $p+1$ points in $\mathbf{R}^p$. These points form the vertices of a *p-simplex* in $\mathbf{R}^p$, which can be defined as the convex hull of these points. Suppose $x_1, ..., x_{p+1}$ and $y_1, ..., y_{p+1}$ are the vertices of two p-simplexes with positive p-dimensional volume. Then there exists a unique affine transformation $\mathbf{R}^p \to \mathbf{R}^p$ of the form $x \to \Lambda x + a$ such that $y_j = \Lambda x_j + a$ for all $j = 1, 2, ..., p+1$.

2.1.3 *Orthogonal Transformations*

A $p \times p$ matrix $\Lambda = (\Lambda_{jk})$ is said to be *orthogonal* if $\Lambda^T = \Lambda^{-1}$, where Λ^T and Λ^{-1} denote the transpose and inverse matrices of Λ respectively. Equivalently, we can say that $\Lambda^T \Lambda = I$, where I is the $p \times p$ identity matrix. By an *orthogonal transformation* from $\mathbf{R}^p$ to itself we shall mean a linear transformation $x \to \Lambda x$ corresponding to multiplication of a p-dimensional column vector on the left by a $p \times p$ orthogonal matrix. For any orthogonal matrix Λ the determinant $\det(\Lambda) = \pm 1$. Those orthogonal matrices with $\det(\Lambda) = 1$ are said to be *special orthogonal matrices*, and their corresponding transformations of $\mathbf{R}^p$ are said to be *special orthogonal transformations*. Special orthogonal transformations can be regarded as generalizations into higher dimensions of the families of rotations about the origin in dimensions two and three. An example of an orthogonal transformation that is not a special orthogonal transformation is the reflection

$$(x_1, x_2, x_3, ..., x_p) \to (-x_1, x_2, x_3, ..., x_p) \tag{2.1}$$

of $\mathbf{R}^p$ through the hyperplane $x_1 = 0$.

Henceforth, we shall let $\mathbf{O}(p)$ and $\mathbf{SO}(p)$ denote the classes of orthogonal and special orthogonal transformations on $\mathbf{R}^p$ respectively.

2.1.4 *Unitary Transformations*

We now describe an analog to the class of orthogonal transformations on $\mathbf{R}^p$. Let $\mathbf{C}$ be the complex plane, and $\mathbf{C}^p$ the space of p-vectors whose entries are elements of $\mathbf{C}$. Linear transformations from $\mathbf{R}^p$ to $\mathbf{R}^p$ can be represented as $x \to \Lambda x$, where Λ is a $p \times p$ matrix of real entries. The complex analogs of these transformations are also of the form $x \to \Lambda x$, with the real entries of the column vector x and the matrix Λ replaced by complex values. These are linear transformations from $\mathbf{C}^p$ to $\mathbf{C}^p$.

Suppose $\Lambda = (\Lambda_{jk})$ is a $p \times p$ matrix of complex values. Let Λ^* be the $p \times p$ matrix whose (j,k)th entry is the complex conjugate of the (k,j)th entry of Λ. Then Λ is said to be a *unitary matrix* if $\Lambda^* \Lambda = I$, where I is the $p \times p$ identity matrix. A linear transformation $x \to \Lambda x$, where x is a column vector of p complex values and Λ is a $p \times p$ unitary matrix, is said to be a *unitary transformation* of $\mathbf{C}^p$. For any unitary matrix Λ the determinant $\det(\Lambda)$ is a complex number with modulus one. We say that Λ is a *special unitary matrix* provided $\det(\Lambda) = 1$.

Just as the complex plane $\mathbf{C}$ can be identified with $\mathbf{R}^2$, so the unitary transformations of $\mathbf{C}^p$ can be identified with particular orthogonal transformations of $\mathbf{R}^{2p}$. The $2p \times 2p$ matrix of real values corresponding to the unitary matrix Λ is found by replacing each complex entry Λ_{jk} by the 2×2 block of real values

$$\begin{pmatrix} \Re(\Lambda_{jk}) & -\Im(\Lambda_{jk}) \\ \Im(\Lambda_{jk}) & \Re(\Lambda_{jk}) \end{pmatrix} \tag{2.2}$$

where $\Re(z)$ and $\Im(z)$ are the real and imaginary parts of the complex number z respectively. Thus every unitary transformation of $\mathbf{C}^p$ can be regarded as an orthogonal transformation of $\mathbf{R}^{2p}$. Under this identification, the determinant of the $2p \times 2p$ orthogonal matrix will be the modulus of the determinant of its $p \times p$ unitary counterpart. While every unitary matrix or transformation can be identified with an orthogonal transformation, the converse is not true. This follows easily from the previous observation that the determinant of its $2p \times 2p$ orthogonal counterpart equals one, being the modulus of a complex number on the unit circle of $\mathbf{C}$. Thus reflections of $\mathbf{R}^{2p}$, and many other orthogonal transformations, cannot be represented as unitary transformations.

Henceforth, we shall let $\mathbf{U}(p)$ and $\mathbf{SU}(p)$ denote, respectively, the classes of unitary and special unitary transformations on $\mathbf{C}^p$.

2.1.5 *Singular Value Decompositions*

Let Λ be a matrix of dimension $q \times p$ that has rank r. Then $\Lambda\Lambda^T$ (or equivalently $\Lambda^T\Lambda$) has r nonzero eigenvalues. It is easy to check that the eigenvalues of $\Lambda\Lambda^T$ are nonnegative. Therefore we can write the eigenvalues as $\lambda_1^2, \lambda_2^2, ..., \lambda_r^2$. We define the matrix $\Gamma = (\Gamma_{jk})$ to be a $q \times p$ matrix for which $\Gamma_{jj} = |\lambda_j|$ for $j = 1, 2, ..., r$ and whose other elements are zero.

Then Λ can be written as

$$\Lambda = \Psi \Gamma \Psi' \tag{2.3}$$

where Ψ and Ψ' are orthogonal matrices of dimension $q \times q$ and $p \times p$ respectively. This decomposition is called a *singular value decomposition* of Λ. The eigenvalues $|\lambda_1|, ..., |\lambda_r|$ are called the *singular values* of the matrix Λ. Note that the singular value decomposition of Λ is not unique, although the set of singular values of Λ is uniquely determined.

to $1/5$ one of the four points will lie in the triangle formed by the other three.

5. In formula (1.21) we encountered the *Procrustean metric*. A *metric* $d(x, y)$ between points x, y of a set is a nonnegative real valued function satisfying
(i) $d(x, y) = 0$ if and only if $x = y$;
(ii) $d(x, y) = d(y, x)$ for all x and y;
(iii) $d(x, z) \leq d(x, y) + d(y, z)$ for all x, y, and z.
Show that the Procrustean metric d defined in Section 1.3 satisfies these properties on the set Σ_2^n.

2

Background Concepts and Definitions

2.1 Transformations on Euclidean Space

In this section, we shall begin with some preliminary definitions relevant to shape analysis.

2.1.1 Properties of Sets

Let $\mathbf{R}^p$ be the usual p-dimensional Euclidean space. A subset $A \subset \mathbf{R}^p$ is said to be *open* if for every $x \in A$, there is some $\epsilon > 0$ such that $y \in A$ whenever $||x - y|| < \epsilon$. A subset A is said to be *closed* if its complement A^c in $\mathbf{R}^p$ is open. By the *interior* A° of any subset A we mean the largest open subset of A, possibly the empty set. The interior of A is found as the union of all open subsets of A.

A subset $A \subset \mathbf{R}^p$ is said to be *convex* if for every $x, y \in A$, the line segment with endpoints at x and y lies entirely in A. The *convex hull* of any given $A \subset \mathbf{R}^p$ is the smallest convex set that contains A. The convex hull of A is found as the intersection of all convex sets that contain the set A.

2.1.2 Affine Transformations

Let $\Lambda = (\Lambda_{jk})$ be a $q \times p$ matrix. By a *linear transformation* from $\mathbf{R}^p$ to $\mathbf{R}^q$ we shall mean a mapping of the form $x \to \Lambda x$, where x is a $p \times 1$ column vector. Linear transformations are special cases of *affine*

transformations, which have the general form $x \to \Lambda x + a$, where a is any $p \times 1$ column vector.

Suppose that $x_1, ..., x_{p+1}$ are $p+1$ points in $\mathbf{R}^p$. These points form the vertices of a *p-simplex* in $\mathbf{R}^p$, which can be defined as the convex hull of these points. Suppose $x_1, ..., x_{p+1}$ and $y_1, ..., y_{p+1}$ are the vertices of two p-simplexes with positive p-dimensional volume. Then there exists a unique affine transformation $\mathbf{R}^p \to \mathbf{R}^p$ of the form $x \to \Lambda x + a$ such that $y_j = \Lambda x_j + a$ for all $j = 1, 2, ..., p+1$.

2.1.3 *Orthogonal Transformations*

A $p \times p$ matrix $\Lambda = (\Lambda_{jk})$ is said to be *orthogonal* if $\Lambda^T = \Lambda^{-1}$, where Λ^T and Λ^{-1} denote the transpose and inverse matrices of Λ respectively. Equivalently, we can say that $\Lambda^T \Lambda = I$, where I is the $p \times p$ identity matrix. By an *orthogonal transformation* from $\mathbf{R}^p$ to itself we shall mean a linear transformation $x \to \Lambda x$ corresponding to multiplication of a p-dimensional column vector on the left by a $p \times p$ orthogonal matrix. For any orthogonal matrix Λ the determinant $\det(\Lambda) = \pm 1$. Those orthogonal matrices with $\det(\Lambda) = 1$ are said to be *special orthogonal matrices*, and their corresponding transformations of $\mathbf{R}^p$ are said to be *special orthogonal transformations*. Special orthogonal transformations can be regarded as generalizations into higher dimensions of the families of rotations about the origin in dimensions two and three. An example of an orthogonal transformation that is not a special orthogonal transformation is the reflection

$$(x_1, x_2, x_3, ..., x_p) \to (-x_1, x_2, x_3, ..., x_p) \tag{2.1}$$

of $\mathbf{R}^p$ through the hyperplane $x_1 = 0$.

Henceforth, we shall let $\mathbf{O}(p)$ and $\mathbf{SO}(p)$ denote the classes of orthogonal and special orthogonal transformations on $\mathbf{R}^p$ respectively.

2.1.4 *Unitary Transformations*

We now describe an analog to the class of orthogonal transformations on $\mathbf{R}^p$. Let $\mathbf{C}$ be the complex plane, and $\mathbf{C}^p$ the space of p-vectors whose entries are elements of $\mathbf{C}$. Linear transformations from $\mathbf{R}^p$ to $\mathbf{R}^p$ can be represented as $x \to \Lambda x$, where Λ is a $p \times p$ matrix of real entries. The complex analogs of these transformations are also of the form $x \to \Lambda x$, with the real entries of the column vector x and the matrix Λ replaced by complex values. These are linear transformations from $\mathbf{C}^p$ to $\mathbf{C}^p$.

Suppose $\Lambda = (\Lambda_{jk})$ is a $p \times p$ matrix of complex values. Let Λ^* be the $p \times p$ matrix whose (j,k)th entry is the complex conjugate of the (k,j)th entry of Λ. Then Λ is said to be a *unitary matrix* if $\Lambda^* \Lambda = I$, where I is the $p \times p$ identity matrix. A linear transformation $x \to \Lambda x$, where x is a column vector of p complex values and Λ is a $p \times p$ unitary matrix, is said to be a *unitary transformation* of $\mathbf{C}^p$. For any unitary matrix Λ the determinant $\det(\Lambda)$ is a complex number with modulus one. We say that Λ is a *special unitary matrix* provided $\det(\Lambda) = 1$.

Just as the complex plane $\mathbf{C}$ can be identified with $\mathbf{R}^2$, so the unitary transformations of $\mathbf{C}^p$ can be identified with particular orthogonal transformations of $\mathbf{R}^{2p}$. The $2p \times 2p$ matrix of real values corresponding to the unitary matrix Λ is found by replacing each complex entry Λ_{jk} by the 2×2 block of real values

$$\begin{pmatrix} \Re(\Lambda_{jk}) & -\Im(\Lambda_{jk}) \\ \Im(\Lambda_{jk}) & \Re(\Lambda_{jk}) \end{pmatrix} \tag{2.2}$$

where $\Re(z)$ and $\Im(z)$ are the real and imaginary parts of the complex number z respectively. Thus every unitary transformation of $\mathbf{C}^p$ can be regarded as an orthogonal transformation of $\mathbf{R}^{2p}$. Under this identification, the determinant of the $2p \times 2p$ orthogonal matrix will be the modulus of the determinant of its $p \times p$ unitary counterpart. While every unitary matrix or transformation can be identified with an orthogonal transformation, the converse is not true. This follows easily from the previous observation that the determinant of its $2p \times 2p$ orthogonal counterpart equals one, being the modulus of a complex number on the unit circle of $\mathbf{C}$. Thus reflections of $\mathbf{R}^{2p}$, and many other orthogonal transformations, cannot be represented as unitary transformations.

Henceforth, we shall let $\mathbf{U}(p)$ and $\mathbf{SU}(p)$ denote, respectively, the classes of unitary and special unitary transformations on $\mathbf{C}^p$.

2.1.5 *Singular Value Decompositions*

Let Λ be a matrix of dimension $q \times p$ that has rank r. Then $\Lambda\Lambda^T$ (or equivalently $\Lambda^T\Lambda$) has r nonzero eigenvalues. It is easy to check that the eigenvalues of $\Lambda\Lambda^T$ are nonnegative. Therefore we can write the eigenvalues as $\lambda_1^2, \lambda_2^2, ..., \lambda_r^2$. We define the matrix $\Gamma = (\Gamma_{jk})$ to be a $q \times p$ matrix for which $\Gamma_{jj} = |\lambda_j|$ for $j = 1, 2, ..., r$ and whose other elements are zero.

Then Λ can be written as

$$\Lambda = \Psi \Gamma \Psi' \tag{2.3}$$

where Ψ and Ψ' are orthogonal matrices of dimension $q \times q$ and $p \times p$ respectively. This decomposition is called a *singular value decomposition* of Λ. The eigenvalues $|\lambda_1|, ..., |\lambda_r|$ are called the *singular values* of the matrix Λ. Note that the singular value decomposition of Λ is not unique, although the set of singular values of Λ is uniquely determined.

A case that will be of particular interest to us occurs when $p = q$ and Λ is of full rank. In this case, Γ is a square diagonal matrix whose diagonal elements are the singular values. The singular value decomposition allows us to represent a matrix in diagonal form, with Ψ and Ψ' serving to provide a change of coordinate systems for the purpose.

The singular value decomposition has an important geometric interpretation that will be of use in the next chapter. Suppose x is a 2×1 column vector and that Λ is a 2×2 matrix of full rank. Under the linear transformation $x \to \Lambda x$ the unit circle in the plane $\mathbf{R}^2$ is mapped to an ellipse. The lengths of the semimajor and semiminor axes of this ellipse are seen from equation (2.3) to be the singular values of Λ.

This geometric interpretation generalizes into higher dimensions. A $p \times p$ matrix Λ of full rank will have p singular values. If x is a $p \times 1$ column vector, then $x \to \Lambda x$ will map the unit sphere in $\mathbf{R}^p$ to an ellipsoid with p principal axes. The singular values of Λ can be seen to be one half the lengths of the principal axes of the ellipsoid.

2.1.6 *Inner Products*

The *inner product* between two elements $x = (x_1, ..., x_p)$ and $y = (y_1, ..., y_p)$ of $\mathbf{R}^p$ is defined as

$$< x, y > = \sum_{j=1}^{p} x_i y_j \tag{2.4}$$

Its complex counterpart for $\mathbf{C}^p$ is called the *Hermitian inner product*. We encountered the Hermitian inner product in Chapter 1 when defining the distance between two shapes in formula (1.21). We define the Hermitian inner product between two vectors $x = (x_1, ..., x_p)$ and $y = (y_1, ..., y_p)$ of complex coordinates to be

$$\ll x, y \gg = \sum_{j=1}^{p} x_j y_j^* \tag{2.5}$$

where y_j^* is the complex conjugate of $y_j \in \mathbf{C}$. Under the identification of $\mathbf{C}$ with $\mathbf{R}^2$ the inner product on $\mathbf{R}^{2p}$ can be defined from the Hermitian inner product on $\mathbf{C}^p$ by noting that $< ., . >= \Re \ll ., . \gg$.

Orthogonal transformations can be characterized as linear transformations that preserve inner products. Thus if $\Lambda = (\Lambda_{jk})$ is an orthogonal matrix, then representing $x, y \in \mathbf{R}^p$ as column vectors, we have

$$< \Lambda x, \ \Lambda y > = < x, y > \tag{2.6}$$

for all x and y in $\mathbf{R}^p$. Similarly, Hermitian inner products on $\mathbf{C}^p$ are preserved under unitary transformations.

2.1.7 *Groups of Transformations*

The classes $\mathbf{O}(p)$, $\mathbf{SO}(p)$, and their complex analogs $\mathbf{U}(p)$ and $\mathbf{SU}(p)$ are classes of transformations of a space to itself. We now summarize some definitions and properties of *groups*, of which these classes are examples.

Let h_1 and h_2 be any two transformations from $\mathbf{R}^p$ to $\mathbf{R}^p$. By the *composition* of h_1 and h_2 we shall mean the function $h_2 \circ h_1$ from $\mathbf{R}^p$ to $\mathbf{R}^p$ defined by

$$(h_2 \circ h_1)(x) \ = \ h_2[h_1(x)] \tag{2.7}$$

Suppose h is a 1-1 function that maps $\mathbf{R}^p$ onto itself. We shall let h^{-1} denote the *inverse function*, where $h^{-1}(y) = x$ whenever $h(x) = y$. There is nothing special about $\mathbf{R}^p$ in these definitions, as $\mathbf{R}^p$ can be replaced by $\mathbf{C}^p$ or any other set.

Definition 2.1.1. *A nonempty collection $\mathbf{H} = \{h\}$ of 1-1 transformations on a set is said to be a* group *provided that it is closed under composition and inversion of transformations.*

In order for a nonempty collection $\mathbf{H}$ of transformations on a set to be a group, it is necessary and sufficient that for any h_1, h_2 in $\mathbf{H}$ the transformation $h_1 \circ h_2^{-1}$ be in $\mathbf{H}$. Setting $h_1 = h_2$ we see that the identity transformation e is always an element of $\mathbf{H}$.

Definition 2.1.2. *Two transformations h_1 and h_2 are said to* commute *when $h_2 \circ h_1 = h_1 \circ h_2$. We say that a group $\mathbf{H}$ is* commutative *or* Abelian *provided that any two elements of $\mathbf{H}$ commute.*

By the *center* of a group $\mathbf{H}$ we mean the set of elements of $\mathbf{H}$ that commute with every other element of $\mathbf{H}$. Obviously, a group $\mathbf{H}$ is commutative if and only if the center of $\mathbf{H}$ is $\mathbf{H}$ itself.

The class of orthogonal matrices is closed under matrix multiplication as well as matrix inversion. Similarly, the class $\mathbf{O}(p)$ of orthogonal transformations is closed under function composition and function inversion. Thus the class of orthogonal transformations is a group, that is commutative only for the cases where $p = 1, 2$. The class $\mathbf{SO}(p)$ of special orthogonal transformations is a *subgroup* of the group of orthogonal transformations. That is, it is a subset of $\mathbf{O}(p)$ that is a group in its own right, being also closed under composition and inversion. When $p = 1$ this subgroup is the trivial group consisting of the identity transformation alone. Similar results hold true for the class of unitary transformations of $\mathbf{C}^p$. The class $\mathbf{U}(p)$ is also a group, containing the subgroup $\mathbf{SU}(p)$.

2.1.8 *Euclidean Motions and Isometries*

By a *Euclidean motion* of $\mathbf{R}^p$ we shall mean a transformation $h : \mathbf{R}^p \to \mathbf{R}^p$ that can be written as the composition of a special orthogonal transformation and a translation of $\mathbf{R}^p$. The class $\mathbf{Euc}(p)$ of Euclidean motions of $\mathbf{R}^p$ is a group and is commutative only for the case where $p = 1$. The group of Euclidean motions allows us to define the concept of *congruence* between subsets of $\mathbf{R}^p$. Two subsets A_1 and A_2 of $\mathbf{R}^p$ are said to be *congruent* if there exists a Euclidean motion $h \in \mathbf{Euc}(p)$ such that $h(A_1) = A_2$, or equivalently $h^{-1}(A_2) = A_1$.

The concept of congruence between sets forms the basis for Euclidean geometry, which involves the investigation of the geometric properties of subsets of Euclidean space $\mathbf{R}^p$. A property of a subset is said to be a *geometric property* if it is shared by any subset that is congruent to it.

The definition of a Euclidean motion of $\mathbf{R}^p$ can be generalized to arbitrary *metric spaces*. A metric space $\mathbf{M}$ is a set on which a metric $d(x, y)$ is defined, where d satisfies the abstract properties (i), (ii), and (iii) of Problem 5 in Chapter 1.

Definition 2.1.3. *A 1-1 correspondence $h : \mathbf{M} \to \mathbf{N}$ between metric spaces is said to be an* isometry *if $d(x, y) = d[h(x), h(y)]$ for all $x, y \in \mathbf{M}$. Two metric spaces are said to be* isometric *if there is an isometry mapping one to the other. When $\mathbf{M}$ and $\mathbf{N}$ are isometric, we shall write $\mathbf{M} \cong \mathbf{N}$.*

In particular, the class of all isometries from $\mathbf{M}$ to itself shall be denoted $\mathbf{Iso}(\mathbf{M})$. It is immediate that the identity transformation on $\mathbf{M}$ is an isometry, and it can be checked that the transformations of $\mathbf{Iso}(\mathbf{M})$ form a group. On $\mathbf{R}^p$, for example, the class of Euclidean motions $\mathbf{Euc}(p)$ forms a subgroup of $\mathbf{Iso}(\mathbf{R}^p)$. This subgroup is a proper subgroup, because the Eucidean motions of $\mathbf{R}^p$ do not include reflections through a $(p-1)$-dimensional hyperplane.

We may also speak of a *linear isometry* between vector spaces.

Definition 2.1.4. *A linear transformation of full rank between two vector spaces is said to be a* linear isometry *if it preserves the lengths of vectors.*

Clearly, an orthogonal rotation of $\mathbf{R}^p$ is an example of a linear isometry from $\mathbf{R}^p$ to itself.

2.1.9 *Similarity Transformations and the Shape of Sets*

Let $(x_1, ..., x_p)$ be an element of $\mathbf{R}^p$. A transformation $(x_1, ..., x_p) \to (\lambda x_1, ..., \lambda x_p)$, where $\lambda > 0$, is said to be an *isotropic rescaling* or simply a *rescaling* of $\mathbf{R}^p$. By a *shape-preserving transformation* or a *similarity transformation* of $\mathbf{R}^p$, we shall mean a transformation that can be represented as the composition of a rigid Euclidean motion and a rescaling of $\mathbf{R}^p$. Once again, it can be checked that the class of similarity transformations forms a group under composition. Henceforth, we shall denote the class of similarity transformations of $\mathbf{R}^p$ by $\mathbf{Sim}(p)$.

The group of similarity transformations has a special representation when $p = 1, 2$. In these cases, additional algebraic structure is available from multiplication of real and complex numbers respectively. In the latter case, we can again identify $\mathbf{R}^2$ with the complex plane $\mathbf{C}$. Then we can write transformations in $\mathbf{Sim}(1)$ and $\mathbf{Sim}(2)$ in the form $x \to ax + b$, where $a \neq 0$ and b are arbitrary elements of $\mathbf{R}$ or $\mathbf{C}$ in the respective dimensions. Multiplication and addition are the usual algebraic operations.

Just as the group of Euclidean motions leads to the concept of congruence between sets, so the group of similarity transformations leads to the concept of *similar* sets.

Definition 2.1.5. *Two subsets A_1 and A_2 of $\mathbf{R}^p$ are said to be* similar *or to have the same* shape *if there exists a similarity transformation $h \in \mathbf{Sim}(p)$ such that $h(A_1) = A_2$ or equivalently if $h^{-1}(A_2) = A_1$. If A_1 and A_2 are similar, then we shall write $A_1 \sim A_2$.*

We shall also be concerned with labeled figures or sets. For example, a triangle is often labeled at its vertices in Euclidean geometry in order to compare corresponding points on different triangles or simply to clarify a construction. The definitions of congruent and similar sets have obvious extensions to labeled sets, provided the labels correspond.

Definition 2.1.6. *We shall say that two correspondingly labeled sets have the same shape if one set can be transformed by a similarity transformation to the other set in such a way that labeled points are mapped to the corresponding points of the other figure.*

For example, two triangles $x_1x_2x_3$ and $y_1y_2y_3$ have the same shape if the angle at vertex x_j equals the angle at y_j for $j = 1$, 2, and 3.

While the distinction between labeled and unlabeled sets can be regarded as a mathematical convenience in defining shapes, it is a more substantial distinction for the comparison of shape differences, as we noted in Chapter 1. An attempt to discover the shape differences between sets will typically involve a matching of the sets to determine how differences in the coordinates of corresponding points can be explained through similarity transformations. Any residual differences that cannot be explained through similarity transformations can be understood to be due to differences in shape. The problem of constructing an appropriate correspondence between unlabeled sets (or unparametrized sets in general) is the *problem of homology*,

A case that will be of particular interest to us occurs when $p = q$ and Λ is of full rank. In this case, Γ is a square diagonal matrix whose diagonal elements are the singular values. The singular value decomposition allows us to represent a matrix in diagonal form, with Ψ and Ψ' serving to provide a change of coordinate systems for the purpose.

The singular value decomposition has an important geometric interpretation that will be of use in the next chapter. Suppose x is a 2×1 column vector and that Λ is a 2×2 matrix of full rank. Under the linear transformation $x \to \Lambda x$ the unit circle in the plane $\mathbf{R}^2$ is mapped to an ellipse. The lengths of the semimajor and semiminor axes of this ellipse are seen from equation (2.3) to be the singular values of Λ.

This geometric interpretation generalizes into higher dimensions. A $p \times p$ matrix Λ of full rank will have p singular values. If x is a $p \times 1$ column vector, then $x \to \Lambda x$ will map the unit sphere in $\mathbf{R}^p$ to an ellipsoid with p principal axes. The singular values of Λ can be seen to be one half the lengths of the principal axes of the ellipsoid.

2.1.6 *Inner Products*

The *inner product* between two elements $x = (x_1, ..., x_p)$ and $y = (y_1, ..., y_p)$ of $\mathbf{R}^p$ is defined as

$$< x, y > = \sum_{j=1}^{p} x_i y_j \tag{2.4}$$

Its complex counterpart for $\mathbf{C}^p$ is called the *Hermitian inner product.* We encountered the Hermitian inner product in Chapter 1 when defining the distance between two shapes in formula (1.21). We define the Hermitian inner product between two vectors $x = (x_1, ..., x_p)$ and $y = (y_1, ..., y_p)$ of complex coordinates to be

$$\ll x, y \gg = \sum_{j=1}^{p} x_j y_j^* \tag{2.5}$$

where y_j^* is the complex conjugate of $y_j \in \mathbf{C}$. Under the identification of $\mathbf{C}$ with $\mathbf{R}^2$ the inner product on $\mathbf{R}^{2p}$ can be defined from the Hermitian inner product on $\mathbf{C}^p$ by noting that $< ., . >= \Re \ll ., . \gg$.

Orthogonal transformations can be characterized as linear transformations that preserve inner products. Thus if $\Lambda = (\Lambda_{jk})$ is an orthogonal matrix, then representing $x, y \in \mathbf{R}^p$ as column vectors, we have

$$< \Lambda x, \ \Lambda y > = < x, y > \tag{2.6}$$

for all x and y in $\mathbf{R}^p$. Similarly, Hermitian inner products on $\mathbf{C}^p$ are preserved under unitary transformations.

2.1.7 *Groups of Transformations*

The classes $\mathbf{O}(p)$, $\mathbf{SO}(p)$, and their complex analogs $\mathbf{U}(p)$ and $\mathbf{SU}(p)$ are classes of transformations of a space to itself. We now summarize some definitions and properties of *groups*, of which these classes are examples.

Let h_1 and h_2 be any two transformations from $\mathbf{R}^p$ to $\mathbf{R}^p$. By the *composition* of h_1 and h_2 we shall mean the function $h_2 \circ h_1$ from $\mathbf{R}^p$ to $\mathbf{R}^p$ defined by

$$(h_2 \circ h_1)(x) \ = \ h_2[h_1(x)] \tag{2.7}$$

Suppose h is a 1-1 function that maps $\mathbf{R}^p$ onto itself. We shall let h^{-1} denote the *inverse function*, where $h^{-1}(y) = x$ whenever $h(x) = y$. There is nothing special about $\mathbf{R}^p$ in these definitions, as $\mathbf{R}^p$ can be replaced by $\mathbf{C}^p$ or any other set.

Definition 2.1.1. *A nonempty collection $\mathbf{H} = \{h\}$ of 1-1 transformations on a set is said to be a* group *provided that it is closed under composition and inversion of transformations.*

In order for a nonempty collection $\mathbf{H}$ of transformations on a set to be a group, it is necessary and sufficient that for any h_1, h_2 in $\mathbf{H}$ the transformation $h_1 \circ h_2^{-1}$ be in $\mathbf{H}$. Setting $h_1 = h_2$ we see that the identity transformation e is always an element of $\mathbf{H}$.

Definition 2.1.2. *Two transformations h_1 and h_2 are said to* commute *when $h_2 \circ h_1 = h_1 \circ h_2$. We say that a group $\mathbf{H}$ is* commutative *or* Abelian *provided that any two elements of $\mathbf{H}$ commute.*

By the *center* of a group $\mathbf{H}$ we mean the set of elements of $\mathbf{H}$ that commute with every other element of $\mathbf{H}$. Obviously, a group $\mathbf{H}$ is commutative if and only if the center of $\mathbf{H}$ is $\mathbf{H}$ itself.

The class of orthogonal matrices is closed under matrix multiplication as well as matrix inversion. Similarly, the class $\mathbf{O}(p)$ of orthogonal transformations is closed under function composition and function inversion. Thus the class of orthogonal transformations is a group, that is commutative only for the cases where $p = 1, 2$. The class $\mathbf{SO}(p)$ of special orthogonal transformations is a *subgroup* of the group of orthogonal transformations. That is, it is a subset of $\mathbf{O}(p)$ that is a group in its own right, being also closed under composition and inversion. When $p = 1$ this subgroup is the trivial group consisting of the identity transformation alone. Similar results hold true for the class of unitary transformations of $\mathbf{C}^p$. The class $\mathbf{U}(p)$ is also a group, containing the subgroup $\mathbf{SU}(p)$.

2.1.8 *Euclidean Motions and Isometries*

By a *Euclidean motion* of $\mathbf{R}^p$ we shall mean a transformation $h : \mathbf{R}^p \to \mathbf{R}^p$ that can be written as the composition of a special orthogonal transformation and a translation of $\mathbf{R}^p$. The class $\mathbf{Euc}(p)$ of Euclidean motions of $\mathbf{R}^p$ is a group and is commutative only for the case where $p = 1$. The group of Euclidean motions allows us to define the concept of *congruence* between subsets of $\mathbf{R}^p$. Two subsets A_1 and A_2 of $\mathbf{R}^p$ are said to be *congruent* if there exists a Euclidean motion $h \in \mathbf{Euc}(p)$ such that $h(A_1) = A_2$, or equivalently $h^{-1}(A_2) = A_1$.

The concept of congruence between sets forms the basis for Euclidean geometry, which involves the investigation of the geometric properties of subsets of Euclidean space $\mathbf{R}^p$. A property of a subset is said to be a *geometric property* if it is shared by any subset that is congruent to it.

The definition of a Euclidean motion of $\mathbf{R}^p$ can be generalized to arbitrary *metric spaces*. A metric space $\mathbf{M}$ is a set on which a metric $d(x, y)$ is defined, where d satisfies the abstract properties (i), (ii), and (iii) of Problem 5 in Chapter 1.

Definition 2.1.3. *A 1-1 correspondence $h : \mathbf{M} \to \mathbf{N}$ between metric spaces is said to be an* isometry *if $d(x, y) = d[h(x), h(y)]$ for all $x, y \in \mathbf{M}$. Two metric spaces are said to be* isometric *if there is an isometry mapping one to the other. When $\mathbf{M}$ and $\mathbf{N}$ are isometric, we shall write $\mathbf{M} \cong \mathbf{N}$.*

In particular, the class of all isometries from $\mathbf{M}$ to itself shall be denoted $\mathbf{Iso}(\mathbf{M})$. It is immediate that the identity transformation on $\mathbf{M}$ is an isometry, and it can be checked that the transformations of $\mathbf{Iso}(\mathbf{M})$ form a group. On $\mathbf{R}^p$, for example, the class of Euclidean motions $\mathbf{Euc}(p)$ forms a subgroup of $\mathbf{Iso}(\mathbf{R}^p)$. This subgroup is a proper subgroup, because the Eucidean motions of $\mathbf{R}^p$ do not include reflections through a $(p-1)$-dimensional hyperplane.

We may also speak of a *linear isometry* between vector spaces.

Definition 2.1.4. *A linear transformation of full rank between two vector spaces is said to be a* linear isometry *if it preserves the lengths of vectors.*

Clearly, an orthogonal rotation of $\mathbf{R}^p$ is an example of a linear isometry from $\mathbf{R}^p$ to itself.

2.1.9 *Similarity Transformations and the Shape of Sets*

Let $(x_1, ..., x_p)$ be an element of $\mathbf{R}^p$. A transformation $(x_1, ..., x_p) \to (\lambda x_1, ..., \lambda x_p)$, where $\lambda > 0$, is said to be an *isotropic rescaling* or simply a *rescaling* of $\mathbf{R}^p$. By a *shape-preserving transformation* or a *similarity transformation* of $\mathbf{R}^p$, we shall mean a transformation that can be represented as the composition of a rigid Euclidean motion and a rescaling of $\mathbf{R}^p$. Once again, it can be checked that the class of similarity transformations forms a group under composition. Henceforth, we shall denote the class of similarity transformations of $\mathbf{R}^p$ by $\mathbf{Sim}(p)$.

The group of similarity transformations has a special representation when $p = 1, 2$. In these cases, additional algebraic structure is available from multiplication of real and complex numbers respectively. In the latter case, we can again identify $\mathbf{R}^2$ with the complex plane $\mathbf{C}$. Then we can write transformations in $\mathbf{Sim}(1)$ and $\mathbf{Sim}(2)$ in the form $x \to ax + b$, where $a \neq 0$ and b are arbitrary elements of $\mathbf{R}$ or $\mathbf{C}$ in the respective dimensions. Multiplication and addition are the usual algebraic operations.

Just as the group of Euclidean motions leads to the concept of congruence between sets, so the group of similarity transformations leads to the concept of *similar* sets.

Definition 2.1.5. *Two subsets A_1 and A_2 of $\mathbf{R}^p$ are said to be* similar *or to have the same* shape *if there exists a similarity transformation $h \in \mathbf{Sim}(p)$ such that $h(A_1) = A_2$ or equivalently if $h^{-1}(A_2) = A_1$. If A_1 and A_2 are similar, then we shall write $A_1 \sim A_2$.*

We shall also be concerned with labeled figures or sets. For example, a triangle is often labeled at its vertices in Euclidean geometry in order to compare corresponding points on different triangles or simply to clarify a construction. The definitions of congruent and similar sets have obvious extensions to labeled sets, provided the labels correspond.

Definition 2.1.6. *We shall say that two correspondingly labeled sets have the same shape if one set can be transformed by a similarity transformation to the other set in such a way that labeled points are mapped to the corresponding points of the other figure.*

For example, two triangles $x_1x_2x_3$ and $y_1y_2y_3$ have the same shape if the angle at vertex x_j equals the angle at y_j for $j = 1$, 2, and 3.

While the distinction between labeled and unlabeled sets can be regarded as a mathematical convenience in defining shapes, it is a more substantial distinction for the comparison of shape differences, as we noted in Chapter 1. An attempt to discover the shape differences between sets will typically involve a matching of the sets to determine how differences in the coordinates of corresponding points can be explained through similarity transformations. Any residual differences that cannot be explained through similarity transformations can be understood to be due to differences in shape. The problem of constructing an appropriate correspondence between unlabeled sets (or unparametrized sets in general) is the *problem of homology*,

discussed in Section 1.5.

2.2 Differential Geometry

2.2.1 *Homeomorphisms and Diffeomorphisms of Euclidean Space*

Let $h: U \to V$ be a continuous function between two open sets $U \subset \mathbf{R}^p$ and $V \subset \mathbf{R}^q$. Let us write $(y_1, y_2, ..., y_q) = h(x_1, x_2, ..., x_p)$. We say that h is a *smooth*, or *differentiable*, mapping on U provided that h possesses finite partial derivatives $\partial y_k / \partial x_j$ for all $j = 1, ..., p$ and all $k = 1, ..., q$. If all these partial derivatives are continuous functions, then we say that h is a $\mathcal{C}^1$-function on U.

This definition can be extended to higher order derivatives. We say that h is a $\mathcal{C}^r$-function on U for any $r = 1, 2, ...$ if h has continuous partial derivatives

$$\frac{\partial^{r_1+r_2+...+r_p} y_k}{\partial x_1^{r_1} \partial x_2^{r_2} ... \partial x_p^{r_p}} \tag{2.8}$$

for all $k = 1, ..., q$ and all nonnegative integers $r_1, r_2, ..., r_p$ such that $r_1+r_2+...+r_p \leq r$. Clearly, any function that is $\mathcal{C}^r$ on U is a $\mathcal{C}^s$-function for any $s < r$.

If h is a $\mathcal{C}^r$-function for all $r \geq 1$, then we say that h is a $\mathcal{C}^\infty$-function. By convention, $\mathcal{C}^0$-functions are understood to be the class of continuous functions on U.

Associated with any smooth function $h: U \to V$ and any point $x = (x_1, ..., x_p)$ in U is the *Jacobian matrix*. This is the matrix of partial derivatives

$$\Lambda = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_p} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \frac{\partial y_q}{\partial x_1} & \cdots & \frac{\partial y_q}{\partial x_p} \end{pmatrix} \tag{2.9}$$

It defines a linear transformation $u \to \Lambda u$ where u is a $p \times 1$ column vector. This linear transformation

$$(\mathcal{D}h)_x \ : \ \mathbf{R}^p \to \mathbf{R}^q \tag{2.10}$$

is called the *derivative of* h *at* x. The Jacobian matrix can be regarded as a coordinate representation of the derivative of h. The derivative $\mathcal{D}h$ is the second term in the Taylor approximation to the function h at x, namely

$$h(x+u) \ = \ h(x) \ + \ (\mathcal{D}h)_x(u) \ + \ o(||u||) \tag{2.11}$$

When $p = q$, the Jacobian matrix becomes a $p \times p$ square matrix. The determinant

$$(\mathcal{J}h)_x \ = \ \det(\Lambda) \tag{2.12}$$

is simply called the *Jacobian* of h at x, at the risk of some confusion. Note that the Jacobian matrix is a matrix valued function at each point $x \in U$ while the Jacobian at x is a real valued function. The Jacobian measures the rate of change of volume induced by the transformation $x \to h(x)$ locally around x.

Suppose that $p = q$ and that h is a 1-1 correspondence from U to V. Then h is said to be a *homeomorphism* from U to V provided that both h and h^{-1} are continuous. When a homeomorphism can be established between U and V we say that U and V are *homeomorphic*. A homeomorphism h is called a $\mathcal{C}^r$-*diffeomorphism* between U and V if both h and h^{-1} are $\mathcal{C}^r$-functions. We will normally refer to a $\mathcal{C}^\infty$-diffeomorphism simply as a *diffeomorphism*. When a diffeomorphism can be established between U and V we shall say that U and V are *diffeomorphic*.

2.2.2 *Topological Spaces*

The properties of continuity and differentiability on $\mathbf{R}^p$ can be abstracted to more general sets, leading to the concepts of the *topological space*, the *topological manifold*, and the *differential manifold*. Suppose $\mathbf{M}$ is a set endowed with a collection of subsets $\mathcal{U} = \{U\}$. We say that $\mathcal{U}$ is a *topology* on the set $\mathbf{M}$ provided that (i) the empty set and $\mathbf{M}$ itself are both elements of $\mathcal{U}$, (ii) any arbitrary union of elements of $\mathcal{U}$ is an element of $\mathcal{U}$, and (iii) any finite intersection of elements of $\mathcal{U}$ is an element of $\mathcal{U}$.

The set $\mathbf{M}$, endowed with a topology, is called a *topological space*, and the elements of $\mathcal{U}$ are called the *open sets* of $\mathbf{M}$. A subset of $\mathbf{M}$ is said to be *closed* if its complement is open. The standard example of a topological space, which we have already considered, is when $\mathbf{M}$ is Euclidean space $\mathbf{R}^p$ and $\mathcal{U}$ is the class of open sets of $\mathbf{R}^p$. Let $\mathbf{M}$ and $\mathbf{N}$ be topological spaces endowed with topologies $\mathcal{U}_1$ and $\mathcal{U}_2$ respectively. A function $h: \mathbf{M} \to \mathbf{N}$ is said to be *continuous* if $h^{-1}(U) \in \mathcal{U}_1$ for all $U \in \mathcal{U}_2$. If h is both 1-1 and onto, then we say that h is a homeomorphism provided that both h and h^{-1} are continuous.

In $\mathbf{R}^p$ a subset that is both closed and bounded has the property of *compactness*. This can be generalized to an arbitrary topological space. A subset A of a topological space $\mathbf{M}$ is said to be *compact* if every collection of open sets whose union contains A has a finite subcollection whose union also contains A. The *Heine-Borel theorem* states that a subset of $\mathbf{R}^p$ is compact if and only if it is closed and bounded. For our purposes in this and subsequent chapters, only a few properties of compactness will be used. Important among these properties is the fact that the continuous image of

a compact set is compact. More specifically, if $\mathbf{M}$ and $\mathbf{N}$ are topological spaces and $h : \mathbf{M} \to \mathbf{N}$ is a continuous function then $h(A)$ is compact for all compact subsets $A \subset \mathbf{M}$.

2.2.3 *Introduction to Manifolds*

A manifold is a generalization of our understanding of a curved surface in three dimensions. We usually think of a curved surface as a subset of three-dimensional Euclidean space $\mathbf{R}^3$ that inherits its geometric properties from the geometric structure of the Euclidean space in which it lies.

The representation of a space as a subset of another space is formally called an *embedding*. However, our intuition, being limited to objects in dimensions less than or equal to three, has trouble visualizing curvature of sets or spaces that cannot be embedded in three-dimensional Euclidean space. The formal definition of a differential manifold has no such constraint. As much of calculus involves local constructions, differential manifolds, which locally resemble Euclidean space, become a natural domain for operations such as taking a gradient of a function, calculating tangent vectors, and other constructions from multivariable calculus.

Examples of differential manifolds are common. A torus (the surface of a doughnut) is a differential manifold, as is a sphere or a flat plane. Some very small two-dimensional being situated in a torus would have trouble distinguishing the space around it from the space of a two-dimensional sphere or a plane. This is because curved surfaces look approximately flat when viewed over a small region. The immediate vicinity of the being provides local information about the surface but little in the way of information about global properties of the surface that distinguish spheres from tori. To find global information, the being would have the walk around both surfaces and be very careful to check angles and distances. If the being were nearsighted and could not check distances and angles, then its examination of the local vicinity, or neighborhood, would fail to detect any local distortions due to the curvature of the surface. It might then conclude that the surrounding space was Euclidean, or flat, in nature. This is what we mean when we say that a differential manifold looks locally like $\mathbf{R}^p$.

Our two-dimensional being might well consult an *atlas* to find its way around the geography of these two-dimensional worlds. We are used to seeing the surface of the Earth displayed in an atlas. However, we know that because the Earth is a sphere, we cannot get all points plotted on a single page or *chart* without tearing the picture and destroying the natural continuity between neighboring points. Just as a portion of the surface of the Earth can be described by a chart, so a portion of a differential manifold is described by a chart, here understood in a mathematical sense. Just as a single page of an atlas cannot cover the entire surface of the Earth without disrupting continuity, so a single chart cannot usually cover the entire region of a differential manifold. The mathematical charts used to

describe a manifold must also be collected together into an atlas. Of course, such charts, if they cover the manifold, will overlap in places. Thus they are not arbitrarily related, but must, in a certain sense, describe the same smoothness on the region of overlap. If the same town appeared on two different pages of a geographical atlas, we would expect the local descriptions on the two pages to be compatible, even if not identical. On a differential manifold, that notion of compatibility is described using a *diffeomorphism.*

2.2.4 *Topological and Differential Manifolds*

Let $\mathbf{M}^p$ be a topological space with a collection of open subsets

$$\{U_\alpha \ : \ \alpha \in A\} \tag{2.13}$$

such that

$$\bigcup_{\alpha \in A} U_\alpha \ = \ \mathbf{M}^p \tag{2.14}$$

and a collection of functions

$$c_\alpha : U_\alpha \to \mathbf{R}^p \tag{2.15}$$

that are all homeomorphisms onto the open subsets $h(U_\alpha)$ of $\mathbf{R}^p$. Note that we do *not* assume $\{U_\alpha \ : \ \alpha \in A\}$ is the entire topology on $\mathbf{M}^p$. Then we say that the functions c_α are *charts* on $\mathbf{M}^p$ provided that

$$c_\beta \circ c_\alpha^{-1} \ : \ c_\alpha(U_\alpha \cap U_\beta) \to c_\beta(U_\alpha \cap U_\beta) \tag{2.16}$$

is a homeomorphism from $c_\alpha(U_\alpha \cap U_\beta)$ to $c_\beta(U_\alpha \cap U_\beta)$ for all α and β in A. See Figure 2.1. We can think of the charts $\{c_\alpha\}_{\alpha \in A}$ as providing local coordinate systems on $\mathbf{M}^p$. Formula (2.16) provides a *patching* criterion, telling us that these different coordinate systems can be glued together in a topologically consistent way.

Definition 2.2.1. *The collection of subsets* $\{U_\alpha\}_{\alpha \in A}$ *with the charts* $\{c_\alpha\}_{\alpha \in A}$ *is said to form an* atlas *on* $\mathbf{M}^p$. *The set* $\mathbf{M}^p$ *together with its atlas* $\{(U_\alpha, c_\alpha) : \alpha \in A\}$ *is called a* topological manifold *of dimension* p.

A subset $V \subset \mathbf{M}^p$ is open if $c_\alpha(V \cap U_\alpha)$ is an open subset of $\mathbf{R}^p$ for every $\alpha \in A$. This definition formalizes our basic understanding that a topological manifold is a space that is locally homeomorphic to Euclidean space.

Definition 2.2.2. *If the functions* $c_\beta \circ c_\alpha^{-1}$ *in (2.16) are also required to be* C^r*-diffeomorphisms then the topological manifold* $\mathbf{M}^p$ *is said to be a* C^r-differential manifold.

discussed in Section 1.5.

2.2 Differential Geometry

2.2.1 *Homeomorphisms and Diffeomorphisms of Euclidean Space*

Let $h : U \to V$ be a continuous function between two open sets $U \subset \mathbf{R}^p$ and $V \subset \mathbf{R}^q$. Let us write $(y_1, y_2, ..., y_q) = h(x_1, x_2, ..., x_p)$. We say that h is a *smooth*, or *differentiable*, mapping on U provided that h possesses finite partial derivatives $\partial y_k / \partial x_j$ for all $j = 1, ..., p$ and all $k = 1, ..., q$. If all these partial derivatives are continuous functions, then we say that h is a C^1-function on U.

This definition can be extended to higher order derivatives. We say that h is a C^r-function on U for any $r = 1, 2, ...$ if h has continuous partial derivatives

$$\frac{\partial^{r_1+r_2+...+r_p} y_k}{\partial x_1^{r_1} \partial x_2^{r_2} ... \partial x_p^{r_p}} \tag{2.8}$$

for all $k = 1, ..., q$ and all nonnegative integers $r_1, r_2, ..., r_p$ such that $r_1 + r_2 + ... + r_p \leq r$. Clearly, any function that is C^r on U is a C^s-function for any $s < r$.

If h is a C^r-function for all $r \geq 1$, then we say that h is a C^∞-function. By convention, C^0-functions are understood to be the class of continuous functions on U.

Associated with any smooth function $h : U \to V$ and any point $x = (x_1, ..., x_p)$ in U is the *Jacobian matrix*. This is the matrix of partial derivatives

$$\Lambda = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_p} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \frac{\partial y_q}{\partial x_1} & \cdots & \frac{\partial y_q}{\partial x_p} \end{pmatrix} \tag{2.9}$$

It defines a linear transformation $u \to \Lambda u$ where u is a $p \times 1$ column vector. This linear transformation

$$(\mathcal{D}h)_x \ : \ \mathbf{R}^p \to \mathbf{R}^q \tag{2.10}$$

is called the *derivative of* h *at* x. The Jacobian matrix can be regarded as a coordinate representation of the derivative of h. The derivative $\mathcal{D}h$ is the second term in the Taylor approximation to the function h at x, namely

$$h(x + u) \ = \ h(x) \ + \ (\mathcal{D}h)_x(u) \ + \ o(||u||) \tag{2.11}$$

When $p = q$, the Jacobian matrix becomes a $p \times p$ square matrix. The determinant

$$(\mathcal{J}h)_x \ = \ \det(\Lambda) \tag{2.12}$$

is simply called the *Jacobian* of h at x, at the risk of some confusion. Note that the Jacobian matrix is a matrix valued function at each point $x \in U$ while the Jacobian at x is a real valued function. The Jacobian measures the rate of change of volume induced by the transformation $x \to h(x)$ locally around x.

Suppose that $p = q$ and that h is a 1-1 correspondence from U to V. Then h is said to be a *homeomorphism* from U to V provided that both h and h^{-1} are continuous. When a homeomorphism can be established between U and V we say that U and V are *homeomorphic*. A homeomorphism h is called a *C^r-diffeomorphism* between U and V if both h and h^{-1} are C^r-functions. We will normally refer to a C^∞-diffeomorphism simply as a *diffeomorphism*. When a diffeomorphism can be established between U and V we shall say that U and V are *diffeomorphic*.

2.2.2 *Topological Spaces*

The properties of continuity and differentiability on $\mathbf{R}^p$ can be abstracted to more general sets, leading to the concepts of the *topological space*, the *topological manifold*, and the *differential manifold*. Suppose $\mathbf{M}$ is a set endowed with a collection of subsets $\mathcal{U} = \{U\}$. We say that $\mathcal{U}$ is a *topology* on the set $\mathbf{M}$ provided that (i) the empty set and $\mathbf{M}$ itself are both elements of $\mathcal{U}$, (ii) any arbitrary union of elements of $\mathcal{U}$ is an element of $\mathcal{U}$, and (iii) any finite intersection of elements of $\mathcal{U}$ is an element of $\mathcal{U}$.

The set $\mathbf{M}$, endowed with a topology, is called a *topological space*, and the elements of $\mathcal{U}$ are called the *open sets* of $\mathbf{M}$. A subset of $\mathbf{M}$ is said to be *closed* if its complement is open. The standard example of a topological space, which we have already considered, is when $\mathbf{M}$ is Euclidean space $\mathbf{R}^p$ and $\mathcal{U}$ is the class of open sets of $\mathbf{R}^p$. Let $\mathbf{M}$ and $\mathbf{N}$ be topological spaces endowed with topologies $\mathcal{U}_1$ and $\mathcal{U}_2$ respectively. A function $h : \mathbf{M} \to \mathbf{N}$ is said to be *continuous* if $h^{-1}(U) \in \mathcal{U}_1$ for all $U \in \mathcal{U}_2$. If h is both 1-1 and onto, then we say that h is a homeomorphism provided that both h and h^{-1} are continuous.

In $\mathbf{R}^p$ a subset that is both closed and bounded has the property of *compactness*. This can be generalized to an arbitrary topological space. A subset A of a topological space $\mathbf{M}$ is said to be *compact* if every collection of open sets whose union contains A has a finite subcollection whose union also contains A. The *Heine-Borel theorem* states that a subset of $\mathbf{R}^p$ is compact if and only if it is closed and bounded. For our purposes in this and subsequent chapters, only a few properties of compactness will be used. Important among these properties is the fact that the continuous image of

a compact set is compact. More specifically, if $\mathbf{M}$ and $\mathbf{N}$ are topological spaces and $h : \mathbf{M} \to \mathbf{N}$ is a continuous function then $h(A)$ is compact for all compact subsets $A \subset \mathbf{M}$.

2.2.3 *Introduction to Manifolds*

A manifold is a generalization of our understanding of a curved surface in three dimensions. We usually think of a curved surface as a subset of three-dimensional Euclidean space $\mathbf{R}^3$ that inherits its geometric properties from the geometric structure of the Euclidean space in which it lies.

The representation of a space as a subset of another space is formally called an *embedding*. However, our intuition, being limited to objects in dimensions less than or equal to three, has trouble visualizing curvature of sets or spaces that cannot be embedded in three-dimensional Euclidean space. The formal definition of a differential manifold has no such constraint. As much of calculus involves local constructions, differential manifolds, which locally resemble Euclidean space, become a natural domain for operations such as taking a gradient of a function, calculating tangent vectors, and other constructions from multivariable calculus.

Examples of differential manifolds are common. A torus (the surface of a doughnut) is a differential manifold, as is a sphere or a flat plane. Some very small two-dimensional being situated in a torus would have trouble distinguishing the space around it from the space of a two-dimensional sphere or a plane. This is because curved surfaces look approximately flat when viewed over a small region. The immediate vicinity of the being provides local information about the surface but little in the way of information about global properties of the surface that distinguish spheres from tori. To find global information, the being would have the walk around both surfaces and be very careful to check angles and distances. If the being were nearsighted and could not check distances and angles, then its examination of the local vicinity, or neighborhood, would fail to detect any local distortions due to the curvature of the surface. It might then conclude that the surrounding space was Euclidean, or flat, in nature. This is what we mean when we say that a differential manifold looks locally like $\mathbf{R}^p$.

Our two-dimensional being might well consult an *atlas* to find its way around the geography of these two-dimensional worlds. We are used to seeing the surface of the Earth displayed in an atlas. However, we know that because the Earth is a sphere, we cannot get all points plotted on a single page or *chart* without tearing the picture and destroying the natural continuity between neighboring points. Just as a portion of the surface of the Earth can be described by a chart, so a portion of a differential manifold is described by a chart, here understood in a mathematical sense. Just as a single page of an atlas cannot cover the entire surface of the Earth without disrupting continuity, so a single chart cannot usually cover the entire region of a differential manifold. The mathematical charts used to

describe a manifold must also be collected together into an atlas. Of course, such charts, if they cover the manifold, will overlap in places. Thus they are not arbitrarily related, but must, in a certain sense, describe the same smoothness on the region of overlap. If the same town appeared on two different pages of a geographical atlas, we would expect the local descriptions on the two pages to be compatible, even if not identical. On a differential manifold, that notion of compatibility is described using a *diffeomorphism.*

2.2.4 *Topological and Differential Manifolds*

Let $\mathbf{M}^p$ be a topological space with a collection of open subsets

$$\{U_\alpha \ : \ \alpha \in A\} \tag{2.13}$$

such that

$$\bigcup_{\alpha \in A} U_\alpha \ = \ \mathbf{M}^p \tag{2.14}$$

and a collection of functions

$$c_\alpha : U_\alpha \to \mathbf{R}^p \tag{2.15}$$

that are all homeomorphisms onto the open subsets $h(U_\alpha)$ of $\mathbf{R}^p$. Note that we do *not* assume $\{U_\alpha \ : \ \alpha \in A\}$ is the entire topology on $\mathbf{M}^p$. Then we say that the functions c_α are *charts* on $\mathbf{M}^p$ provided that

$$c_\beta \circ c_\alpha^{-1} \ : \ c_\alpha(U_\alpha \cap U_\beta) \to c_\beta(U_\alpha \cap U_\beta) \tag{2.16}$$

is a homeomorphism from $c_\alpha(U_\alpha \cap U_\beta)$ to $c_\beta(U_\alpha \cap U_\beta)$ for all α and β in A. See Figure 2.1. We can think of the charts $\{c_\alpha\}_{\alpha \in A}$ as providing local coordinate systems on $\mathbf{M}^p$. Formula (2.16) provides a *patching* criterion, telling us that these different coordinate systems can be glued together in a topologically consistent way.

Definition 2.2.1. *The collection of subsets* $\{U_\alpha\}_{\alpha \in A}$ *with the charts* $\{c_\alpha\}_{\alpha \in A}$ *is said to form an* atlas *on* $\mathbf{M}^p$. *The set* $\mathbf{M}^p$ *together with its atlas* $\{(U_\alpha, c_\alpha) : \alpha \in A\}$ *is called a* topological manifold *of dimension* p.

A subset $V \subset \mathbf{M}^p$ is open if $c_\alpha(V \cap U_\alpha)$ is an open subset of $\mathbf{R}^p$ for every $\alpha \in A$. This definition formalizes our basic understanding that a topological manifold is a space that is locally homeomorphic to Euclidean space.

Definition 2.2.2. *If the functions* $c_\beta \circ c_\alpha^{-1}$ *in (2.16) are also required to be* C^r*-diffeomorphisms then the topological manifold* $\mathbf{M}^p$ *is said to be a* C^r-differential manifold.

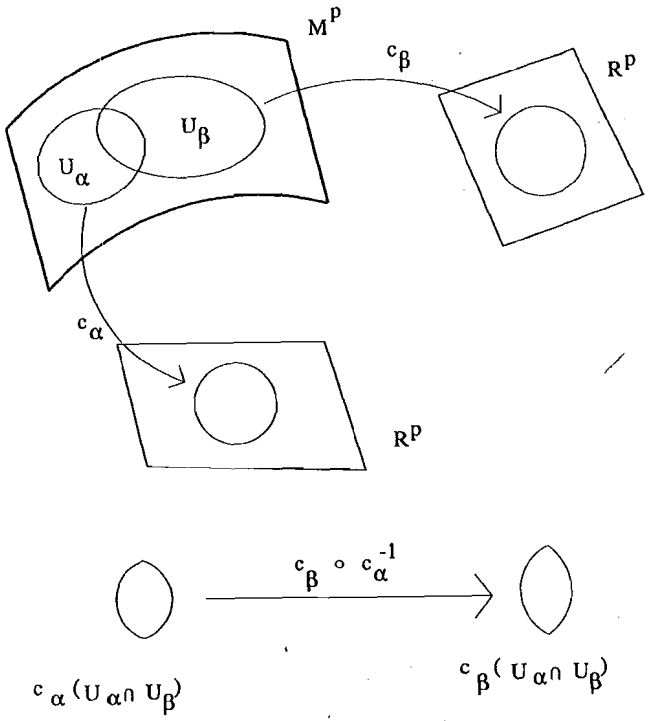

FIGURE 2.1. *Charts on a manifold. A chart provides a coordinate system on a manifold. In order to ensure that the coordinate systems are consistent with each other, a patching criterion is required on the sets of the manifold where the coordinate systems overlap. For the figure shown, the patching criterion requires that* $c_\beta \circ c_\alpha^{-1}$ *be a diffeomorphism between subsets of* $\mathbf{R}^p$. *A set of compatible charts that cover the manifold is called an* atlas. *In* $\mathbf{R}^p$ *it is often useful to change coordinate systems for the convenience of calculations. The same is true for differential manifolds. As it is the charts that provide coordinates for points in the manifold, a change of coordinate systems about a point* $x \in \mathbf{M}^p$ *is simply a change in the choice of the chart* c_α *that provides coordinates for* x. *If the change in coordinates is to be compatible with the differential structure defined on* $\mathbf{M}^p$, *then the new chart* c_β *will need to satisfy the patching criterion above. Such a criterion will automatically be satisfied if the chart* c_β *belongs to the atlas on* $\mathbf{M}^p$. *However, the new chart is not required to belong to the atlas. If the new chart satisfies the patching criterion, it can be included in the atlas, and thereby enlarge the atlas.*

For convenience, we shall refer to a C^∞-differential manifold simply as a differential manifold. Again, informally we can say that a differential manifold is a space that is locally diffeomorphic to Euclidean space.

Now let $\mathbf{M}^p$ and $\mathbf{N}^q$ be differential manifolds of dimension p and q respectively. A continuous function

$$h : \mathbf{M}^p \to \mathbf{N}^q \tag{2.17}$$

is said to be *differentiable*, or *smooth*, if for every $x \in \mathbf{M}^p$ there exists a chart (U_α, c_α) on $\mathbf{M}^p$ and a chart (V_β, c_β) on $\mathbf{N}^q$ such that $x \in U_\alpha$, $h(x) \in V_\beta$, and such that the mapping

$$c_\beta \circ h \circ c_\alpha^{-1} : c_\alpha[h^{-1}(V_\beta) \cap U_\alpha] \to \mathbf{R}^q \tag{2.18}$$

is differentiable. Similarly, we will say that h is a C^r-function provided that the function defined in formula (2.18) is a C^r-function. If $p = q$ and h is 1-1 and onto, then h is called a *C^r-diffeomorphism* provided that h and h^{-1} are C^r. Once again, when h is a C^∞-diffeomorphism, then we shall simply refer to h as a diffeomorphism. When a C^r-diffeomorphism can be established between two manifolds $\mathbf{M}^p$ and $\mathbf{N}^p$ then $\mathbf{M}^p$ and $\mathbf{N}^p$ are said to be *C^r-diffeomorphic*. If $r = \infty$ then we shall simply say that $\mathbf{M}^p$ and $\mathbf{N}^p$ are *diffeomorphic*.

Atlases provide coordinate systems for manifolds. For example, if x is a point in U_α then the coordinates of $c_\alpha(x)$ in the Euclidean space $\mathbf{R}^p$ can be used to locate the point. Unfortunately, there is usually no single chart that can provide a nondegenerate coordinate system simultaneously for the entire manifold, as charts have to be patched together to cover the manifold. However, in many cases, the points of degeneracy of coordinate systems introduced by charts need not be a hindrance to calculations. For this reason, we often suppress the chart notation, and say that point x has coordinates $(x_1, x_2, \ldots, x_p)$ rather than the more precise statement that these coordinates belong to $c_\alpha(x)$.

The *intrinsic* properties of a manifold are those that are invariant under a change of coordinates that is compatible with the differential structure, as explained in Figure 2.1. On the other hand, those properties that are dependent upon the coordinate system are called *extrinsic* properties of the manifold.

As we defined a differential manifold to be a space that is locally diffeomorphic to Euclidean space, it is not surprising that Euclidean space $\mathbf{R}^p$ turns out to be a differential manifold. To do this, we make the atlas consist of a *single* chart, with $U = \mathbf{R}^p$ and $c = e$, where e is the identity transformation from $\mathbf{R}^p$ to $\mathbf{R}^p$. With this construction, it becomes a routine matter to check that $\mathbf{R}^p$ satisfies the definition of a differential manifold.

2.2.5 An Introduction to Tangent Vectors

Let us return to our intuitive example of a differential manifold, namely a surface embedded in Euclidean space $\mathbf{R}^3$. A typical way in which a surface can be defined is as the solution set to an equation of the form

$$h(x_1,\ x_2,\ x_3)\ =\ 0 \tag{2.19}$$

where h is a real valued function defined on $\mathbf{R}^3$. Let us denote this surface by $\mathbf{M}^2$ as shown in Figure 2.2. Suppose that $x = (x_1, x_2, x_3)$ is a point on this surface. Now if the gradient vector

$$\nabla h(x)\ =\ \left(\frac{\partial h}{\partial x_1},\ \frac{\partial h}{\partial x_2},\ \frac{\partial h}{\partial x_3}\right) \tag{2.20}$$

is nonvanishing, it will point in a direction perpendicular to the surface. Tangent vectors to the surface at the point x will then be those vectors in $\mathbf{R}^3$ that are orthogonal to this normal vector. The set of all vectors that are tangent to the surface at x is said to be the *tangent space* of the surface at x. Thus a vector $v = (v_1, v_2, v_3)$ is a tangent vector to the surface at the point x if and only if

$$< v,\ \nabla h(x) > = \sum_{j=1}^{3} v_j \frac{\partial h}{\partial x_j}\ =\ 0 \tag{2.21}$$

When we turn to general differential manifolds this construction unfortunately does not generalize. Nevertheless, the space of tangent vectors can be defined in a more abstract sense, despite the fact that a normal vector to a surface is a property of the embedding in $\mathbf{R}^3$ and not intrinsic to the differential geometry of that surface. A key insight in generalizing the concept of a tangent vector is to note that on a surface, the tangent vectors at a point x can be placed in 1-1 correspondence with *equivalence classes of paths* through x, which we shall now consider.

Let x_0 be a point on the surface $\mathbf{M}^2$. Now let

$$x(t)\ =\ (x_1(t),\ x_2(t),\ x_3(t)) \tag{2.22}$$

be a path in the surface passing through a point x_0 at $t = 0$ and defined for values of t in some open interval $(-\epsilon, \epsilon)$. For each t, define the vector $\dot{x}(t)$ by

$$\dot{x}(t)\ =\ \frac{dx(t)}{dt}\ =\ \left(\frac{dx_1(t)}{dt},\ \frac{dx_2(t)}{dt},\ \frac{dx_3(t)}{dt}\right) \tag{2.23}$$

Then it can be seen that the vector $\dot{x}(0)$ is a tangent vector to the surface at the point x_0. See Figure 2.2. Thus every smooth path through x_0 defines a tangent vector at x_0. This tangent vector is not unique to the path, as there exist many paths through x_0 having the same tangent vector at that

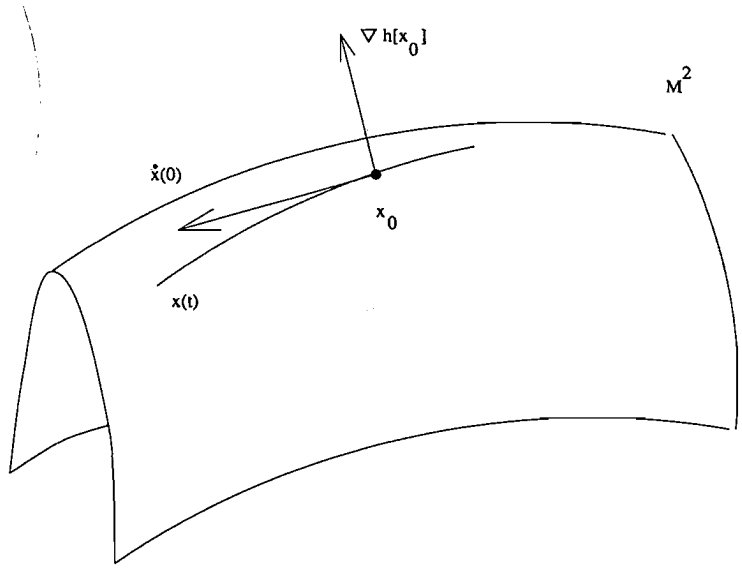

FIGURE 2.2. *Tangent and normal vectors to a surface. At any point on a surface, the tangent vectors to the surface are perpendicular to a normal vector that is the gradient of the defining equation.*

point. However, all paths through x_0 having the same tangent vector at x_0 form an *equivalence class* of paths. It is this equivalence class that we will formally identify with the tangent vector at x_0 in the next section.

We close this section by considering how tangent vectors to a surface can be used to represent infinitesimal displacements of points within the surface. Consider Figure 2.3. Along a smooth path in a surface, position two points x and y. From the point x draw a vector in $\mathbf{R}^3$ out to y. This vector is called a *secant vector* because it points along a secant line segment whose endpoints are the two points x and y in the surface. Secant vectors point in the direction of the displacement from x to y, but are represented in the Euclidean space $\mathbf{R}^3$ rather than the surface itself. When the displacement from x to y becomes infinitesimally small, then the secant vector in limiting form becomes a tangent vector to the surface. Thus we can write $dx\ =\ \dot{x}(t)\ dt$ where $dx = x(t + dt) - x(t)$ and $\dot{x}(t)$ is, once again, the tangent to the curve at $x = x(t)$. Thus the length ds of the displacement dx is

$$ds\ =\ ||\dot{x}(t)||\ dt \tag{2.24}$$

2.2.6 Tangent Vectors and Tangent Spaces

Henceforth, we shall assume that $\mathbf{M}^p$ is a differential manifold. Let $x(t)$ and $y(t)$ be two smooth paths in $\mathbf{M}^p$ passing through a common point x_0

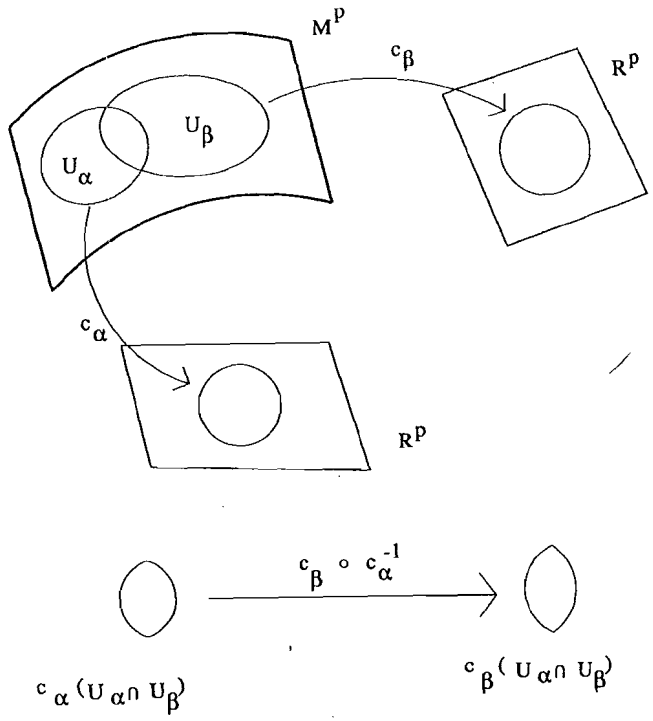

FIGURE 2.1. *Charts on a manifold. A chart provides a coordinate system on a manifold. In order to ensure that the coordinate systems are consistent with each other, a patching criterion is required on the sets of the manifold where the coordinate systems overlap. For the figure shown, the patching criterion requires that* $c_\beta \circ c_\alpha^{-1}$ *be a diffeomorphism between subsets of* $\mathbf{R}^p$. *A set of compatible charts that cover the manifold is called an* atlas. *In* $\mathbf{R}^p$ *it is often useful to change coordinate systems for the convenience of calculations. The same is true for differential manifolds. As it is the charts that provide coordinates for points in the manifold, a change of coordinate systems about a point* $x \in \mathbf{M}^p$ *is simply a change in the choice of the chart* c_α *that provides coordinates for* x. *If the change in coordinates is to be compatible with the differential structure defined on* $\mathbf{M}^p$, *then the new chart* c_β *will need to satisfy the patching criterion above. Such a criterion will automatically be satisfied if the chart* c_β *belongs to the atlas on* $\mathbf{M}^p$. *However, the new chart is not required to belong to the atlas. If the new chart satisfies the patching criterion, it can be included in the atlas, and thereby enlarge the atlas.*

For convenience, we shall refer to a C^∞-differential manifold simply as a differential manifold. Again, informally we can say that a differential manifold is a space that is locally diffeomorphic to Euclidean space.

Now let $\mathbf{M}^p$ and $\mathbf{N}^q$ be differential manifolds of dimension p and q respectively. A continuous function

$$h : \mathbf{M}^p \to \mathbf{N}^q \tag{2.17}$$

is said to be *differentiable*, or *smooth*, if for every $x \in \mathbf{M}^p$ there exists a chart (U_α, c_α) on $\mathbf{M}^p$ and a chart (V_β, c_β) on $\mathbf{N}^q$ such that $x \in U_\alpha$, $h(x) \in V_\beta$, and such that the mapping

$$c_\beta \circ h \circ c_\alpha^{-1} : c_\alpha[h^{-1}(V_\beta) \cap U_\alpha] \to \mathbf{R}^q \tag{2.18}$$

is differentiable. Similarly, we will say that h is a C^r-function provided that the function defined in formula (2.18) is a C^r-function. If $p = q$ and h is 1-1 and onto, then h is called a *C^r-diffeomorphism* provided that h and h^{-1} are C^r. Once again, when h is a C^∞-diffeomorphism, then we shall simply refer to h as a diffeomorphism. When a C^r-diffeomorphism can be established between two manifolds $\mathbf{M}^p$ and $\mathbf{N}^p$ then $\mathbf{M}^p$ and $\mathbf{N}^p$ are said to be *C^r-diffeomorphic*. If $r = \infty$ then we shall simply say that $\mathbf{M}^p$ and $\mathbf{N}^p$ are *diffeomorphic*.

Atlases provide coordinate systems for manifolds. For example, if x is a point in U_α then the coordinates of $c_\alpha(x)$ in the Euclidean space $\mathbf{R}^p$ can be used to locate the point. Unfortunately, there is usually no single chart that can provide a nondegenerate coordinate system simultaneously for the entire manifold, as charts have to be patched together to cover the manifold. However, in many cases, the points of degeneracy of coordinate systems introduced by charts need not be a hindrance to calculations. For this reason, we often suppress the chart notation, and say that point x has coordinates $(x_1, x_2, \ldots, x_p)$ rather than the more precise statement that these coordinates belong to $c_\alpha(x)$.

The *intrinsic* properties of a manifold are those that are invariant under a change of coordinates that is compatible with the differential structure, as explained in Figure 2.1. On the other hand, those properties that are dependent upon the coordinate system are called *extrinsic* properties of the manifold.

As we defined a differential manifold to be a space that is locally diffeomorphic to Euclidean space, it is not surprising that Euclidean space $\mathbf{R}^p$ turns out to be a differential manifold. To do this, we make the atlas consist of a *single* chart, with $U = \mathbf{R}^p$ and $c = e$, where e is the identity transformation from $\mathbf{R}^p$ to $\mathbf{R}^p$. With this construction, it becomes a routine matter to check that $\mathbf{R}^p$ satisfies the definition of a differential manifold.

2.2.5 An Introduction to Tangent Vectors

Let us return to our intuitive example of a differential manifold, namely a surface embedded in Euclidean space $\mathbf{R}^3$. A typical way in which a surface can be defined is as the solution set to an equation of the form

$$h(x_1,\ x_2,\ x_3)\ =\ 0 \tag{2.19}$$

where h is a real valued function defined on $\mathbf{R}^3$. Let us denote this surface by $\mathbf{M}^2$ as shown in Figure 2.2. Suppose that $x = (x_1, x_2, x_3)$ is a point on this surface. Now if the gradient vector

$$\nabla h(x)\ =\ \left(\frac{\partial h}{\partial x_1},\ \frac{\partial h}{\partial x_2},\ \frac{\partial h}{\partial x_3}\right) \tag{2.20}$$

is nonvanishing, it will point in a direction perpendicular to the surface. Tangent vectors to the surface at the point x will then be those vectors in $\mathbf{R}^3$ that are orthogonal to this normal vector. The set of all vectors that are tangent to the surface at x is said to be the *tangent space* of the surface at x. Thus a vector $v = (v_1, v_2, v_3)$ is a tangent vector to the surface at the point x if and only if

$$< v,\ \nabla h(x) > = \sum_{j=1}^{3} v_j \frac{\partial h}{\partial x_j}\ =\ 0 \tag{2.21}$$

When we turn to general differential manifolds this construction unfortunately does not generalize. Nevertheless, the space of tangent vectors can be defined in a more abstract sense, despite the fact that a normal vector to a surface is a property of the embedding in $\mathbf{R}^3$ and not intrinsic to the differential geometry of that surface. A key insight in generalizing the concept of a tangent vector is to note that on a surface, the tangent vectors at a point x can be placed in 1-1 correspondence with *equivalence classes of paths* through x, which we shall now consider.

Let x_0 be a point on the surface $\mathbf{M}^2$. Now let

$$x(t)\ =\ (x_1(t),\ x_2(t),\ x_3(t)) \tag{2.22}$$

be a path in the surface passing through a point x_0 at $t = 0$ and defined for values of t in some open interval $(-\epsilon, \epsilon)$. For each t, define the vector $\dot{x}(t)$ by

$$\dot{x}(t)\ =\ \frac{dx(t)}{dt}\ =\ \left(\frac{dx_1(t)}{dt},\ \frac{dx_2(t)}{dt},\ \frac{dx_3(t)}{dt}\right) \tag{2.23}$$

Then it can be seen that the vector $\dot{x}(0)$ is a tangent vector to the surface at the point x_0. See Figure 2.2. Thus every smooth path through x_0 defines a tangent vector at x_0. This tangent vector is not unique to the path, as there exist many paths through x_0 having the same tangent vector at that

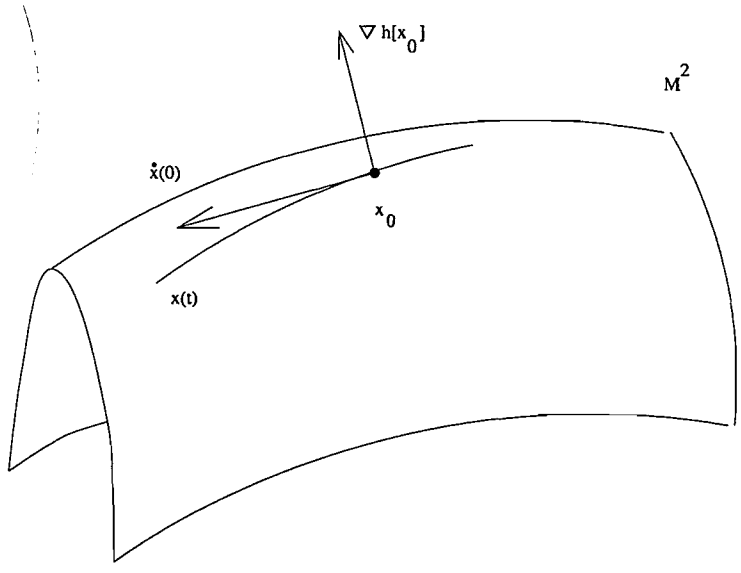

FIGURE 2.2. *Tangent and normal vectors to a surface. At any point on a surface, the tangent vectors to the surface are perpendicular to a normal vector that is the gradient of the defining equation.*

point. However, all paths through x_0 having the same tangent vector at x_0 form an *equivalence class* of paths. It is this equivalence class that we will formally identify with the tangent vector at x_0 in the next section.

We close this section by considering how tangent vectors to a surface can be used to represent infinitesimal displacements of points within the surface. Consider Figure 2.3. Along a smooth path in a surface, position two points x and y. From the point x draw a vector in $\mathbf{R}^3$ out to y. This vector is called a *secant vector* because it points along a secant line segment whose endpoints are the two points x and y in the surface. Secant vectors point in the direction of the displacement from x to y, but are represented in the Euclidean space $\mathbf{R}^3$ rather than the surface itself. When the displacement from x to y becomes infinitesimally small, then the secant vector in limiting form becomes a tangent vector to the surface. Thus we can write $dx\ =\ \dot{x}(t)\ dt$ where $dx = x(t + dt) - x(t)$ and $\dot{x}(t)$ is, once again, the tangent to the curve at $x = x(t)$. Thus the length ds of the displacement dx is

$$ds\ =\ ||\dot{x}(t)||\ dt \tag{2.24}$$

2.2.6 Tangent Vectors and Tangent Spaces

Henceforth, we shall assume that $\mathbf{M}^p$ is a differential manifold. Let $x(t)$ and $y(t)$ be two smooth paths in $\mathbf{M}^p$ passing through a common point x_0

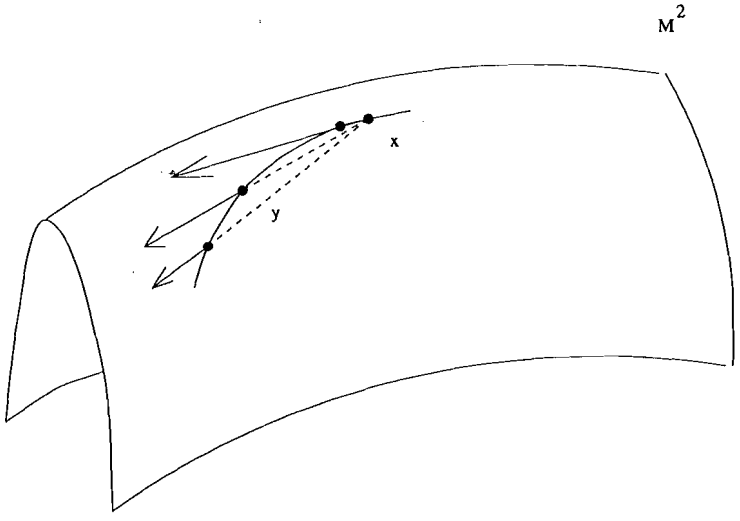

FIGURE 2.3. *Secant vectors to a surface. In the limit, as the displacement between points becomes infinitesimal, the secant vectors converge to a tangent vector.*

at $t = 0$, say. Let us suppose that a coordinate system has been constructed by a chart (U_α, c_α) around x_0 so that the paths have coordinates

$$x(t) = (x_1(t),\ x_2(t),\ ...,\ x_p(t)) \tag{2.25}$$

$$y(t) = (y_1(t),\ y_2(t),\ ...,\ y_p(t)) \tag{2.26}$$

and

$$x_0 = (x_{01},\ x_{02},\ ...,x_{0p}) \tag{2.27}$$

The paths $x(t)$ and $y(t)$ are said to be *smooth* if their coordinates are differentiable functions of the time coordinate t. Henceforth, we shall restrict attention to smooth paths. The paths $x(t)$ and $y(t)$ are said to be *tangent* at x_0 provided that

$$\frac{dx_j(0)}{dt} = \frac{dy_j(0)}{dt} \tag{2.28}$$

for all $j = 1,...,p$. It is important to note that although the condition of tangency is expressed in terms of the coordinate system, the tangency property is independent of the choice of coordinates. This follows from the fact that in $\mathbf{R}^p$, the diffeomorphic images of two tangent paths will also be tangent. Changing coordinate systems on $\mathbf{M}^p$ is equivalent to a diffeomorphism on $\mathbf{R}^p$ as formula (2.16) shows.

Definition 2.2.3. *We define the* tangent vector $\dot{x}$ *to the path* $x(t)$ *at the point* $x_0 = x(0)$ *to be the* equivalence class *of all paths* $y(t)$ *such that* $y(0) = x_0$ *and such that* $y(t)$ *is tangent to* $x(t)$ *at* $t = 0$.

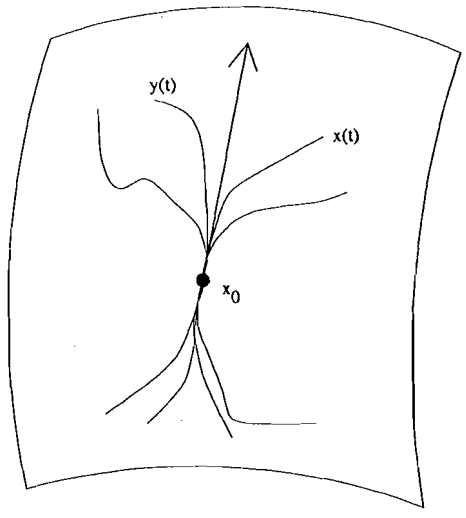

FIGURE 2.4. *A tangent vector represented as an equivalence class of paths through a point on the manifold. At any point* x_0 *in a manifold, we consider all smooth paths passing through* x_0 *at time* $t = 0$. *The property of tangency between two such paths defines an equivalence relation between the paths. The tangent vectors to the manifold at the point* x_0 *are formally defined as the equivalence classes of this relation.*

See Figure 2.4. In order to show that these equivalence classes deserve to be called tangent vectors, it is necessary to show that they have the same properties that vectors have, namely, the ability to be added together and multiplied by a scalar. Suppose that $x(t)$ and $z(t)$ are two paths passing through a point $x_0 \in \mathbf{M}^p$ at $t = 0$. We define the *vector sum* $\dot{x} + \dot{z}$ to be the tangent vector at x_0 to the path whose coordinates are

$$(x_1(t) + z_1(t) - x_{01},\ ...,\ x_p(t) + z_p(t) - x_{0p}) \tag{2.29}$$

which also passes through x_0 at $t = 0$. It is not immediately obvious that this definition of the sum of tangent vectors is well defined. To prove that it is, it is necessary to show that the definition is independent of the coordinate system used to express the paths and is independent of the choice of paths used to represent the tangent vectors $\dot{x}$ and $\dot{y}$. However, this can be done. See Problem 7 at the end of the chapter.

Similarly, we can multiply the vector $\dot{x}$ by a scalar $\lambda \in \mathbf{R}$. Define $\lambda\,\dot{x}$ to be the equivalence class of paths tangent at $t = 0$ to the path with coordinates

$$(\lambda\,[x_1(t) - x_{01}] + x_{01},\ ...,\lambda\,[x_p(t) - x_{0p}] + x_{0p}) \tag{2.30}$$

Scalar multiplication can also be shown to be well defined. Note that we

can add tangent vectors at the same point x_0 but cannot add tangent vectors that are tangent to the manifold at different points.

Definition 2.2.4. *The vector space of all tangent vectors to the manifold* $\mathbf{M}^p$ *at a given point* $x \in \mathbf{M}^p$ *is called the* tangent space *at* x *and is denoted by* $T_x(\mathbf{M}^p)$.

The tangent space $T_x(\mathbf{M}^p)$ can be shown to have the same dimension as the manifold $\mathbf{M}^p$. So $T_x(\mathbf{M}^p)$ is linearly isomorphic to Euclidean space $\mathbf{R}^p$.

Within $T_x(\mathbf{M}^p)$ it is possible to construct a set of basis vectors as follows: For each $j = 1, ..., p$ consider the path

$$t \rightarrow (x_1, x_2, ..., x_{j-1}, x_j + t, x_{j+1}, ..., x_p) \tag{2.31}$$

defined in a neighborhood of $x = (x_1, ..., x_p)$ around $t = 0$. These paths pass through the point x at $t = 0$ and follow the axes of the coordinate system about x. For each $j = 1, ..., n$ we define $\partial_j(x) \in T_x(\mathbf{M}^p)$ to be the tangent vector to the path defined by formula (2.31) at the point x where $t = 0$.

The tangent vectors $\partial_1(x)$, $\partial_2(x)$, ..., $\partial_p(x)$ collectively form a basis for the tangent space $T_x(\mathbf{M}^p)$. That is, any tangent vector in $T_x(\mathbf{M}^p)$ can be written as

$$\sum_{j=1}^{p} a_j(x)\ \partial_j(x) \tag{2.32}$$

where each a_j is a real valued function of $x \in \mathbf{M}^p$. For example, we can write

$$\dot{x}(t) = \sum_{j=1}^{p} \dot{x}_j(t)\ \partial_j[x(t)] \tag{2.33}$$

where

$$\dot{x}_j(t) = \frac{dx_j(t)}{dt} \tag{2.34}$$

It should be noted that the definition of the basis vectors ∂_1, ∂_2, ..., ∂_p depends upon the particular coordinate system used. Under a change in the coordinate system around x, a different set of basis vectors emerges. However, both sets span the same space $T_x(\mathbf{M}^p)$, whose elements are intrinsic to the manifold and not artifacts of the choice of coordinate system.

As the tangent vector of formula (2.32) is a function of x, it defines a tangent vector at every $x \in \mathbf{M}^p$ where the coordinate system is defined. A function that assigns an element of $T_x(\mathbf{M}^p)$ for every $x \in \mathbf{M}^p$ is called a *tangent vector field* on $\mathbf{M}^p$. The tangent vector field is said to be a C^r-vector field provided that when expressed in terms of the basis vectors $\partial_1(x), ..., \partial_p(x)$, the real valued functions a_j are C^r-functions of $x \in \mathbf{M}^p$.

2.2.7 *Metric Tensors and Riemannian Manifolds*

Suppose that

$$g(x) = \begin{pmatrix} g_{11}(x) & g_{12}(x) & \cdots & g_{1p}(x) \\ g_{21}(x) & g_{22}(x) & \cdots & g_{2p}(x) \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ g_{p1}(x) & g_{p2}(x) & \cdots & g_{pp}(x) \end{pmatrix} \tag{2.35}$$

is a positive definite symmetric matrix for all $x \in \mathbf{M}^p$. Then $g(x)$ defines an inner product on $T_x(\mathbf{M}^p)$ as follows. Consider two tangent vectors in $T_x(\mathbf{M}^p)$, namely $\sum_j a_j(x)\partial_j(x)$ and $\sum_k b_k(x)\partial_k(x)$. Then we define the inner product of these tangent vectors to be

$$< \sum_{j=1}^{p} a_j(x)\partial_j(x),\ \sum_{k=1}^{p} b_k(x)\partial_k(x) > = \sum_{j=1}^{p}\sum_{k=1}^{p} g_{jk}(x)a_j(x)b_k(x) \tag{2.36}$$

This notation is cumbersome if used on a regular basis. We shall suppose that g_{jk} is a smoothly varying function in x across the manifold and shall suppress the x, both in g_{jk} and the tangent vectors. Thus we can also write this in more compact form as

$$< \sum_j a_j\partial_j\ ,\ \sum_k b_k\partial_k > = \sum_j\sum_k g_{jk}a_jb_k \tag{2.37}$$

In the classical notation of differential geometry, the notation is even more compact, with equation (2.37) written with the summation signs understood, following the *Einstein summation convention.* This classical notation is not well suited to our purposes here. Therefore we shall continue to use a less compact notation that includes summation signs.

Definition 2.2.5. *The inner product defined on the tangent spaces of the manifold by (2.37) is said to be a* Riemannian metric tensor, *or simply a* metric tensor *on* $\mathbf{M}^p$. *A differential manifold endowed with a smooth metric tensor is said to be a* Riemannian manifold.

Metric tensors allow us to define inner products between tangent vectors at the same point $x \in \mathbf{M}^p$ but do not define inner products between tangent vectors at different points.

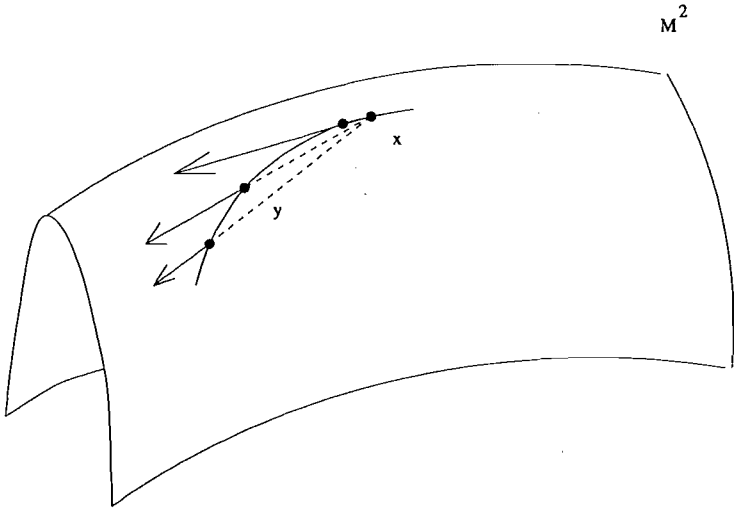

FIGURE 2.3. *Secant vectors to a surface. In the limit, as the displacement between points becomes infinitesimal, the secant vectors converge to a tangent vector.*

at $t = 0$, say. Let us suppose that a coordinate system has been constructed by a chart (U_α, c_α) around x_0 so that the paths have coordinates

$$x(t) = (x_1(t),\ x_2(t),\ ...,\ x_p(t)) \tag{2.25}$$

$$y(t) = (y_1(t),\ y_2(t),\ ...,\ y_p(t)) \tag{2.26}$$

and

$$x_0 = (x_{01},\ x_{02},\ ...,x_{0p}) \tag{2.27}$$

The paths $x(t)$ and $'y(t)$ are said to be *smooth* if their coordinates are differentiable functions of the time coordinate t. Henceforth, we shall restrict attention to smooth paths. The paths $x(t)$ and $y(t)$ are said to be *tangent* at x_0 provided that

$$\frac{dx_j(0)}{dt} = \frac{dy_j(0)}{dt} \tag{2.28}$$

for all $j = 1,...,p$. It is important to note that although the condition of tangency is expressed in terms of the coordinate system, the tangency property is independent of the choice of coordinates. This follows from the fact that in $\mathbf{R}^p$, the diffeomorphic images of two tangent paths will also be tangent. Changing coordinate systems on $\mathbf{M}^p$ is equivalent to a diffeomorphism on $\mathbf{R}^p$ as formula (2.16) shows.

Definition 2.2.3. *We define the* tangent vector $\dot{x}$ *to the path* $x(t)$ *at the point* $x_0 = x(0)$ *to be the* equivalence class *of all paths* $y(t)$ *such that* $y(0) = x_0$ *and such that* $y(t)$ *is tangent to* $x(t)$ *at* $t = 0$.

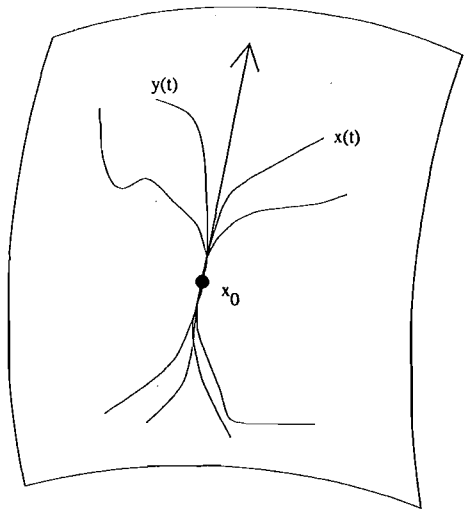

FIGURE 2.4. *A tangent vector represented as an equivalence class of paths through a point on the manifold. At any point* x_0 *in a manifold, we consider all smooth paths passing through* x_0 *at time* $t = 0$. *The property of tangency between two such paths defines an equivalence relation between the paths. The tangent vectors to the manifold at the point* x_0 *are formally defined as the equivalence classes of this relation.*

See Figure 2.4. In order to show that these equivalence classes deserve to be called tangent vectors, it is necessary to show that they have the same properties that vectors have, namely, the ability to be added together and multiplied by a scalar. Suppose that $x(t)$ and $z(t)$ are two paths passing through a point $x_0 \in \mathbf{M}^p$ at $t = 0$. We define the *vector sum* $\dot{x} + \dot{z}$ to be the tangent vector at x_0 to the path whose coordinates are

$$(x_1(t) + z_1(t) - x_{01},\ ...,\ x_p(t) + z_p(t) - x_{0p}) \tag{2.29}$$

which also passes through x_0 at $t = 0$. It is not immediately obvious that this definition of the sum of tangent vectors is well defined. To prove that it is, it is necessary to show that the definition is independent of the coordinate system used to express the paths and is independent of the choice of paths used to represent the tangent vectors $\dot{x}$ and $\dot{y}$. However, this can be done. See Problem 7 at the end of the chapter.

Similarly, we can multiply the vector $\dot{x}$ by a scalar $\lambda \in \mathbf{R}$. Define $\lambda\,\dot{x}$ to be the equivalence class of paths tangent at $t = 0$ to the path with coordinates

$$(\lambda\,[x_1(t) - x_{01}]\ +\ x_{01},\ ...,\lambda\,[x_p(t) - x_{0p}]\ +\ x_{0p}) \tag{2.30}$$

Scalar multiplication can also be shown to be well defined. Note that we

can add tangent vectors at the same point x_0 but cannot add tangent vectors that are tangent to the manifold at different points.

Definition 2.2.4. *The vector space of all tangent vectors to the manifold* $\mathbf{M}^p$ *at a given point* $x \in \mathbf{M}^p$ *is called the* tangent space *at* x *and is denoted by* $T_x(\mathbf{M}^p)$.

The tangent space $T_x(\mathbf{M}^p)$ can be shown to have the same dimension as the manifold $\mathbf{M}^p$. So $T_x(\mathbf{M}^p)$ is linearly isomorphic to Euclidean space $\mathbf{R}^p$.

Within $T_x(\mathbf{M}^p)$ it is possible to construct a set of basis vectors as follows: For each $j = 1, ..., p$ consider the path

$$t \rightarrow (x_1, x_2, ..., x_{j-1}, x_j + t, x_{j+1}, ..., x_p) \tag{2.31}$$

defined in a neighborhood of $x = (x_1, ..., x_p)$ around $t = 0$. These paths pass through the point x at $t = 0$ and follow the axes of the coordinate system about x. For each $j = 1, ..., n$ we define $\partial_j(x) \in T_x(\mathbf{M}^p)$ to be the tangent vector to the path defined by formula (2.31) at the point x where $t = 0$.

The tangent vectors $\partial_1(x)$, $\partial_2(x)$, ..., $\partial_p(x)$ collectively form a basis for the tangent space $T_x(\mathbf{M}^p)$. That is, any tangent vector in $T_x(\mathbf{M}^p)$ can be written as

$$\sum_{j=1}^{p} a_j(x)\ \partial_j(x) \tag{2.32}$$

where each a_j is a real valued function of $x \in \mathbf{M}^p$. For example, we can write

$$\dot{x}(t) = \sum_{j=1}^{p} \dot{x}_j(t)\ \partial_j[x(t)] \tag{2.33}$$

where

$$\dot{x}_j(t) = \frac{dx_j(t)}{dt} \tag{2.34}$$

It should be noted that the definition of the basis vectors ∂_1, ∂_2, ..., ∂_p depends upon the particular coordinate system used. Under a change in the coordinate system around x, a different set of basis vectors emerges. However, both sets span the same space $T_x(\mathbf{M}^p)$, whose elements are intrinsic to the manifold and not artifacts of the choice of coordinate system.

As the tangent vector of formula (2.32) is a function of x, it defines a tangent vector at every $x \in \mathbf{M}^p$ where the coordinate system is defined. A function that assigns an element of $T_x(\mathbf{M}^p)$ for every $x \in \mathbf{M}^p$ is called a *tangent vector field* on $\mathbf{M}^p$. The tangent vector field is said to be a C^r-vector field provided that when expressed in terms of the basis vectors $\partial_1(x), ..., \partial_p(x)$, the real valued functions a_j are C^r-functions of $x \in \mathbf{M}^p$.

2.2.7 *Metric Tensors and Riemannian Manifolds*

Suppose that

$$g(x) = \begin{pmatrix} g_{11}(x) & g_{12}(x) & \cdots & g_{1p}(x) \\ g_{21}(x) & g_{22}(x) & \cdots & g_{2p}(x) \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ g_{p1}(x) & g_{p2}(x) & \cdots & g_{pp}(x) \end{pmatrix} \tag{2.35}$$

is a positive definite symmetric matrix for all $x \in \mathbf{M}^p$. Then $g(x)$ defines an inner product on $T_x(\mathbf{M}^p)$ as follows. Consider two tangent vectors in $T_x(\mathbf{M}^p)$, namely $\sum_j a_j(x)\partial_j(x)$ and $\sum_k b_k(x)\partial_k(x)$. Then we define the inner product of these tangent vectors to be

$$< \sum_{j=1}^{p} a_j(x)\partial_j(x),\ \sum_{k=1}^{p} b_k(x)\partial_k(x) > = \sum_{j=1}^{p}\sum_{k=1}^{p} g_{jk}(x)a_j(x)b_k(x) \tag{2.36}$$

This notation is cumbersome if used on a regular basis. We shall suppose that g_{jk} is a smoothly varying function in x across the manifold and shall suppress the x, both in g_{jk} and the tangent vectors. Thus we can also write this in more compact form as

$$< \sum_j a_j\partial_j\ ,\ \sum_k b_k\partial_k > = \sum_j\sum_k g_{jk}a_jb_k \tag{2.37}$$

In the classical notation of differential geometry, the notation is even more compact, with equation (2.37) written with the summation signs understood, following the *Einstein summation convention*. This classical notation is not well suited to our purposes here. Therefore we shall continue to use a less compact notation that includes summation signs.

Definition 2.2.5. *The inner product defined on the tangent spaces of the manifold by (2.37) is said to be a* Riemannian metric tensor, *or simply a* metric tensor *on* $\mathbf{M}^p$. *A differential manifold endowed with a smooth metric tensor is said to be a* Riemannian manifold.

Metric tensors allow us to define inner products between tangent vectors at the same point $x \in \mathbf{M}^p$ but do not define inner products between tangent vectors at different points.

2.2.8 *Geodesic Paths and Geodesic Distance*

Consider a smooth path $x(t)$ on a Riemannian manifold $\mathbf{M}^p$. The tangent vector to the path at a time t is

$$\dot{x}(t) = \sum_{j=1}^{p} \frac{dx_j(t)}{dt} \partial_j(t) \tag{2.38}$$

where $x_j(t)$ is the jth coordinate of $x(t)$ and $\partial_j(t) = \partial_j[x(t)]$ is the jth basis vector of the tangent space $T_{x(t)}(\mathbf{M}^p)$. In more compact notation, this becomes

$$\dot{x}(t) = \sum_{j=1}^{p} \dot{x}_j(t) \partial_j(t) \tag{2.39}$$

For any t let

$$\gamma(t) = ||\dot{x}(t)|| = \sqrt{< \dot{x}(t), \dot{x}(t) >} \tag{2.40}$$

be the length of the vector $\dot{x}(t)$. The inner product generated by the metric tensor can be calculated using formula (2.37). So we can write

$$\gamma(t) = \sqrt{\sum_{j=1}^{p} \sum_{k=1}^{p} g_{jk}(t) \dot{x}_j(t) \dot{x}_k(t)} \tag{2.41}$$

where $g_{jk}(t)$ is the value of the metric tensor at $x(t)$.

Suppose t undergoes a small increment to $t + dt$. Then, as in formula (2.24), the length ds of the path segment from $x(t)$ to $x(t + dt)$ is

$$ds = \gamma(t)\, dt \tag{2.42}$$

Therefore we can write the length of the path $x(\ .\ )$ from $t = t_0$ to $t = t_1$ as

$$L = \int_{t_0}^{t_1} ds = \int_{t_0}^{t_1} \gamma(t)\, dt\ . \tag{2.43}$$

It should be noted that not only does the metric tensor determine the lengths of arcs, but the metric tensor is also itself determined by the arc length. That is, if ds can be calculated for any increment of a path from $x(t)$ to $x(t + dt)$ then there is at most one metric tensor g that is compatible with this definition. In some cases, we shall determine the structure of a Riemannian manifold by calculating the arc length function ds.

Roughly speaking, a *geodesic* path on a Riemannian manifold is the path between two points that has shortest length. This definition is a bit too narrow to work but serves for the basic intuition. More correctly, we can say that a geodesic $x(t)$ is a path in a Riemannian manifold that is locally shortest. This means that the path can be broken up into pieces such that

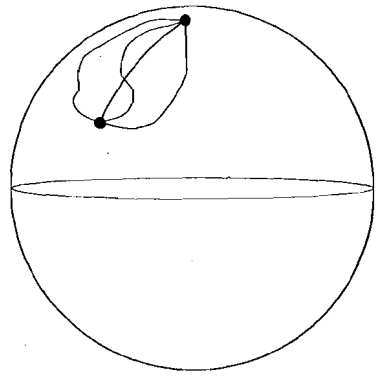

FIGURE 2.5. *The geodesic path on a manifold displayed as the path of locally shortest length. On the sphere we see a variety of paths between two points. The shortest path is a geodesic between the two points, which in this case is an arc of a great circle of the sphere.*

the paths connecting the endpoints of the pieces are all the shortest paths. This definition does not require that the endpoints of the path $x(t)$ be specified in order to determine whether it is a geodesic: the property can be investigated locally along the path. See Figure 2.5. In Euclidean space $\mathbf{R}^p$, the shortest distance between two points is, of course, a straight line. Thus the geodesic paths of a manifold can be regarded as the analogs of straight lines for spaces that are not flat. We can find formulas for geodesic paths by applying the calculus of variations to the arc length formula (2.43) above.

To find a condition to ensure that this path is minimal in length, we consider a perturbation of the path along a coordinate. If the integral is minimized then its derivative with respect to this perturbation is zero. This leads to the following set of equations from the calculus of variations. The path is a geodesic provided the Euler-Lagrange equations are satisfied, namely that

$$\frac{d}{dt}\left(\frac{\partial \gamma}{\partial \dot{x}_j}\right) = \frac{\partial \gamma}{\partial x_j} \tag{2.44}$$

for all $j = 1, 2, \ldots, p$. To interpret the partial derivatives in this formula, note that for fixed t the expression $\gamma(t)$ depends upon $x_1(t), \ldots, x_p(t)$ and $\dot{x}_1(t), \ldots, \dot{x}_p(t)$. Variation in the position coordinates $x_j(t)$ is suppressed in the notation, but arises from the metric tensor g in formula (2.41), which is a function of $(x_1, \ldots, x_p)$. The partial derivatives are understood to be the partial derivatives in each of these $2p$ variables holding the other $2p - 1$ variables fixed. The example in Section 2.2.10 below shows how to interpret this formula in $\mathbf{R}^p$ with the usual coordinate system. Problems 5 and 6 at the end of the chapter ask the reader to check the equations for various settings.

Having defined the concept of a geodesic path in a Riemannian manifold, we are in a position to define the concept of the *geodesic distance* between two points in the manifold.

Definition 2.2.6. *Suppose that a Riemannian manifold* $\mathbf{M}^p$ *is pathwise connected, in the sense that for any two points* $x, y \in \mathbf{M}^p$ *there exists a smooth path* $x(t)$ *such that* $x(t_0) = x$ *and* $x(t_1) = y$. *We define the* geodesic distance *from* x *to* y *to be the length of the shortest path from* x *to* y.

With this definition, a pathwise connected Riemannian manifold becomes a metric space, as was defined in Problem 5 of Chapter 1.

It can be shown that the path of shortest length is a geodesic in $\mathbf{M}^p$. However, the converse does not hold. The length of a geodesic path from x to y can be strictly greater than the geodesic distance from x to y. This can easily be seen by considering the fact that on a sphere any great circle passing through two distinct points can be subdivided into two paths from one point to the other. Both of these paths are geodesics, but their lengths need not be equal. It is the smaller of these two lengths that is the geodesic distance from one point to the other.

2.2.9 *Affine Connections*

Closely related to the concept of a geodesic path is the concept of an affine connection. We noted earlier that the metric tensor allows us to compare the lengths and orientations of tangent vectors within a tangent space $T_x(\mathbf{M}^p)$. However, the metric tensor does not give us a direct method of comparing vectors in different tangent spaces, say $T_x(\mathbf{M}^p)$ and $T_y(\mathbf{M}^p)$. The way we would naturally think of doing this is to rigidly transport a vector from one place in the manifold to another. For example, we could draw a geodesic from x to y and move a vector along the geodesic so that its length remains constant and its angle with respect to the tangent vector of the geodesic path is also constant. A method for transporting tangent vectors is called an *affine connection.* The particular method just described using geodesics and the metric tensor is called the *Levi-Civita connection.* A curious property of connections such as the Levi-Civita connection is that when vectors are transported around the manifold along a sequence of geodesic paths, they can arrive back at their starting place with a different orientation from the one they started with. This is paradoxical when we recall that the method of transport associated with the Levi-Civita connection requires that the orientation remain fixed with respect to the paths. However, the reader can try it on a sphere and observe this, moving a vector from the north pole to the equator, part way around the equator, and back to the north pole again. This change in orientation is a consequence

of the *curvature* of the manifold.

2.2.10 *Example*

We consider the geodesics on $\mathbf{R}^p$ and check that they are straight lines. The usual Cartesian coordinates are used so that the atlas consists of a single chart $(\mathbf{R}^p, e)$, where $e : \mathbf{R}^p \to \mathbf{R}^p$ is the identity map. The metric tensor g is the $p \times p$ identity matrix. Consider a smooth path $x(t)$ in $\mathbf{R}^p$. Then

$$\gamma = \sqrt{\sum_{j=1}^{p} \dot{x}_j^2(t)} \tag{2.45}$$

The partial derivative $\partial\gamma/\partial\dot{x}_j$ on the left-hand side of (2.44) can be computed directly from this formula by holding all other $\dot{x}_k$ constant for $k \neq j$. We obtain

$$\frac{\partial\gamma}{\partial\dot{x}_j} = \frac{\dot{x}_j}{\gamma} = \frac{\dot{x}_j}{||\dot{x}||} \tag{2.46}$$

This is a directional cosine of $\dot{x}$. Thus the left-hand side measures how this directional cosine of the tangent vector along the path changes. On the right-hand side of the Euler-Lagrange equations the partial derivatives are all zero because the metric tensor g_{jk} in the formula for $\gamma(t)$ is a constant function of position $x(t)$. Therefore, we see that the Euler-Lagrange equations reduce to stating that the directional cosines of the path are constant. The path must therefore be a straight line.

2.2.11 *Building New Manifolds From Old: Product Manifolds*

Just as it is possible to build Euclidean spaces of arbitrarily high dimension by taking Cartesian products of $\mathbf{R}$, so it is possible to build differential manifolds by taking Cartesian products of differential manifolds. Suppose $\mathbf{M}^p$ and $\mathbf{N}^q$ are differential manifolds of dimension p and q respectively. We can make $\mathbf{M}^p \times \mathbf{N}^q$ into a differential manifold by using charts of the form

$$(U_\alpha \times V_\beta \ , \ c_\alpha \times c_\beta) \tag{2.47}$$

where (U_α, c_α) is a chart on $\mathbf{M}^p$, (V_β, c_β) is a chart on $\mathbf{N}^q$, and

$$(c_\alpha \times c_\beta) : U_\alpha \times V_\beta \ \to \ \mathbf{R}^{p+q} \tag{2.48}$$

is defined by

$$(c_\alpha \times c_\beta)(x, y) = (c_\alpha(x), c_\beta(y)) \tag{2.49}$$

The manifold $\mathbf{M}^p \times \mathbf{N}^q$ resulting from this definition is of dimension $p+q$. Tangent spaces of $\mathbf{M}^p \times \mathbf{N}^q$ can be identified with Cartesian products of

2.2.8 *Geodesic Paths and Geodesic Distance*

Consider a smooth path $x(t)$ on a Riemannian manifold $\mathbf{M}^p$. The tangent vector to the path at a time t is

$$\dot{x}(t) = \sum_{j=1}^{p} \frac{dx_j(t)}{dt} \partial_j(t) \tag{2.38}$$

where $x_j(t)$ is the jth coordinate of $x(t)$ and $\partial_j(t) = \partial_j[x(t)]$ is the jth basis vector of the tangent space $T_{x(t)}(\mathbf{M}^p)$. In more compact notation, this becomes

$$\dot{x}(t) = \sum_{j=1}^{p} \dot{x}_j(t) \partial_j(t) \tag{2.39}$$

For any t let

$$\gamma(t) = \|\dot{x}(t)\| = \sqrt{< \dot{x}(t), \dot{x}(t) >} \tag{2.40}$$

be the length of the vector $\dot{x}(t)$. The inner product generated by the metric tensor can be calculated using formula (2.37). So we can write

$$\gamma(t) = \sqrt{\sum_{j=1}^{p} \sum_{k=1}^{p} g_{jk}(t) \dot{x}_j(t) \dot{x}_k(t)} \tag{2.41}$$

where $g_{jk}(t)$ is the value of the metric tensor at $x(t)$.

Suppose t undergoes a small increment to $t + dt$. Then, as in formula (2.24), the length ds of the path segment from $x(t)$ to $x(t + dt)$ is

$$ds = \gamma(t)\, dt \tag{2.42}$$

Therefore we can write the length of the path $x(\ .\ )$ from $t = t_0$ to $t = t_1$ as

$$L = \int_{t_0}^{t_1} ds = \int_{t_0}^{t_1} \gamma(t)\, dt\ . \tag{2.43}$$

It should be noted that not only does the metric tensor determine the lengths of arcs, but the metric tensor is also itself determined by the arc length. That is, if ds can be calculated for any increment of a path from $x(t)$ to $x(t + dt)$ then there is at most one metric tensor g that is compatible with this definition. In some cases, we shall determine the structure of a Riemannian manifold by calculating the arc length function ds.

Roughly speaking, a *geodesic* path on a Riemannian manifold is the path between two points that has shortest length. This definition is a bit too narrow to work but serves for the basic intuition. More correctly, we can say that a geodesic $x(t)$ is a path in a Riemannian manifold that is locally shortest. This means that the path can be broken up into pieces such that

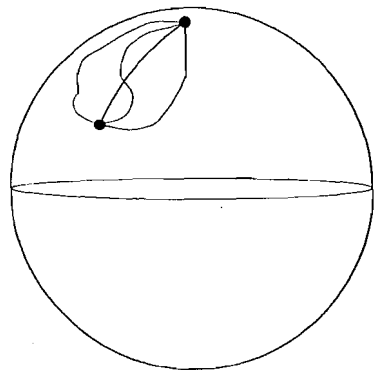

FIGURE 2.5. *The geodesic path on a manifold displayed as the path of locally shortest length. On the sphere we see a variety of paths between two points. The shortest path is a geodesic between the two points, which in this case is an arc of a great circle of the sphere.*

the paths connecting the endpoints of the pieces are all the shortest paths. This definition does not require that the endpoints of the path $x(t)$ be specified in order to determine whether it is a geodesic: the property can be investigated locally along the path. See Figure 2.5. In Euclidean space $\mathbf{R}^p$, the shortest distance between two points is, of course, a straight line. Thus the geodesic paths of a manifold can be regarded as the analogs of straight lines for spaces that are not flat. We can find formulas for geodesic paths by applying the calculus of variations to the arc length formula (2.43) above.

To find a condition to ensure that this path is minimal in length, we consider a perturbation of the path along a coordinate. If the integral is minimized then its derivative with respect to this perturbation is zero. This leads to the following set of equations from the calculus of variations. The path is a geodesic provided the Euler-Lagrange equations are satisfied, namely that

$$\frac{d}{dt}\left(\frac{\partial \gamma}{\partial \dot{x}_j}\right) = \frac{\partial \gamma}{\partial x_j} \tag{2.44}$$

for all $j = 1, 2, \ldots, p$. To interpret the partial derivatives in this formula, note that for fixed t the expression $\gamma(t)$ depends upon $x_1(t), \ldots, x_p(t)$ and $\dot{x}_1(t), \ldots, \dot{x}_p(t)$. Variation in the position coordinates $x_j(t)$ is suppressed in the notation, but arises from the metric tensor g in formula (2.41), which is a function of $(x_1, \ldots, x_p)$. The partial derivatives are understood to be the partial derivatives in each of these $2p$ variables holding the other $2p - 1$ variables fixed. The example in Section 2.2.10 below shows how to interpret this formula in $\mathbf{R}^p$ with the usual coordinate system. Problems 5 and 6 at the end of the chapter ask the reader to check the equations for various settings.

Having defined the concept of a geodesic path in a Riemannian manifold, we are in a position to define the concept of the *geodesic distance* between two points in the manifold.

Definition 2.2.6. *Suppose that a Riemannian manifold* $\mathbf{M}^p$ *is pathwise connected, in the sense that for any two points* $x, y \in \mathbf{M}^p$ *there exists a smooth path* $x(t)$ *such that* $x(t_0) = x$ *and* $x(t_1) = y$. *We define the* geodesic distance *from* x *to* y *to be the length of the shortest path from* x *to* y.

With this definition, a pathwise connected Riemannian manifold becomes a metric space, as was defined in Problem 5 of Chapter 1.

It can be shown that the path of shortest length is a geodesic in $\mathbf{M}^p$. However, the converse does not hold. The length of a geodesic path from x to y can be strictly greater than the geodesic distance from x to y. This can easily be seen by considering the fact that on a sphere any great circle passing through two distinct points can be subdivided into two paths from one point to the other. Both of these paths are geodesics, but their lengths need not be equal. It is the smaller of these two lengths that is the geodesic distance from one point to the other.

2.2.9 *Affine Connections*

Closely related to the concept of a geodesic path is the concept of an affine connection. We noted earlier that the metric tensor allows us to compare the lengths and orientations of tangent vectors within a tangent space $T_x(\mathbf{M}^p)$. However, the metric tensor does not give us a direct method of comparing vectors in different tangent spaces, say $T_x(\mathbf{M}^p)$ and $T_y(\mathbf{M}^p)$. The way we would naturally think of doing this is to rigidly transport a vector from one place in the manifold to another. For example, we could draw a geodesic from x to y and move a vector along the geodesic so that its length remains constant and its angle with respect to the tangent vector of the geodesic path is also constant. A method for transporting tangent vectors is called an *affine connection.* The particular method just described using geodesics and the metric tensor is called the *Levi-Civita connection.* A curious property of connections such as the Levi-Civita connection is that when vectors are transported around the manifold along a sequence of geodesic paths, they can arrive back at their starting place with a different orientation from the one they started with. This is paradoxical when we recall that the method of transport associated with the Levi-Civita connection requires that the orientation remain fixed with respect to the paths. However, the reader can try it on a sphere and observe this, moving a vector from the north pole to the equator, part way around the equator, and back to the north pole again. This change in orientation is a consequence

of the *curvature* of the manifold.

2.2.10 *Example*

We consider the geodesics on $\mathbf{R}^p$ and check that they are straight lines. The usual Cartesian coordinates are used so that the atlas consists of a single chart $(\mathbf{R}^p, e)$, where $e : \mathbf{R}^p \to \mathbf{R}^p$ is the identity map. The metric tensor g is the $p \times p$ identity matrix. Consider a smooth path $x(t)$ in $\mathbf{R}^p$. Then

$$\gamma = \sqrt{\sum_{j=1}^{p} \dot{x}_j^2(t)} \tag{2.45}$$

The partial derivative $\partial\gamma/\partial\dot{x}_j$ on the left-hand side of (2.44) can be computed directly from this formula by holding all other $\dot{x}_k$ constant for $k \neq j$. We obtain

$$\frac{\partial\gamma}{\partial\dot{x}_j} = \frac{\dot{x}_j}{\gamma} = \frac{\dot{x}_j}{||\dot{x}||} \tag{2.46}$$

This is a directional cosine of $\dot{x}$. Thus the left-hand side measures how this directional cosine of the tangent vector along the path changes. On the right-hand side of the Euler-Lagrange equations the partial derivatives are all zero because the metric tensor g_{jk} in the formula for $\gamma(t)$ is a constant function of position $x(t)$. Therefore, we see that the Euler-Lagrange equations reduce to stating that the directional cosines of the path are constant. The path must therefore be a straight line.

2.2.11 *Building New Manifolds From Old: Product Manifolds*

Just as it is possible to build Euclidean spaces of arbitrarily high dimension by taking Cartesian products of $\mathbf{R}$, so it is possible to build differential manifolds by taking Cartesian products of differential manifolds. Suppose $\mathbf{M}^p$ and $\mathbf{N}^q$ are differential manifolds of dimension p and q respectively. We can make $\mathbf{M}^p \times \mathbf{N}^q$ into a differential manifold by using charts of the form

$$(U_\alpha \times V_\beta \ , \ c_\alpha \times c_\beta) \tag{2.47}$$

where (U_α, c_α) is a chart on $\mathbf{M}^p$, (V_β, c_β) is a chart on $\mathbf{N}^q$, and

$$(c_\alpha \times c_\beta) : U_\alpha \times V_\beta \ \to \ \mathbf{R}^{p+q} \tag{2.48}$$

is defined by

$$(c_\alpha \times c_\beta)(x, y) = (c_\alpha(x), c_\beta(y)) \tag{2.49}$$

The manifold $\mathbf{M}^p \times \mathbf{N}^q$ resulting from this definition is of dimension $p+q$. Tangent spaces of $\mathbf{M}^p \times \mathbf{N}^q$ can be identified with Cartesian products of

those of $\mathbf{M}^p$ and $\mathbf{N}^q$ so that

$$T_{(x,y)}(\mathbf{M}^p \times \mathbf{N}^q) = T_x(\mathbf{M}^p) \times T_y(\mathbf{N}^q) \tag{2.50}$$

With this understanding, we can make $\mathbf{M}^p \times \mathbf{N}^q$ into a Riemannian manifold by putting the metric tensor elements as blocks down the main diagonal. If g_M is a metric tensor on $\mathbf{M}^p$ and g_N is a metric tensor on $\mathbf{N}^q$ then an appropriate metric on $\mathbf{M}^p \times \mathbf{N}^q$ is

$$\left(\begin{array}{c|c} g_M & 0 \\ \hline 0 & g_N \end{array}\right) \tag{2.51}$$

2.2.12 *Building New Manifolds From Old: Submanifolds*

It is also possible to construct new manifolds by looking inside a manifold. Suppose $\mathbf{N}^q$ is a subset of a differential manifold $\mathbf{M}^p$. We say that $\mathbf{N}^q$ is a *q-dimensional submanifold* of $\mathbf{M}^p$ for $q < p$ if for every point $y \in \mathbf{N}^q$, there exists a smooth coordinate system $x = (x_1, ..., x_p)$ on some open set $U \subset \mathbf{M}^p$ containing y such that

$$U \cap \mathbf{N}^q = \{x \in U : x_{q+1} = x_{q+2} = ... = x_p = 0\} \tag{2.52}$$

More informally we can say that a q-dimensional submanifold of $\mathbf{M}^p$ is a subset that is locally diffeomorphic to a linear subspace. The submanifold $\mathbf{N}^q$ inherits a coordinate system from this construction. The coordinates

$$x \rightarrow (x_1, x_2, ..., x_q) \tag{2.53}$$

make a local smooth coordinate system of the right dimension on the submanifold.

Using the coordinate system $x = (x_1, x_2, ..., x_p)$ we can set up the basis $\partial_1, \partial_2, ..., \partial_p$ for the tangent space $T_x(\mathbf{M}^p)$. Among these basis vectors, the first q tangent vectors $\partial_1, \partial_2, ..., \partial_q$ form a basis for the tangent space $T_x(\mathbf{N}^q)$. Thus any tangent vector in $T_x(\mathbf{N}^q)$ can be written as $\sum_{j=1}^{q} a_j(x)\partial_j(x)$. If $\mathbf{M}^p$ is a Riemannian manifold, then $\mathbf{N}^q$ can be made into a Riemannian manifold by inheriting the concept of arc length from $\mathbf{M}^p$. A geodesic path in $\mathbf{N}^q$ is simply a path of shortest length in $\mathbf{M}^p$ among those constrained to lie wholly within $\mathbf{N}^q$. If g is the metric tensor associated with the coordinate system $(x_1, ..., x_p)$ then the induced metric tensor on $\mathbf{N}^q$ is constructed as the $q \times q$ matrix consisting of the first q rows and columns of g.

2.2.13 *Derivatives of Functions between Manifolds*

In 2.2.1, we defined the derivative of a differentiable function $h : U \rightarrow V$, where U and V are open sets of $\mathbf{R}^p$ and $\mathbf{R}^q$ respectively. We shall now

extend our definition to the case where h is defined between differential manifolds.

Let

$$h : \mathbf{M}^p \rightarrow \mathbf{N}^q \tag{2.54}$$

be a differentiable function, and suppose that $x(t)$ is a smooth path in $\mathbf{M}^p$. Then $h[x(t)]$ can be seen to be a smooth path in the manifold $\mathbf{N}^q$.

Differentiable mappings preserve tangency. For example, if x_0 is any point on the path $x(t)$, and if $y(t)$ is a path in $\mathbf{M}^p$ that is tangent to $x(t)$ at x_0, then $h[y(t)]$ is tangent to $h[x(t)]$ at the point $h(x_0) \in \mathbf{N}^q$. It follows from this fact that h maps the equivalence class of paths in $\mathbf{M}^p$ tangent to $x(t)$ at x_0 to the equivalence class of paths in $\mathbf{N}^q$ tangent to $h[x(t)]$ at $h(x_0)$. But these equivalence classes are tangent vectors at x_0 and $h(x_0)$ respectively. So this defines a mapping

$$(\mathcal{D}h)_x : T_x(\mathbf{M}^p) \rightarrow T_{h(x)}(\mathbf{N}^q) \tag{2.55}$$

Definition 2.2.7. *The mapping $(\mathcal{D}h)_x$ in formula (2.55) above is called the* derivative *of h at $x \in \mathbf{M}^p$, and can be shown to be a linear transformation between the tangent spaces.*

We can also define $(\mathcal{D}h)_x$ directly using coordinates on the manifold. In terms of the coordinates

$$x = (x_1, x_2, ..., x_p) \tag{2.56}$$

suppose that we can write $h(x)$ as

$$(h_1(x), h_2(x), ..., h_q(x)) \tag{2.57}$$

Let $\partial_1, ..., \partial_p$ be the coordinate basis of $T_x(\mathbf{M}^p)$, and correspondingly, let $\partial'_1, ..., \partial'_q$ be the coordinate basis for $T_{h(x)}(\mathbf{N}^q)$. Then $(\mathcal{D}h)_x$ can be expressed in terms of these basis vectors as

$$\sum_{j=1}^{p} a_j \, \partial_j \rightarrow \sum_{k=1}^{q} b_k \, \partial'_k \tag{2.58}$$

where

$$b_k = \sum_{j=1}^{p} a_j \frac{\partial h_k}{\partial x_j} \tag{2.59}$$

The expression can be seen to be left multiplication

$$a \rightarrow \Lambda a \tag{2.60}$$

where $a = (a_1, ..., a_p)$ is the row vector of coefficients and Λ is the Jacobian matrix of the coordinate transformation from $\mathbf{R}^p$ to $\mathbf{R}^q$.

2.2.14 *Example: The Sphere*

We finish this chapter with some examples of differential manifolds that will be useful in the next chapter. Examples of manifolds that are surfaces in $\mathbf{R}^3$ (and one surface that is not) can be found in Problems 2-6 at the end of the chapter.

In $\mathbf{R}^3$, let $\mathbf{S}^2(r)$, $r > 0$ be the set of all points $x = (x_1, x_2, x_3)$ such that

$$x_1^2 + x_2^2 + x_3^2 = r^2 \tag{2.61}$$

For notational simplicity, we typically let $\mathbf{S}^2$ denote the special case where $\mathbf{S}^2(r)$ has canonical radius $r = 1$. The set $\mathbf{S}^2(r)$ is called the 2-sphere of radius r. We can put an atlas on $\mathbf{S}^2(r)$ using the open sets U_{1+}, U_{1-}, and correspondingly the open sets U_{2+}, U_{2-} and U_{3+}, U_{3-}, where U_{j+} and U_{j-} are the set of points of $\mathbf{S}^2(r)$ with positive and negative x_j-coordinate respectively. To define a chart on U_{1+} we set

$$c_{1+}(x) \;=\; (x_2, x_3) \tag{2.62}$$

Similarly, we define $c_{1-}(x) = (x_2, x_3)$ on U_{1-}. Charts c_{2+}, c_{2-}, c_{3+}, and c_{3-} on the other open sets are defined correspondingly.

Although these coordinate systems establish $\mathbf{S}^2(r)$ as a differential manifold, there are more charts than necessary. A minimum of two charts is necessary to define an appropriate atlas on $\mathbf{S}^2(r)$ that corresponds to our intuitive understanding of the geometry of the sphere. For practical calculations, it is usually sufficient to set up a coordinate system through a single chart. These coordinates are the longitude θ_1 and the colatitude θ_2, defined so that the point

$$(r\,\cos(\theta_1)\,\sin(\theta_2),\; r\,\sin(\theta_1)\,\sin(\theta_2),\; r\,\cos(\theta_2)) \tag{2.63}$$

has coordinates (θ_1, θ_2).

To impose the usual metric of great circle distance on $\mathbf{S}^2(r)$ we introduce the metric tensor $g = (g_{jk})$ for the coordinate system $(\theta_1,\ \theta_2)$ where

$$g_{11} \;=\; r^2\,\sin^2(\theta_2) \tag{2.64}$$

and

$$g_{22} \;=\; r^2 \tag{2.65}$$

The off-diagonal elements $g_{12} = g_{21}$ are set to zero. The geodesics of the manifold can be shown to be *arcs of great circles.*

Extending to arbitrary dimensions is straightforward. In general, the p-sphere of radius r will be denoted $\mathbf{S}^p(r)$ and can be identified with the set of all points $(x_1,\ x_2,\ ...,\ x_p)$ in $\mathbf{R}^p$ such that

$$x_1^2 \;+\; x_2^2 \;+\; ... \;+\; x_p^2 \;=\; r^2 \tag{2.66}$$

Again, we let $\mathbf{S}^p$ denote the sphere of radius $r = 1$. An atlas

$$\begin{array}{cc} (U_{1+}, c_{1+}) & (U_{1-}, c_{1-}) \\ (U_{2+}, c_{2+}) & (U_{2-}, c_{2-}) \\ \cdots & \cdots \\ (U_{(p+1)+}, c_{(p+1)+}) & (U_{(p+1)-}, c_{(p+1)-}) \end{array} \tag{2.67}$$

can be imposed on $\mathbf{S}^p(r)$ in a manner similar to the 2-sphere above. The 1-sphere $\mathbf{S}^1$ is simply the unit circle.

The usual geodesic distance between two points of $\mathbf{S}^p$ is the shorter of the two arcs of the great circle joining the points. This is simply the angle made between the two vectors from the origin to the two points. Thus if x and y are elements of $\mathbf{S}^p \subset \mathbf{R}^{p+1}$ the geodesic distance from x to y is given by

$$d(x, y) \;=\; \cos^{-1}(< x, y >) \tag{2.68}$$

where again $< ., . >$ is the usual inner product on $\mathbf{R}^{p+1}$. More generally, on the sphere $\mathbf{S}^p(r)$, the geodesic distance from x to y is

$$d(x, y) \;=\; r\,\cos^{-1}\left(r^{-2} < x, y >\right) \tag{2.69}$$

The Cartesian product $\mathbf{S}^p \times \mathbf{S}^q$ of two spheres $\mathbf{S}^p$ and $\mathbf{S}^q$ is a generalization of a torus, which becomes the special case when $p = q = 1$. Although the representation of the torus $\mathbf{S}^1 \times \mathbf{S}^1$ is as a subset of $\mathbf{R}^4$, this torus is well known to be diffeomorphic to a surface in $\mathbf{R}^3$ that is the boundary of a doughnut. See Problem 2. However, the next example we shall consider is a two-dimensional manifold or surface that cannot be represented as a subset of $\mathbf{R}^3$.

2.2.15 *Example: Real Projective Spaces*

In $\mathbf{R}^3$, consider the set of all lines passing through the origin. Any such line can be represented as the set of scalar multiples

$$\{(\lambda x_1,\ \lambda x_2,\ \lambda x_3)\ :\ \lambda \in \mathbf{R}\} \tag{2.70}$$

for some nonzero element $(x_1, x_2, x_3) \in \mathbf{R}^3$.

Definition 2.2.8. *We call the set of such lines through the origin* real projective 2-space *and symbolize it as* $\mathbf{RP}^2$. *As any line through the origin meets the unit sphere about the origin in exactly two antipodal points, it can be seen that real projective 2-space is naturally identifiable with the set of all pairs of antipodal points on the unit sphere.*

See Figure 2.6. This representation is particularly useful in making $\mathbf{RP}^2$

those of $\mathbf{M}^p$ and $\mathbf{N}^q$ so that

$$T_{(x,y)}(\mathbf{M}^p \times \mathbf{N}^q) = T_x(\mathbf{M}^p) \times T_y(\mathbf{N}^q) \tag{2.50}$$

With this understanding, we can make $\mathbf{M}^p \times \mathbf{N}^q$ into a Riemannian manifold by putting the metric tensor elements as blocks down the main diagonal. If g_M is a metric tensor on $\mathbf{M}^p$ and g_N is a metric tensor on $\mathbf{N}^q$ then an appropriate metric on $\mathbf{M}^p \times \mathbf{N}^q$ is

$$\left(\begin{array}{c|c} g_M & 0 \\ \hline 0 & g_N \end{array}\right) \tag{2.51}$$

2.2.12 *Building New Manifolds From Old: Submanifolds*

It is also possible to construct new manifolds by looking inside a manifold. Suppose $\mathbf{N}^q$ is a subset of a differential manifold $\mathbf{M}^p$. We say that $\mathbf{N}^q$ is a *q-dimensional submanifold* of $\mathbf{M}^p$ for $q < p$ if for every point $y \in \mathbf{N}^q$, there exists a smooth coordinate system $x = (x_1, ..., x_p)$ on some open set $U \subset \mathbf{M}^p$ containing y such that

$$U \cap \mathbf{N}^q = \{x \in U : x_{q+1} = x_{q+2} = ... = x_p = 0\} \tag{2.52}$$

More informally we can say that a q-dimensional submanifold of $\mathbf{M}^p$ is a subset that is locally diffeomorphic to a linear subspace. The submanifold $\mathbf{N}^q$ inherits a coordinate system from this construction. The coordinates

$$x \to (x_1, x_2, ..., x_q) \tag{2.53}$$

make a local smooth coordinate system of the right dimension on the submanifold.

Using the coordinate system $x = (x_1, x_2, ..., x_p)$ we can set up the basis $\partial_1, \partial_2, ..., \partial_p$ for the tangent space $T_x(\mathbf{M}^p)$. Among these basis vectors, the first q tangent vectors $\partial_1, \partial_2, ..., \partial_q$ form a basis for the tangent space $T_x(\mathbf{N}^q)$. Thus any tangent vector in $T_x(\mathbf{N}^q)$ can be written as $\sum_{j=1}^{q} a_j(x)\partial_j(x)$. If $\mathbf{M}^p$ is a Riemannian manifold, then $\mathbf{N}^q$ can be made into a Riemannian manifold by inheriting the concept of arc length from $\mathbf{M}^p$. A geodesic path in $\mathbf{N}^q$ is simply a path of shortest length in $\mathbf{M}^p$ among those constrained to lie wholly within $\mathbf{N}^q$. If g is the metric tensor associated with the coordinate system $(x_1, ..., x_p)$ then the induced metric tensor on $\mathbf{N}^q$ is constructed as the $q \times q$ matrix consisting of the first q rows and columns of g.

2.2.13 *Derivatives of Functions between Manifolds*

In 2.2.1, we defined the derivative of a differentiable function $h : U \to V$, where U and V are open sets of $\mathbf{R}^p$ and $\mathbf{R}^q$ respectively. We shall now

extend our definition to the case where h is defined between differential manifolds.

Let

$$h : \mathbf{M}^p \to \mathbf{N}^q \tag{2.54}$$

be a differentiable function, and suppose that $x(t)$ is a smooth path in $\mathbf{M}^p$. Then $h[x(t)]$ can be seen to be a smooth path in the manifold $\mathbf{N}^q$.

Differentiable mappings preserve tangency. For example, if x_0 is any point on the path $x(t)$, and if $y(t)$ is a path in $\mathbf{M}^p$ that is tangent to $x(t)$ at x_0, then $h[y(t)]$ is tangent to $h[x(t)]$ at the point $h(x_0) \in \mathbf{N}^q$. It follows from this fact that h maps the equivalence class of paths in $\mathbf{M}^p$ tangent to $x(t)$ at x_0 to the equivalence class of paths in $\mathbf{N}^q$ tangent to $h[x(t)]$ at $h(x_0)$. But these equivalence classes are tangent vectors at x_0 and $h(x_0)$ respectively. So this defines a mapping

$$(\mathcal{D}h)_x : T_x(\mathbf{M}^p) \to T_{h(x)}(\mathbf{N}^q) \tag{2.55}$$

Definition 2.2.7. *The mapping $(\mathcal{D}h)_x$ in formula (2.55) above is called the* derivative *of h at $x \in \mathbf{M}^p$, and can be shown to be a linear transformation between the tangent spaces.*

We can also define $(\mathcal{D}h)_x$ directly using coordinates on the manifold. In terms of the coordinates

$$x = (x_1, x_2, ..., x_p) \tag{2.56}$$

suppose that we can write $h(x)$ as

$$(h_1(x), h_2(x), ..., h_q(x)) \tag{2.57}$$

Let $\partial_1, ..., \partial_p$ be the coordinate basis of $T_x(\mathbf{M}^p)$, and correspondingly, let $\partial_1', ..., \partial_q'$ be the coordinate basis for $T_{h(x)}(\mathbf{N}^q)$. Then $(\mathcal{D}h)_x$ can be expressed in terms of these basis vectors as

$$\sum_{j=1}^{p} a_j \, \partial_j \to \sum_{k=1}^{q} b_k \, \partial_k' \tag{2.58}$$

where

$$b_k = \sum_{j=1}^{p} a_j \frac{\partial h_k}{\partial x_j} \tag{2.59}$$

The expression can be seen to be left multiplication

$$a \to \Lambda a \tag{2.60}$$

where $a = (a_1, ..., a_p)$ is the row vector of coefficients and Λ is the Jacobian matrix of the coordinate transformation from $\mathbf{R}^p$ to $\mathbf{R}^q$.

2.2.14 *Example: The Sphere*

We finish this chapter with some examples of differential manifolds that will be useful in the next chapter. Examples of manifolds that are surfaces in $\mathbf{R}^3$ (and one surface that is not) can be found in Problems 2-6 at the end of the chapter.

In $\mathbf{R}^3$, let $\mathbf{S}^2(r)$, $r > 0$ be the set of all points $x = (x_1, x_2, x_3)$ such that

$$x_1^2 + x_2^2 + x_3^2 = r^2 \tag{2.61}$$

For notational simplicity, we typically let $\mathbf{S}^2$ denote the special case where $\mathbf{S}^2(r)$ has canonical radius $r = 1$. The set $\mathbf{S}^2(r)$ is called the 2-sphere of radius r. We can put an atlas on $\mathbf{S}^2(r)$ using the open sets U_{1+}, U_{1-}, and correspondingly the open sets U_{2+}, U_{2-} and U_{3+}, U_{3-}, where U_{j+} and U_{j-} are the set of points of $\mathbf{S}^2(r)$ with positive and negative x_j-coordinate respectively. To define a chart on U_{1+} we set

$$c_{1+}(x) = (x_2, x_3) \tag{2.62}$$

Similarly, we define $c_{1-}(x) = (x_2, x_3)$ on U_{1-}. Charts c_{2+}, c_{2-}, c_{3+}, and c_{3-} on the other open sets are defined correspondingly.

Although these coordinate systems establish $\mathbf{S}^2(r)$ as a differential manifold, there are more charts than necessary. A minimum of two charts is necessary to define an appropriate atlas on $\mathbf{S}^2(r)$ that corresponds to our intuitive understanding of the geometry of the sphere. For practical calculations, it is usually sufficient to set up a coordinate system through a single chart. These coordinates are the longitude θ_1 and the colatitude θ_2, defined so that the point

$$(r\cos(\theta_1)\sin(\theta_2),\ r\sin(\theta_1)\sin(\theta_2),\ r\cos(\theta_2)) \tag{2.63}$$

has coordinates (θ_1, θ_2).

To impose the usual metric of great circle distance on $\mathbf{S}^2(r)$ we introduce the metric tensor $g = (g_{jk})$ for the coordinate system (θ_1, θ_2) where

$$g_{11} = r^2 \sin^2(\theta_2) \tag{2.64}$$

and

$$g_{22} = r^2 \tag{2.65}$$

The off-diagonal elements $g_{12} = g_{21}$ are set to zero. The geodesics of the manifold can be shown to be *arcs of great circles.*

Extending to arbitrary dimensions is straightforward. In general, the p-sphere of radius r will be denoted $\mathbf{S}^p(r)$ and can be identified with the set of all points $(x_1, x_2, \ldots, x_p)$ in $\mathbf{R}^p$ such that

$$x_1^2 + x_2^2 + \ldots + x_p^2 = r^2 \tag{2.66}$$

Again, we let $\mathbf{S}^p$ denote the sphere of radius $r = 1$. An atlas

$$\begin{array}{cc} (U_{1+}, c_{1+}) & (U_{1-}, c_{1-}) \\ (U_{2+}, c_{2+}) & (U_{2-}, c_{2-}) \\ \ldots & \ldots \\ (U_{(p+1)+}, c_{(p+1)+}) & (U_{(p+1)-}, c_{(p+1)-}) \end{array} \tag{2.67}$$

can be imposed on $\mathbf{S}^p(r)$ in a manner similar to the 2-sphere above. The 1-sphere $\mathbf{S}^1$ is simply the unit circle.

The usual geodesic distance between two points of $\mathbf{S}^p$ is the shorter of the two arcs of the great circle joining the points. This is simply the angle made between the two vectors from the origin to the two points. Thus if x and y are elements of $\mathbf{S}^p \subset \mathbf{R}^{p+1}$ the geodesic distance from x to y is given by

$$d(x, y) = \cos^{-1}(< x, y >) \tag{2.68}$$

where again $< ., . >$ is the usual inner product on $\mathbf{R}^{p+1}$. More generally, on the sphere $\mathbf{S}^p(r)$, the geodesic distance from x to y is

$$d(x, y) = r\cos^{-1}(r^{-2} < x, y >) \tag{2.69}$$

The Cartesian product $\mathbf{S}^p \times \mathbf{S}^q$ of two spheres $\mathbf{S}^p$ and $\mathbf{S}^q$ is a generalization of a torus, which becomes the special case when $p = q = 1$. Although the representation of the torus $\mathbf{S}^1 \times \mathbf{S}^1$ is as a subset of $\mathbf{R}^4$, this torus is well known to be diffeomorphic to a surface in $\mathbf{R}^3$ that is the boundary of a doughnut. See Problem 2. However, the next example we shall consider is a two-dimensional manifold or surface that cannot be represented as a subset of $\mathbf{R}^3$.

2.2.15 *Example: Real Projective Spaces*

In $\mathbf{R}^3$, consider the set of all lines passing through the origin. Any such line can be represented as the set of scalar multiples

$$\{(\lambda x_1, \lambda x_2, \lambda x_3) : \lambda \in \mathbf{R}\} \tag{2.70}$$

for some nonzero element $(x_1, x_2, x_3) \in \mathbf{R}^3$.

Definition 2.2.8. *We call the set of such lines through the origin* real projective 2-space *and symbolize it as* $\mathbf{RP}^2$. *As any line through the origin meets the unit sphere about the origin in exactly two antipodal points, it can be seen that real projective 2-space is naturally identifiable with the set of all pairs of antipodal points on the unit sphere.*

See Figure 2.6. This representation is particularly useful in making $\mathbf{RP}^2$

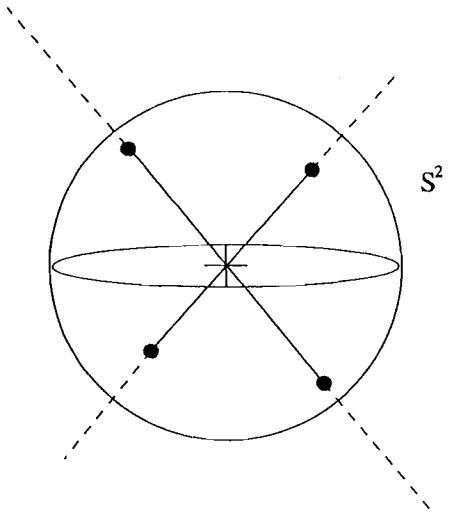

FIGURE 2.6. *Real projective 2-space represented as the space of lines passing through the origin in 3-dimensional Euclidean space. Each line defines a pair of antipodal points on the unit sphere. Therefore a 1-1 correspondence exists between such lines and pairs of antipodal points.*

into a differential manifold. Note that there is a *natural mapping*

$$\mathcal{A} \; : \; \mathbf{S}^2 \rightarrow \mathbf{RP}^2 \tag{2.71}$$

that maps any point of the unit sphere to the set of two antipodal points of which it is an element. This mapping is an example of a special type of differentiable function between manifolds called a *covering mapping*. The image under $\mathcal{A}$ of any point $x = (x_1, x_2, x_3) \in \mathbf{S}^2$ is the pair of antipodal points

$$\mathcal{A}(x) \; = \; \{x \; , \; -x\} \tag{2.72}$$

To make $\mathbf{RP}^2$ into a differential manifold, we can modify the charts of Example 2.2.1. Note that

$$\mathcal{A}(U_{1+}) \; = \; \mathcal{A}(U_{1-}) \tag{2.73}$$

Similarly, we have

$$\mathcal{A}(U_{2+}) \; = \; \mathcal{A}(U_{2-}) \text{ and } \mathcal{A}(U_{3+}) \; = \; \mathcal{A}(U_{3-}) \tag{2.74}$$

These three sets are open in the natural topology on $\mathbf{RP}^2$. For any real number $a \neq 0$, let $\text{sgn}(a)$ denote the sign of the number a. We construct charts

$$(U_1, \; c_1) \; , \; \; (U_2, \; c_2) \; , \; \; (U_3, c_3) \tag{2.75}$$

on $\mathbf{RP}^2$ by defining the three open sets

$$U_1 = \mathcal{A}(U_{1+}), \; U_2 = \mathcal{A}(U_{2+}), \; U_3 = \mathcal{A}(U_{3+}) \tag{2.76}$$

where U_{j+} and U_{j-} are as defined for the sphere $\mathbf{S}^2$ above. Define $c_1 : U_1 \rightarrow \mathbf{R}^2$ to be

$$\mathcal{A}(x) \; \rightarrow \; (\text{sgn}(x_1)x_2, \; \text{sgn}(x_1)x_3) \tag{2.77}$$

Similarly, we define $c_2 : U_2 \rightarrow \mathbf{R}^2$ to be

$$\mathcal{A}(x) \; \rightarrow \; (\text{sgn}(x_2)x_1, \; \text{sgn}(x_2)x_3) \tag{2.78}$$

and $c_3 : U_3 \rightarrow \mathbf{R}^2$ to be

$$\mathcal{A}(x) \; \rightarrow \; (\text{sgn}(x_1)x_2, \; \text{sgn}(x_1)x_3) \tag{2.79}$$

This particular differential structure on $\mathbf{RP}^2$ has the property that a function $f : \mathbf{RP}^2 \rightarrow \mathbf{R}$ is differentiable if and only if the function

$$f \circ \mathcal{A} : \mathbf{S}^2 \; \rightarrow \; \mathbf{R} \tag{2.80}$$

is differentiable. Another property of the covering mapping $\mathcal{A}$ is that its derivative

$$(\mathcal{D}\mathcal{A})_x \; : \; T_x(\mathbf{S}^2) \; \rightarrow \; T_{\mathcal{A}(x)}(\mathbf{RP}^2) \tag{2.81}$$

is a linear transformation of full rank, i.e., is onto, at all points $x \in \mathbf{S}^2$.

This fact can be used to motivate a particular choice of metric tensor on $\mathbf{RP}^2$. As $(\mathcal{D}\mathcal{A})_x$ maps onto $T_{\mathcal{A}(x)}(\mathbf{RP}^2)$, we can write any element of this tangent space as

$$(\mathcal{D}\mathcal{A})_x(a_1 \; \partial_1 \; + \; a_2 \; \partial_2) \; = \; a_1 \; (\mathcal{D}\mathcal{A})_x(\partial_1) \; + \; a_2 \; (\mathcal{D}\mathcal{A})_x(\partial_2) \tag{2.82}$$

where ∂_1 and ∂_2 form a coordinate basis for $T_x(\mathbf{S}^2)$. Let $\partial_j' = (\mathcal{D}\mathcal{A})_x(\partial_j)$ for $j = 1, 2$. Then ∂_1' and ∂_2' form a basis for $T_{\mathcal{A}(x)}(\mathbf{RP}^2)$. If in addition, we set

$$< \partial_j', \; \partial_k' > = < \partial_j, \; \partial_k > = \; g_{jk} \tag{2.83}$$

then

$$(\mathcal{D}\mathcal{A})_x \; : \; T_x(\mathbf{S}^2) \; \rightarrow \; T_{\mathcal{A}(x)}(\mathbf{RP}^2) \tag{2.84}$$

becomes a *linear isometry* between tangent spaces.

With this metric tensor on $\mathbf{RP}^2$, the covering map $\mathcal{A}$ maps the geodesic great circles of $\mathbf{S}^2$ to geodesic paths in $\mathbf{RP}^2$ and $\mathcal{A}$ becomes a local isometry between Riemannian manifolds. That is, if x and y are points of $\mathbf{S}^2$ separated by a geodesic distance of less than $\pi/2$ then the geodesic distance from x to y in $\mathbf{S}^2$ equals the geodesic distance from $\mathcal{A}(x)$ to $\mathcal{A}(y)$ in $\mathbf{RP}^2$. However, the two manifolds are not isometric because $\mathcal{A}$ is not 1-1.

Extensions, some straightforward and others more substantial, are possible.

Definition 2.2.9. *We define* real projective p-space, *denoted by* $\mathbf{RP}^p$, *to be the space of lines through the origin in* $\mathbf{R}^{p+1}$. *This space can be interpreted as the set of antipodal pairs of points on the* p*-dimensional unit sphere* $\mathbf{S}^p \subset \mathbf{R}^{p+1}$.

The constructions above generalize in a natural way. Again, the covering map $\mathcal{A} : \mathbf{S}^p \to \mathbf{RP}^p$ establishes a *local isometry* between $\mathbf{S}^p$ and $\mathbf{RP}^p$ that is not an isometry. To visualize what this means, consider the case where $p = 1$. In this case, it can be shown that $\mathbf{RP}^1$ is a circle of radius $1/2$. We can establish a local isometry between a circle

$$\mathbf{S}^1 = \mathbf{S}^1(1) \quad \text{and} \quad \mathbf{RP}^1 \cong \mathbf{S}^1(1/2) \tag{2.85}$$

by noting that the covering map $\mathcal{A}$ wraps the circle $\mathbf{S}^1$, whose circumference has length 2π twice around the circle $\mathbf{S}^1(1/2)$, whose circumference has length π. This is akin to winding a thread tightly twice around a spool and then joining the ends of the thread to form a loop. The change of radius by a factor of one half is a natural consequence of the fact that the covering mapping $\mathcal{A}$ is 2 to 1.

A consequence of this construction is that the geodesic distance between two points $\mathcal{A}(x)$ and $\mathcal{A}(y)$ in $\mathbf{RP}^p$ is found to be the smaller of the two geodesic distances from $d(x, y)$ and $d(x, -y)$ in $\mathbf{S}^p$. This can be used as a definition of the metric on $\mathbf{RP}^p$ without reference to the metric tensor. Furthermore, the image under the mapping $\mathcal{A}$ of a geodesic great circle path in $\mathbf{S}^p$ is a geodesic path in $\mathbf{RP}^p$.

The projective spaces $\mathbf{RP}^p$ have a role in the representation of shapes. In Section 1.4.1, we noted that the pre-shape of n landmarks along a line can be represented as a point on a sphere of dimension $n - 2$. Antipodal points on this sphere represent the pre-shapes of reflected configurations. For example, if x_1, x_2, and x_3 are three landmarks on the real line, then the pre-shapes of (x_1, x_2, x_3) and $(-x_1, -x_2, -x_3)$ are antipodal points on the circle. This can be seen in Figure 1.3, where the three pairs of antipodal points displayed are the pre-shapes of equally spaced landmarks. So the projective spaces $\mathbf{RP}^{n-2}$ are appropriate manifolds for representing the pre-shapes of aligned landmarks when the distinction between a configuration and its reflection is ignored. This is particularly appropriate for the example in Section 1.4.1, where a reflection along the line of alignment corresponds to a rotation by $180°$ in the plane in which the images lie. We can summarize this by saying that *the shapes of* $n \geq 3$ *aligned landmarks in the plane can be naturally represented as elements of the real projective space* $\mathbf{RP}^{n-2}$.

A variant of real projective space, which we shall consider next, is obtained by replacing the real coordinates of Euclidean space $\mathbf{R}^{p+1}$ with complex coordinates. We shall encounter this space in the context of shape manifolds in Section 3.2 in the next chapter.

2.2.16 *Example: Complex Projective Spaces*

The manifolds that we shall consider next will be written with complex coordinates in what follows. However, they can be understood as examples of the differential manifolds that we have been discussing up to now. This can be seen through the identification of $\mathbf{R}^2$ with the complex plane $\mathbf{C}$. The differential manifold $\mathbf{CP}^p$ that we shall consider will have p complex dimensions or equivalently $2p$ real dimensions. It can be regarded as a collection of complex lines through the origin in $\mathbf{C}^{p+1}$ or as a collection of planes through the origin in $\mathbf{R}^{2p+2}$. Note that the latter interpretation has to be made with some care after identifying $\mathbf{R}^2$ with $\mathbf{C}$. Every complex line through the origin of $\mathbf{C}^p$ can be considered as a plane through the origin in $\mathbf{R}^{2p+2}$. However, the converse is not true. Similarly, we saw earlier that every unitary transformation of $\mathbf{C}^p$ is an orthogonal transformation of $\mathbf{R}^{2p}$, without the converse holding.

Let $\mathbf{CP}^p$ be the collection of all complex lines

$$\{(\lambda z_1, \lambda z_2, ..., \lambda z_{p+1}) \ : \ \text{for all} \ \ \lambda \in \mathbf{C}\} \tag{2.86}$$

found by taking a point $(z_1, z_2, ..., z_{p+1}) \in \mathbf{C}^{p+1}$ distinct from the origin and drawing the complex line through this point and the origin $(0, 0, ..., 0)$. Any such complex line intersects the sphere

$$\mathbf{S}^{2p+1} = \left\{(z_1, z_2, ..., z_{p+1}) \ : \ \sum |z_j|^2 = 1\right\} \tag{2.87}$$

in a great circle. These circles partition the sphere, so that any point $(z_1, z_2, ..., z_{p+1})$ in $\mathbf{S}^{2p+1}$ will be an element of a unique circle of the form

$$\{(\lambda z_1, \lambda z_2, ..., \lambda z_{p+1}) : \text{for all } \lambda \in \mathbf{C} \text{ such that } |\lambda| = 1\} \tag{2.88}$$

Let us call this circle $\mathcal{O}(z_1, z_2, ..., z_{p+1})$. The use of the symbol $\mathcal{O}$ reflects the fact that these great circles are *orbits*, or equivalence classes, in the terminology of differential geometry. To a certain extent our low-dimensional intuition fails us here, because we are used to having geodesic great circles of the 2-sphere $\mathbf{S}^2$ always intersecting. However, the spheres we are considering are 3-spheres or of higher dimension. The additional room that this provides allows for the partition of the spheres (in certain cases) into great circles.

We can build charts on the set of such great circles as follows: For $j = 1, 2, ..., p + 1$ let U_j be the set

$$U_j \ = \ \{\mathcal{O}(z_1, ..., z_j, ..., z_{p+1}) \in \mathbf{CP}^p \ : \ z_j \neq 0\} \tag{2.89}$$

On U_j we can set up the coordinates

$$\mathcal{O}(z_1, ..., z_j, ..., z_{p+1}) \to (z_1/z_j, ..., z_{j-1}/z_j, z_{j+1}/z_j, ..., z_{p+1}/z_j) \quad (2.90)$$

This coordinate system maps the open set U_j onto $\mathbf{C}^p \cong \mathbf{R}^{2p}$. Patching these charts together makes $\mathbf{CP}^p$ into a differential manifold.

We can summarize this as follows:

Definition 2.2.10. *We define* complex projective p-space, *denoted by* $\mathbf{CP}^p$, *to be the set of complex lines through the origin in* $\mathbf{C}^{p+1}$ *as in formula (2.86) above. This space can be naturally identified with the set of great circles of* $\mathbf{S}^{2p+1}$ *defined by formula (2.88).*

It remains to construct a metric on the manifold $\mathbf{CP}^p$. Rather than beginning at the local level, so to speak, with the construction of the metric tensor, it will be more convenient to define geodesic distance globally on $\mathbf{CP}^p$ and to note that it leads to a Riemannian geometry on the differential manifold. Let us contract our notation a bit more here by letting z stand for the full vector $(z_1, ..., z_{p+1})$, which lies in $\mathbf{S}^{2p+1}$. Similarly $\mathcal{O}(z)$ will be the element of $\mathbf{CP}^p$ in which z lies. Now suppose we wish to define the geodesic distance between two elements $\mathcal{O}(z)$ and $\mathcal{O}(w)$ of $\mathbf{CP}^p$. Write $w = (w_1, ..., w_{p+1})$. We could naturally define the distance between $\mathcal{O}(z)$ and $\mathcal{O}(w)$ to be

$$d[\mathcal{O}(z), \mathcal{O}(w)] = \inf [d(x, y)] : x \in \mathcal{O}(z); y \in \mathcal{O}(w)] \quad (2.91)$$

where $d(x, y)$ is the geodesic distance on $\mathbf{S}^{2p+1}$ from x to y. We can intuitively think of this formula as saying that the distance from one great circle to another is the shortest gap between them. See Figure 2.7. Now, while this is a perfectly well-defined quantity, there is no reason *a priori* to suppose that this satisfies the properties that a distance measure, or metric, has. In particular, the triangle inequality has to be checked carefully. The triangle inequality does hold, in a sense, because of the symmetry of the sphere $\mathbf{S}^{2p+1}$. The minimum can be achieved at every value of x by minimizing over y, or correspondingly, at every value of y by minimizing over x.

The reader should note the similarity between our construction here and the Procrustean minimization of formula (1.18) in Section 1.3 of the previous chapter. The differences in notation and context should not disguise the fact that the geometric situations are equivalent. In Section 1.2, the points on the sphere were pre-shapes and the orbits or great circles were shapes. We proceed similarly. Writing the geodesic distance on $\mathbf{S}^{2p+1}$ explicitly, we have

$$d[\mathcal{O}(z), \mathcal{O}(w)] = \inf [\cos^{-1}(< x, y >) : x \in \mathcal{O}(z), y \in \mathcal{O}(w)] \quad (2.92)$$

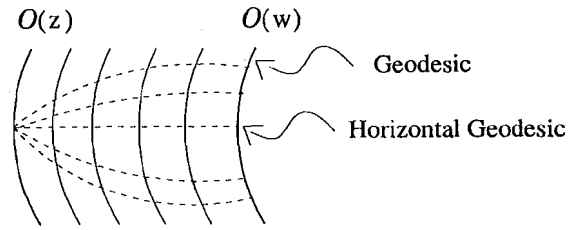

FIGURE 2.7. *Complex projective space represented as a space of circles on the sphere. In this picture, a small portion of a sphere is seen with circular arcs (the orbits) displayed vertically. From a given point on the left orbit of the picture, a variety of geodesics (great circle arcs) can be drawn to the right orbit. A "horizontal" geodesic will have the shortest length and will meet "vertical" orbits at right angles. The distance between two orbits is the shortest great circle path from one arc to another.*

The inner product we are working with here is the inner product on $\mathbf{S}^{2p+1}$ as embedded in $\mathbf{R}^{2p+2}$. We can write this in terms of the Hermitian inner product on $\mathbf{S}^{2p+1}$ as embedded in $\mathbf{C}^{p+1}$. This becomes

$$< x, y > = \Re(\ll x, y \gg) = \Re\left(\sum_j x_j y_j^*\right) \quad (2.93)$$

where x_j and y_j are the jth complex coordinates of x and y respectively. The next observation we make is that the minimization can be achieved by fixing $x = z$ and writing $y_j = e^{i\theta} w_j$, where $i = \sqrt{-1}$, minimizing over $0 \leq \theta < 2\pi$. Thus

$$d[\mathcal{O}(z), \mathcal{O}(w)] = \inf \left\{ \cos^{-1} \left[\Re \sum_{j=1}^{p+1} z_j (e^{-i\theta} w_j^*) \right] : 0 \leq \theta < 2\pi \right\} \quad (2.94)$$

We can perform the minimization by maximizing the sum with respect to θ. Now

$$\Re\left[e^{-i\theta}(z_j w_j^*)\right] = \cos(\theta)\Re(z_j w_j^*) + \sin(\theta)\Im(z_j w_j^*) \quad (2.95)$$

and so the maximum can be found by differentiating with respect to θ and setting the result equal to zero. This yields

$$e^{i\theta} = \frac{\sum_{j=1}^{p+1} z_j w_j^*}{|\sum_{j=1}^{p+1} z_j w_j^*|} \quad (2.96)$$

Plugging this in, we see that

$$d[\mathcal{O}(z_1), \mathcal{O}(z_2)] = \cos^{-1}\left(|\sum_{j=1}^{p+1} z_j w_j^*|\right) \quad (2.97)$$

This is the famous *Fubini-Study metric* on $\mathbf{CP}^p$. As in Section 1.3, when considering distances between shapes, we note that the maximum distance between elements of $\mathbf{CP}^p$ is $\pi/2$. In addition, the right-hand side in this distance formula does not depend upon the specific choice of z and w within the orbits $\mathcal{O}(z)$ and $\mathcal{O}(w)$. The modulus operation nullifies the effect of this selection, which corresponds to multiplication of the coordinates by a common complex factor of modulus one.

As this provides us with a metric on $\mathbf{CP}^p$ we can now consider the geodesics on this manifold. In Section 2.2.15, we found that the geodesics on $\mathbf{RP}^p$ were images under the covering map $\mathcal{A}$ of geodesic great circle paths of $\mathbf{S}^p$. It is natural to consider whether this is the case here. In fact, the geodesics of $\mathbf{CP}^p$ are images of geodesics on $\mathbf{S}^{2p+1}$. However, they are images of particular geodesics called *horizontal* geodesics. Intuitively, we think of the orbits of $\mathbf{S}^{2p+1}$ as arranged *vertically* with the mapping $\mathbf{S}^{2p+1} \rightarrow \mathbf{CP}^p$ as mapping *downwards*. Thus the horizontal geodesics are always perpendicular to the orbits. See Figure 2.7. These geodesics are great circle paths of $\mathbf{S}^{2p+1}$ with the property that they intersect the orbits $\mathcal{O}(z)$ orthogonally. More precisely, we can say that a great circle path $z(t)$ is horizontal if for every t the tangent vector $\dot{z}(t)$ is orthogonal to the vectors of the tangent space of $\mathcal{O}[z(t)]$. It is not the case, in general, that any two points in $\mathbf{S}^{2p+1}$ can be joined by a horizontal geodesic. However, if z and w are chosen from $\mathcal{O}(z)$ and $\mathcal{O}(w)$, respectively, so as to minimize the geodesic distance as above, then z and w can be joined by a horizontal geodesic. The construction of horizontal geodesics will play an important role in Chapter 3, where we shall consider them in greater detail.

2.2.17 *Example: Hyperbolic Half Spaces*

Consider the Riemannian manifold consisting of the upper half space in $\mathbf{R}^p$ given by

$$\mathbf{HS}^p = \{(x_1, x_2, ..., x_p) : x_p > 0\} \tag{2.98}$$

and metric tensor

$$g_{jj}(x_1, ..., x_p) = x_p^{-2} \tag{2.99}$$

for all $j = 1, ..., p$ and $g_{jk} = 0$ for all $1 \leq j \neq k \leq p$. The reader will notice the similarity between this space and ordinary Euclidean space. The major difference is the appearance of the last coordinate in the denominator of the diagonal terms of the metric tensor.

Definition 2.2.11. *The space* $\mathbf{HS}^p$ *with the metric tensor of formula (2.99) is called the* hyperbolic half space *of dimension* p.

The family of hyperbolic half spaces $\mathbf{HS}^p$ represents the negative curvature counterpart of the family of positively curved spheres $\mathbf{S}^p$. Solving

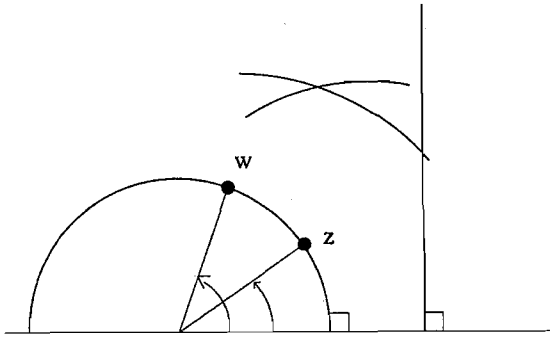

FIGURE 2.8. *The Poincaré Plane. Geodesic paths in the Poincaré Plane are the arcs of circles that meet the x-axis at right angles. In the limiting form, as the radius goes to infinity, these circles become vertical lines, which are also geodesics.*

the Euler-Lagrange equations, we find that the geodesic paths of $\mathbf{HS}^p$ are half circles or lines that meet the boundary $x_p = 0$ orthogonally. An important special case is $p = 2$, which is called the *Poincaré Plane*. See Figure 2.8.

It is convenient to represent $\mathbf{HS}^2$ using complex coordinates as

$$\mathbf{HS}^2 = \{z \in \mathbf{C} : \Im(z) > 0\} \tag{2.100}$$

with

$$ds^2 = \frac{|dz|^2}{[\Im(z)]^2} \tag{2.101}$$

Using this complex notation, we can calculate the geodesic distance between two points z and w in $\mathbf{HS}^2$ by integrating ds, given by formula (2.101), along a geodesic path from z to w. As we noted above, these geodesics are circles that are orthogonal to the real axis (with vertical straight lines as the limiting case). Let z and w lie on a geodesic circle centered at a with radius r. As the circle is orthogonal to the real axis, the point a must be a real number. Let rays be drawn from a to z and w making counterclockwise angles β_z and β_w with the real axis. A simple calculation will show that the geodesic distance from z to w is a function of β_z and β_w alone, the quantities a and r disappearing from the final answer. To see this, note that we can write

$$z = a + r\cos(\beta_z) + i\, r \sin(\beta_z) \tag{2.102}$$

and

$$w = a + r\cos(\beta_w) + i\, r \sin(\beta_w) \tag{2.103}$$

where $i = \sqrt{-1}$. Then the geodesic distance from z to w is given by

$$\int_z^w ds = \int_{\beta_z}^{\beta_w} \csc(\beta)\, d\beta = \log\left\{\frac{[1-\cos(\beta_w)]\sin(\beta_z)}{[1-\cos(\beta_z)]\sin(\beta_w)}\right\} \tag{2.104}$$

where we choose the direction of integration so that $0 \leq \beta_z < \beta_w \leq \pi$.

From formula (2.104), we can see that the real axis $\Im(z) = 0$ is not really a boundary at all, but rather an infinite horizon. Half circles that are geodesics in the upper half plane $\mathbf{HS}^2$ have finite length as measured by Euclidean geometry, but have infinite length when measured using the hyperbolic formula of (2.104). The difference is a consequence of the appearance of $\Im(z)$ in the denominator of the formula for the metric tensor. This has the effect of greatly inflating distances compared to the Euclidean metric between points close to the real axis.

As we noted above, the geodesic curves of $\mathbf{HS}^2$ include not only the circles of the half plane that meet the real axis orthogonally, but also vertical lines of the form $\Re(z) =$ constant. These can be thought of as geodesic circles that have infinite radius. For points z and w connected by such a geodesic, formula (2.46) must be interpreted with some care. As $\Re(z) = \Re(w)$, we can simply integrate formula (2.101) along the imaginary coordinates on which they differ. Alternatively, we can take the limiting form of formula (2.104). In either case, we find that the geodesic distance from z to w is equal to

$$\int_z^w ds = \int_{\Im(z)}^{\Im(w)} \frac{du}{u} = \log\left[\frac{\Im(w)}{\Im(z)}\right] \tag{2.105}$$

where $\Im(w) \geq \Im(z) > 0$. This particular formula will have an important role to play when we examine shape variation due to affine transformations in Chapter 3.

The transformation

$$z \rightarrow i\,\frac{z-i}{z+i} \tag{2.106}$$

maps the points of the Poincaré Plane $\mathbf{HS}^2$ onto the the *Poincaré Disk*

$$\mathbf{HD}^2 = \{w \in \mathbf{C} \,:\, |w| < 1\} \tag{2.107}$$

See Figure 2.9. This mapping defines an isometry between $\mathbf{HS}^2$ and $\mathbf{HD}^2$ when the disk is endowed with the metric

$$ds^2 = \frac{4\,|dw|^2}{[1-|w|^2]^2} \tag{2.108}$$

This formula can be derived in a straightforward manner by doing a change of variables from z to w on formula (2.101).

It can be checked that the real axis of the Poincaré Plane is mapped to the circle $|w| = 1$ on the Poincaré Disk. This circle becomes its *circle at infinity*. In addition, the geodesic half circles and lines of $\mathbf{HS}^2$ are mapped to circles and lines in $\mathbf{HD}^2$ that are orthogonal to the circle $|w| = 1$.

As an additional way of representing the geometry of $\mathbf{HS}^2$, we might wish to construct a curved surface in $\mathbf{R}^3$ and a correspondence between

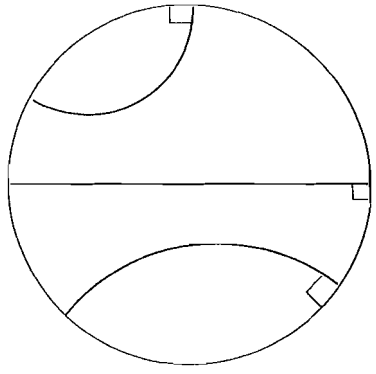

FIGURE 2.9. *The Poincaré Disk. Geodesic paths in the Poincaré Disk are the arcs of circles that meet the boundary of the disk at right angles. In the limiting form, as the radius of these arcs goes to infinity, they become diameters of the disk and are also geodesics.*

this surface and $\mathbf{HS}^2$ so that the geodesics of $\mathbf{HS}^2$ correspond to the geodesics of the curved surface. To do this, we can construct the *Poincaré Trumpet* $\mathbf{HT}^2$. It is convenient to use real coordinates (x_1, x_2) on the upper half plane of $\mathbf{R}^2$ in this case. For arbitrary $\epsilon > 0$, define the functions

$$f_1, f_2 : \mathbf{R} \rightarrow \mathbf{R} \tag{2.109}$$

by

$$f_1(x) = \frac{\epsilon}{x} \tag{2.110}$$

and

$$f_2(x) = \frac{-\sqrt{x^2-\epsilon^2}}{x} + \log\,(x + \sqrt{x^2-\epsilon^2}) \tag{2.111}$$

Now define

$$u_1 = f_2(x_2) \quad u_2 = f_1(x_2)\cos(x_1/\epsilon) \quad u_3 = f_1(x_2)\sin(x_1/\epsilon) \tag{2.112}$$

Formula (2.112) maps the region of the Poincaré Plane where $-\epsilon\pi < x_1 \leq \epsilon\pi$ to a surface

$$\mathbf{HT}^2 = \{(u_1, u_2, u_3) \,:\, -\epsilon\pi < x_1 \leq \epsilon\pi\} \tag{2.113}$$

in $\mathbf{R}^3$. See Figure 2.10. While this representation is perhaps the most intuitive way to represent a space of constant negative curvature, the Poincaré Trumpet is the least satisfactory in other respects. If the representation is extended to the entire half plane then the mapping ceases to be 1-1. The mapping of the entire half plane onto the trumpet is, in fact, a covering map that wraps the half plane infinitely many times around the trumpet. Thus the correspondence is only locally correct.

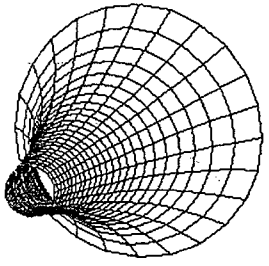

FIGURE 2.10. *Hyperbolic geometry representation in three dimensions: the Poincaré Trumpet.*

2.3 Notes

The reader looking for a good introduction to differential geometry may be somewhat overwhelmed by the variety of books that are formal introductions to the subject but make few concessions to the reader who is not trained in abstract mathematics. Such a reader would be well served by looking at the book by Guillemin and Pollack [78] and the book by Morgan [122]. For a general overall introduction to differential geometry, see Spivak [163].

2.4 Problems

1. The Hairy Ball Theorem says that for any continuous tangent vector field on a sphere $\mathbf{S}^2$ there is some point on the sphere at which the vector field vanishes. Is the analogous result true for the torus $\mathbf{S}^1 \times \mathbf{S}^1$?

2. We can construct a two-dimensional surface that is diffeomorphic to the torus $\mathbf{S}^1 \times \mathbf{S}^1$ as follows: Let $\mathbf{T}$ be the set of all points $(x_1, x_2, x_3) \in \mathbf{R}^3$ such that

$$(\sqrt{x_1^2 + x_2^2} - 2)^2 + x_3^2 = 1 \qquad (2.114)$$

This is the standard doughnut shape. Show that $\mathbf{T}$ is diffeomorphic to $\mathbf{S}^1 \times \mathbf{S}^1$.

3. An interesting surface called the *Moebius strip* can be embedded in the interior of the doughnut $\mathbf{T}$ from Problem 2 above. Let $\mathbf{M}^2$ be the set of all (x_1, x_2, x_3) such that

$$(r-2)^2 + x_3^2 \leq 1 \qquad x_3 \sin(\theta/2) = (r-2)\cos(\theta/2) \qquad (2.115)$$

where (r, θ) are the polar coordinates of (x_1, x_2). This is, in fact, a manifold with boundary. The manifold proper is constructed with strict inequality above. Show that the boundary of $\mathbf{M}^2$ is diffeomorphic to $\mathbf{S}^1$. (If we glue the boundaries of two separate copies of a Moebius strip together we also get a manifold without boundary. This manifold is called the Klein bottle $\mathbf{K}^2$.)

4. Following from Problem 3 above, we note that another manifold with boundary whose boundary is $\mathbf{S}^1$ is the disk $\mathbf{D}^2$. This is the set of all (x_1, x_2) such that $x_1^2 + x_2^2 \leq 1$. As the boundary of $\mathbf{D}^2$ is diffeomorphic to the boundary of $\mathbf{M}^2$ from Problem 3 above, in principle (given four dimensions to do it in), we could glue the boundaries together by fusing diffeomorphic points. If the two surfaces were cut out from paper we could try to tape their boundaries together. However, as we progressed with the taping in three dimensions we would simply run out of room to do it in. In four dimensions there is enough room. Show that the resulting manifold without boundary is diffeomorphic to the projective plane $\mathbf{RP}^2$.

5. Show that the geodesic paths on the sphere $\mathbf{S}^2$ are arcs of great circles found by slicing the sphere with a plane through the center of the sphere.

6. Consider the cylindrical surface in $\mathbf{R}^3$ defined as the set of all (x_1, x_2, x_3) such that $x_1^2 + x_2^2 = 1$ with $-\infty < x_3 < +\infty$. This surface is also represented as $\mathbf{S}^1 \times \mathbf{R}$. Show that the geodesics of $\mathbf{S}^1 \times \mathbf{R}$ are helixes of the form

$$x_1(t) = \cos(at) \qquad x_2(t) = \sin(at) \qquad x_3(t) = bt \qquad (2.116)$$

for arbitrary real values a and b.

7. Prove that formulas (2.29) and (2.30) make tangent vector summation and scalar multiplication well defined. That is, show that the equivalence classes of paths defined for $\dot{x}(t_0) + \dot{z}(t_0)$ and $\lambda\, \dot{x}(t_0)$ do not depend upon the coordinate system used. Furthermore, show that if $\dot{y}(t_0) = \dot{x}(t_0)$ and $\dot{w}(t_0) = \dot{z}(t_0)$ then as defined by (2.28) and (2.29) we have

$$\dot{x}(t_0) + \dot{z}(t_0) = \dot{y}(t_0) + \dot{w}(t_0) \qquad (2.117)$$

and

$$\lambda\, \dot{x}(t_0) = \lambda\, \dot{y}(t_0) \qquad (2.118)$$

3

Shape Spaces

3.1 The Sphere of Triangle Shapes

In this and the next two sections, we shall develop a geometric theory of shape due to Kendall [90].

Consider three landmarks

$$x_j = \Re(x_j) + i\,\Im(x_j), \quad j = 1, 2, 3 \tag{3.1}$$

in the complex plane such that at least two of the three landmarks are distinct. We shall now consider how to naturally represent the shape of the triangle with vertices at x_1, x_2, and x_3. It can easily be seen that the shape of the triangle can be represented as the complex number

$$z = \frac{2x_3 - (x_1 + x_2)}{x_2 - x_1} \tag{3.2}$$

provided that $x_2 \neq x_1$. The point z in the complex plane has the following interpretation. The triangle $x_1x_2x_3$ has the same shape as the triangle whose vertices lie at the three points -1, $+1$, and z. Thus to encode the shape of the triangle we need only move two points, say x_1 and x_2, to standard positions using a similarity transformation and record the position of the third point under this transformation. The real and imaginary coordinates of z, which determine the shape of the triangle, are called *Bookstein coordinates*, after F. Bookstein [19], who popularized them. See Figure 3.1.

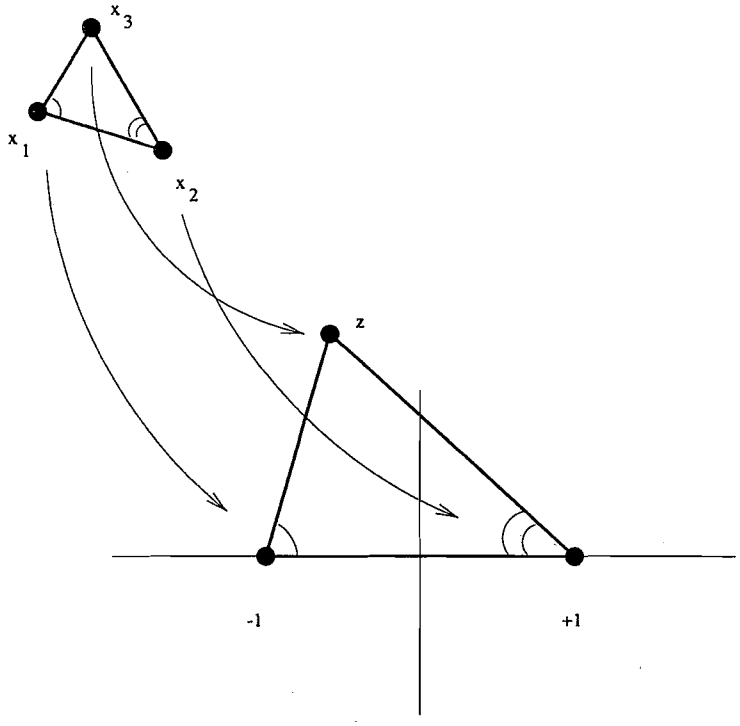

FIGURE 3.1. *Bookstein coordinates for three planar points. A triangle of landmarks* x_1, x_2, *and* x_3 *is translated, rotated, and rescaled so that the base points* x_1 *and* x_2 *are mapped to* -1 *and* $+1$, *respectively, in the complex plane. The third point* x_3 *is then mapped to a point* z *that encodes the shape information in the triangle. The real and imaginary parts of* z *are called the Bookstein coordinates.*

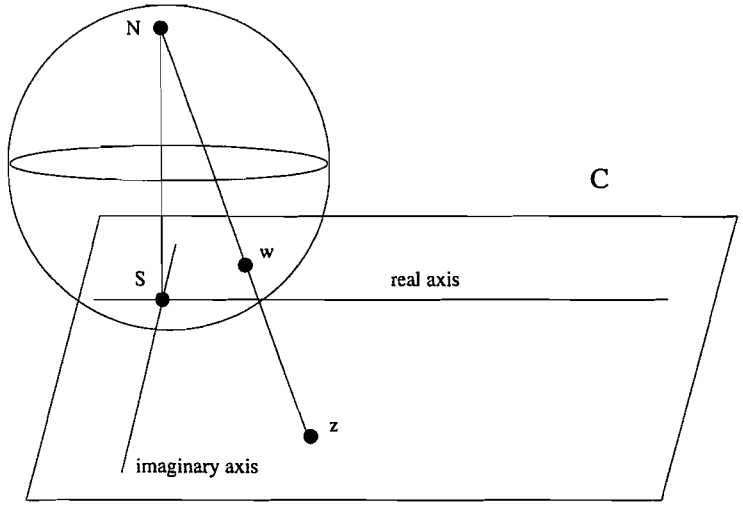

FIGURE 3.2. *The stereographic projection. On the complex plane* $\mathbf{C}$ *a sphere sits so that its south pole is at the origin. For any point* $z \in \mathbf{C}$ *we draw the line passing from the north pole of the sphere through the surface of the sphere at* w *and meeting the complex plane at* z. *The stereographic projection thereby puts the points of the closed plane and the sphere into 1-1 correspondence by mapping* z *to* w. *The stereographic projection maps the point at infinity in the complex plane to the north pole of the sphere.*

Such a coordinatization is not without its deficiencies, and it is these that we shall now consider. The most obvious difficulty in using the coordinates of z is that the representation breaks down if $x_1 = x_2$. Note that if $x_1 = x_2$ and x_3 is a distinct point, then the shape of the triangle is perfectly well defined even if the Bookstein coordinates are not. A related problem is that the use of x_1 and x_2 to standardize a side of the triangle is rather arbitrary. One of the other pairs of points could just as well be chosen.

Now suppose $x_1 = x_2$ and that x_3 is distinct from the other two points. Then the shape of this triangle is most naturally interpreted as $z = \infty$, the point at infinity in the complex plane. So the representation of shape is non-degenerate for all shapes provided this point is included. The complex plane, with the point at infinity added, is topologically equivalent to a sphere. Putting this another way, we could say that if a point is removed from the sphere $\mathbf{S}^2$ the resulting set is homeomorphic to the plane. The complex plane, together with its point at infinity, is called the *closed complex plane*. A standard mathematical tool that puts the closed complex plane into 1-1 correspondence with the points of a sphere is the *stereographic projection*. See Figure 3.2.

As the point z in Figure 3.2 follows the locus of a circle in the complex

plane, the point w on the sphere also follows the locus of a circle, although not necessarily a great circle. As lines in the closed complex plane can be regarded as circles of infinite radius passing through ∞, we find that as z follows the locus of a line in the plane, the corresponding point w follows the locus of a circle passing through the north pole. The class of circles generated by such loci for w is the full class of all circles on the sphere.

An interesting class of transformations emerges when we look at rotations of the sphere. Suppose we rotate the sphere so that w goes to some point w'. Correspondingly, the point z will move to some point z' elsewhere in the closed complex plane. The transformation $z \to z'$ is an example of a type of transformation called a *Moebius transformation* or a *linear fractional transformation*, whose general form is

$$z \to \frac{a\,z + b}{c\,z + d} \tag{3.3}$$

where a, b, c, and d are complex numbers such that $ad \neq bc$. Just as the class of rotations of the sphere maps circles to circles, so the Moebius transformations map circles in the plane, including straight lines as circles of infinite radius, to circles or straight lines.

Using the stereographic projection, we can represent the shape of any triangle $x_1x_2x_3$ as a point on a sphere in a topologically natural way. For any such triangle, we compute the point z given by formula (3.2) whose real and imaginary parts are the Bookstein coordinates of the shape. We then map z to a point w on the sphere by a stereographic projection. This takes us partway towards the goal stated in Chapter 1, namely the representation of shapes as points on manifolds.

However, we are not yet finished. Using the stereographic projection we can make a strong case for the argument that the space of triangle shapes should be homeomorphic to a sphere. However, topological considerations can tell us nothing about distances between shapes. In order to construct a satisfactory representation of triangle shapes as points on spheres we need to find a representation of triangle shapes such that the Procrustean metric of formula (1.21) in Chapter 1 is equivalent to geodesic distance on a sphere. At this stage we have no guarantee that this can be done, and even less of a guarantee that the stereographic projection will be instrumental in the construction.

If there is a representation on a sphere that works, we can easily see what the radius of the sphere must be. In Chapter 1, we found that the maximum Procrustean distance between any two triangle shapes was $\pi/2$. If this is interpreted as a geodesic distance on a sphere, then the radius of the sphere would be equal to $1/2$, and such shapes would be antipodal points on the sphere. In fact, we can find two such shapes. Two triangles whose Bookstein coordinates are $z = \pm\sqrt{3}i$ are equilateral triangles that are reflections of each other and have a Procrustean distance of $\pi/2$ from each other. Thus we seek a representation on a sphere in which these two

points are antipodal.

We can also look for clues to the role of the stereographic projection. Note that the Procrustean distance given in formula (1.21) is indifferent as to the labeling of the landmarks x_1, x_2, and x_3, provided all triangles are relabeled consistently. For example, using $n = 3$ in formula (1.21), we could interchange τ_{11} and τ_{12} and the distance $d(\sigma_1, \sigma_2)$ would not change provided that we similarly interchanged τ_{21} and τ_{22}. Another way of saying this is that the group of relabelings of landmarks is an isometry of the shape space Σ_2^n. So let us consider how the group of relabelings of $x_1x_2x_3$ induces transformations on Bookstein coordinates for triangle shapes. If we switch x_1 and x_2 then z, as defined by formula (3.2), is mapped to $-z$. This is an isometry of the complex plane. However, if we switch x_1 and x_3 then the point z is mapped by the transformation

$$z \to \frac{z + 3}{z - 1} \tag{3.4}$$

which is an example of a Moebius transformation of the complex plane. In a similar way to the above, if we switch the triangle points x_2 and x_3 then the induced transformation of shape becomes

$$z \to \frac{3 - z}{1 + z} \tag{3.5}$$

which is also a Moebius transformation. In fact, our first transformation $z \to -z$ is also a special case of a Moebius transformation. The group of relabelings of shapes is the set of six transformations of the complex plane that can be written as the arbitrary composition of these three Moebius transformations.

It is no coincidence that the Moebius transformations of the complex plane arise in relabeling triangle landmarks and also as the images under stereographic projection of rotations of the sphere. In both cases we are dealing with isometries – in the former case the isometries of Σ_2^3 and in the latter case isometries of the sphere. The type of transformation that we are seeking should be a stereographic projection from the closed complex plane onto a sphere of radius $1/2$ taking the two equilateral triangles into antipodal points!

Suppose the shape of triangle $x_1x_2x_3$ is displayed by Bookstein coordinates as a point z in the closed complex plane. Now define

$$w_1 = \frac{1 - |z|^2/3}{2(1 + |z|^2/3)}, \quad w_2 = \frac{\Re(z)/\sqrt{3}}{1 + |z|^2/3}, \quad w_3 = \frac{\Im(z)/\sqrt{3}}{1 + |z|^2/3} \tag{3.6}$$

Then

$$z \to (w_1,\ w_2,\ w_3) \tag{3.7}$$

is a stereographic projection of the triangle shape in Bookstein coordinates onto a sphere of radius $1/2$ centered at the origin in $\mathbf{R}^3$. The mapping

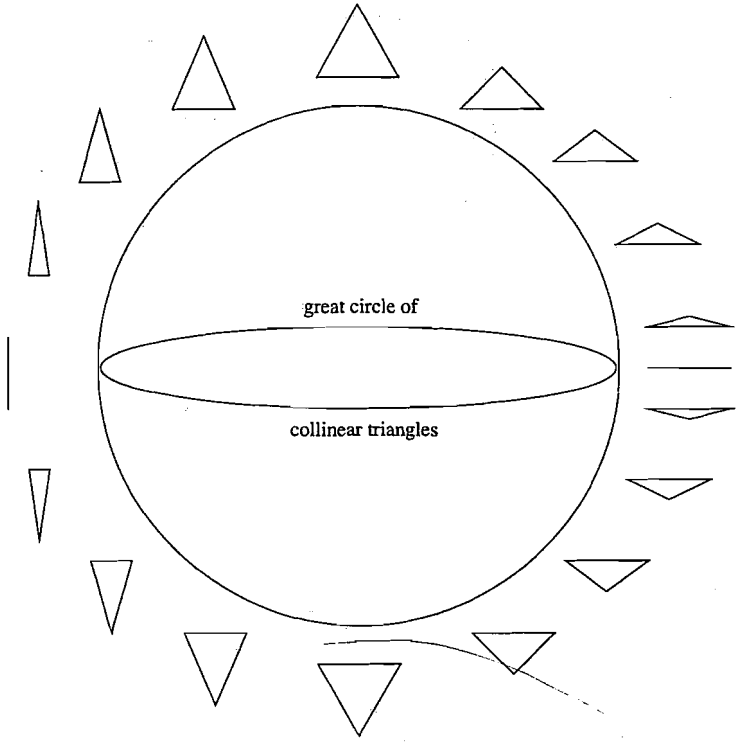

FIGURE 3.3. *Spherical geometry for three planar points. The shape of any triangle $x_1x_2x_3$ is encoded in Bookstein coordinates z as a point in the closed complex plane and then mapped by a particular stereographic projection to the sphere. There are two antipodal points on the sphere that correspond to the two equilateral triangles of landmarks in the plane. Passing through these two antipodal points are three great circles that correspond to the isosceles triangles of landmarks – each great circle characterized by the choice of vertex at which the isosceles angle occurs. A family of isosceles triangles around one such great circle is displayed around the outside of the sphere. Triangles of aligned landmarks (i.e., collinear triangles) are to be found on the great circle of the sphere that is equidistant from the two equilateral triangles and orthogonal to the great circles of isosceles triangles.*

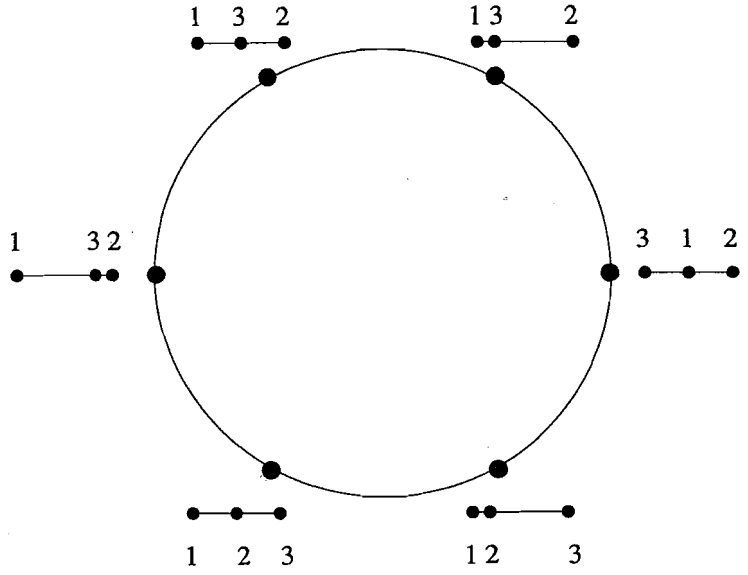

FIGURE 3.4. *A close-up look at the great circle of collinear triangles from Figure 3.3. Three points separated by arcs of* 120° *mark the collinear triangles where two landmarks are coincident. Halfway between these points are the collinear triangles of equally spaced points.*

from the pre-shape of $x_1x_2x_3$ to the shape of the triangle is a mapping

$$\mathbf{S}^3 \rightarrow \mathbf{S}^2(1/2) \tag{3.8}$$

It is helpful to study the sphere $\mathbf{S}^2(1/2)$ by finding the coordinates of interesting triangles on it. For example, there are two equilateral triangles represented by antipodal points on the sphere at $w_3 = \pm(1/2)$. That there are two equilateral triangles rather than one is a consequence of the fact that triangle shapes are not identified with their reflections. Halfway between these antipodal points are the shapes corresponding to $w_3 = 0$. These shapes lie on a great circle of $\mathbf{S}^2(1/2)$ that is the set of collinear triangles. In other words, these are the triangles that have a straight angle at one of the vertices. Included in this set are the three shapes corresponding to triangles where two of the points are coincident and the third point is distinct. These shape points are equally spaced at angles of $2\pi/3$ radians around the great circle $w_3 = 0$. See Figure 3.4.

The reader should make a careful comparison of Figures 1.3 and 3.4. In both figures, we see collinear triangles of landmarks displayed as points around a circle. However, there is an important difference. In Figure 1.3, the pre-shapes of triangles that are reflections of each other are distinct antipodal points of the circle. However, as we argued earlier in Example 2.2.15, the shapes of collinear triangles in the plane lie naturally on a real projective space and not a sphere. As it happens, the real projective space $\mathbf{RP}^1$ is isometric to the circle $\mathbf{S}^1(1/2)$. So each pair of antipodal points of Figure 1.3 is represented as a single point in Figure 3.4. For example, the earlier figure has six points that represent the shapes of triangles where one landmark is at the midpoint between the other two. However, Figure 3.4 has only three such points around the circle. The pre-shape space of Figure 1.3 is the unit circle $\mathbf{S}^1$, whereas the space of collinear shapes in Figure 3.4 is the real projective space $\mathbf{RP}^1 \cong \mathbf{S}^1(1/2)$.

To obtain the great circle distance on $\mathbf{S}^2(1/2)$ between any two shapes, we use the inner product between vectors on $\mathbf{S}^2(1/2)$. If $u = (u_1, u_2, u_3)$ and $v = (v_1, v_2, v_3)$ are two points on $\mathbf{S}^2(1/2)$ then the geodesic distance from u to v is given by formula (2.69) using $r = 1/2$. We obtain

$$d(u, v) = \frac{1}{2} \cos^{-1}(4 < u, v >) \tag{3.9}$$

We leave the reader to check that the great circle distance defined by this formula is equivalent to that of Chapter 1. See Problem 2.

Before we turn to the study of Σ_2^n it is worth considering some of the geometry of the sphere and its relationship to Bookstein coordinates. The geodesic paths are the shortest paths between points. As we noted, these are the arcs of great circles on $\mathbf{S}^2(1/2)$. To find the corresponding paths in Bookstein coordinates, we need to construct the images of the great circles under stereographic projection. Any circle on the sphere $\mathbf{S}^2(1/2)$ is mapped by the inverse of the stereographic projection defined by (3.6) to a circle or a straight line in the plane. Among these, the images of the great circles are a subset. The x-axis of collinear shapes is an example of a geodesic in Bookstein coordinates. To find the others, note that any two great circles of $\mathbf{S}^2(1/2)$ will intersect in antipodal points. In fact, we can characterize a great circle of the sphere as a circle meeting the equator of collinear shapes in antipodal points. Now in Bookstein coordinates, two points $z_1, z_2 \in \mathbf{C}$ are images of antipodal points on the sphere if $z_2 = -3/z_1^*$. This can be checked by plugging $z = -3/z_1^*$ and $z = z_1$ into the coordinates of the stereographic projection in formula (3.6). After some rearranging, we see that the resulting stereographic coordinates become the negatives of each other. Therefore, any circle in the plane of Bookstein coordinates that passes through points of the form a and $-3/a$ on the real axis will be the stereographic image of a great circle of $\mathbf{S}^2(1/2)$.

3.2 Complex Projective Spaces of Shapes

In this section we shall study the spaces Σ_2^n where $n \geq 3$. As we shall see, the sphere of triangle shapes described in the previous section is a special case of a complex projective space having two real dimensions.

We will continue to identify landmarks x_j in the plane with elements of the complex plane $\mathbf{C}$. Suppose $(x_1, x_2, \ldots, x_n)$ are n such landmarks, at most $n-1$ of that are coincident. To discover the information in this configuration of landmarks that is invariant under $\mathbf{Sim}(2)$, we first remove the effect of translations by centering the points about their centroid $\bar{x}$ yielding

$$(x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}) \tag{3.10}$$

This vector lies in a subspace of $\mathbf{C}^n$ having $n-1$ complex dimensions or $2n-2$ real dimensions. The effect of multiplication of these variables by a complex nonzero quantity λ

$$[\lambda(x_1 - \bar{x}), \lambda(x_2 - \bar{x}), \ldots, \lambda(x_n - \bar{x})] \tag{3.11}$$

is to scale the centered points by $|\lambda|$ and rotate them by $\arg(\lambda)$. To remove the effect of complex multiplication, we identify all such multiples and declare them to lie in the same equivalence class.

So, the shape space Σ_2^n can be identified with the set of *complex lines* through the origin in the subspace

$$\mathbf{F}^{n-1} = \{(x_1, \ldots, x_n) \in \mathbf{C}^n : \sum_{j=1}^{n} x_j = 0\} \tag{3.12}$$

which has $n-1$ complex dimensions. This looks very similar to complex projective space $\mathbf{CP}^{n-2}$, as given in Definition 2.2.10. The difference is

that we are considering complex lines through the subspace $\mathbf{F}^{n-1}$ rather than $\mathbf{C}^n$. However, this difference turns out to be superficial, because we can construct a linear isometry from $\mathbf{F}^{n-1}$ to $\mathbf{C}^{n-1}$ that maps complex lines through the origin in the subspace $\mathbf{F}^{n-1}$ to complex lines through the origin in $\mathbf{C}^{n-1}$. To construct this linear isometry, we define

$$y_j = [jx_{j+1} - (x_1 + x_2 + ... + x_j)] / \sqrt{j^2 + j} \tag{3.13}$$

for $1 \leq j \leq n-1$. The mapping

$$(x_1, x_2, ..., x_n) \rightarrow (y_1, y_2, ..., y_{n-1}) \tag{3.14}$$

is a linear isometry from $\mathbf{F}^{n-1}$ to $\mathbf{C}^{n-1}$ that preserves the complex lines of (3.11) above.

Under the identification established by (3.14), we can see that the definition of the Procrustean metric in Section 1.3 is completely parallel to the definition of the Fubini-Study metric in Section 2.2. In particular, formulas (2.92) and (1.18) yield equivalent metrics under the identification of (3.14). Thus we have proved the following result:

Proposition 3.2.1. *The shape space Σ_2^n endowed with the Procrustean metric is isometric to the complex projective space $\mathbf{CP}^{n-2}$.*

So, the shape spaces Σ_2^n are Riemannian manifolds such that geodesic distance between points in the shape space is equivalent to the Procrustean metric defined in Chapter 1. A technical note on this point is that the Gaussian curvature of Σ_2^n is a constant throughout the manifold and equal to 4. By contrast, the sphere $\mathbf{S}^{2n-3}$ of pre-shapes has a constant positive curvature equal to 1. (The reader who is not familiar with Gaussian curvature on manifolds should rest assured that this notion will not play a large role in our exposition of shape geometry.) A special case of this was seen previously for the sphere of shapes, which is required to have radius 1/2. In general, the mapping

$$\mathbf{S}_\star^{2n-3} \rightarrow \Sigma_2^n \cong \mathbf{CP}^{n-2} \tag{3.15}$$

of each pre-shape into its shape equivalence class becomes what is known as a Riemannian submersion, a local projection that will be described in greater detail in the next section where we shall consider the general spaces Σ_p^n.

A special case of our construction is quite famous in differential geometry. We have seen that Σ_2^3 is isometric to the sphere $\mathbf{S}^2(1/2)$. Thus the mapping from each pre-shape to its corresponding shape is equivalent to a mapping

$$\mathbf{S}_\star^3 \rightarrow \mathbf{S}^2(1/2) \tag{3.16}$$

The change of radius is a secondary consideration here. A continuous function from $\mathbf{S}^3$ to $\mathbf{S}^2$ of this kind is an example of what is known as a *Hopf fibration* between the spheres. In general, it is impossible to find continuous mappings from a sphere of one dimension to a sphere of a lower dimension that are locally projections of this kind. However, there are special dimensions for which it is possible. From three dimensions to two dimensions is one such case. Such limitations in dimensions already give us a clue that the Procrustean approach to the shape of a general number of points in general dimensions will not be as smooth a theory as for the shapes of points in dimension 2.

3.3 Landmarks in Three and Higher Dimensions

3.3.1 Introduction

So far, we have only considered the shapes of landmarks in two dimensions. However, the shapes of solid objects are of common interest, and are most naturally represented by landmarks in three dimensions. Landmarks in four and higher dimensions are of interest in multivariate statistics, where the shapes of multivariate data sets provide information about normality, linearity, and correlation between variables.

Let $x_1, x_2, ..., x_n$ be $n \geq 3$ landmarks in $\mathbf{R}^p$, where $p \geq 3$. We shall suppose that at least two of these landmarks are distinct, so that $\sum ||x_j - \bar{x}||^2 > 0$. The standardization of the location and scale of these n landmarks can proceed in a manner similar to the two-dimensional case. The pre-shape τ can be constructed by centering the landmarks about their centroid $\bar{x}$ and by rescaling the centered configuration of landmarks so that $\sum ||x_j - \bar{x}||^2 = 1$. Thus the pre-shape τ of $x_1, ..., x_n$ can be seen to be an element of the sphere

$$\mathbf{S}_\star^{np-p-1} = \{(y_1, ..., y_n) \in \mathbf{R}^{np} : \sum y_j = 0, \sum ||y_j||^2 = 1\} \tag{3.17}$$

The space Σ_p^n of shapes of $x_1, ..., x_n$ can now be formally identified with the collection of equivalence classes in $\mathbf{S}_\star^{np-p-1}$ of all pre-shapes sharing a common shape. For any pre-shape $\tau \in \mathbf{S}_\star^{np-p-1}$, let $\mathcal{O}(\tau)$ be be the set of all pre-shapes τ' that have the same shape as τ.

For example, in dimension $p = 3$ the special orthogonal group $\mathbf{SO}(3)$ is simply the group of rotations about the origin in three-dimensional space. Let $h \in \mathbf{SO}(3)$. Suppose that the landmarks $x_1, ..., x_n$ have pre-shape τ. For $j = 1, ..., n$, let $x_j' = h(x_j)$ be the jth landmark rotated by h. So the landmarks $x_1', ..., x_n'$ are a rotated version of $x_1, ..., x_n$. If τ' is the pre-shape of $x_1', ..., x_n'$, then τ' will be an element of $\mathcal{O}(\tau)$. The converse will also follow. If $x_1', ..., x_n'$ have a pre-shape $\tau' \in \mathcal{O}(\tau)$, then there will exist a rotation $h \in \mathbf{SO}(3)$ such that $x_j' = h(x_j)$ for all j.

Such equivalence classes can be defined similarly in higher dimensions p using the special orthogonal group $\mathbf{SO}(p)$. Then we can define

$$\Sigma_p^n = \{\mathcal{O}(\tau) : \tau \in \mathbf{S}_\star^{np-p-1}\} \tag{3.18}$$

We can also define the function

$$s_{pn} : \mathbf{S}_\star^{np-p-1} \to \Sigma_p^n \tag{3.19}$$

taking each pre-shape τ to its corresponding equivalence class, or shape, $\mathcal{O}(\tau)$.

Now any set of n landmarks in $\mathbf{R}^p$ can be identified with an element of $\mathbf{R}^{np}$. Since the elements of the group $\mathbf{SO}(p)$ transform the landmarks individually, we can regard $\mathbf{SO}(p)$ as a group of transformations on $\mathbf{R}^{np}$. Each $h \in \mathbf{SO}(p)$ maps a point $(x_1, ..., x_n)$ in $\mathbf{R}^{np} = (\mathbf{R}^p)^n$ by the rule

$$(x_1, x_2, ..., x_n) \to [h(x_1), h(x_2), ..., h(x_n)] \tag{3.20}$$

Interpreted in this way, the group $\mathbf{SO}(p)$ becomes a subgroup of the group of special orthogonal transformations on $\mathbf{R}^{np}$, namely $\mathbf{SO}(np)$.

The next thing to note is that $\mathbf{S}_\star^{np-p-1}$ is a subset of $\mathbf{R}^{np}$, and that transforming according to (3.20), the transformations $h \in \mathbf{SO}(p)$ map $\mathbf{S}_\star^{np-p-1}$ onto itself. So $\mathbf{SO}(p)$ is a class of isometries of the sphere $\mathbf{S}_\star^{np-p-1}$. Moreover, we can write

$$\mathcal{O}(\tau) = \{h(\tau) \ : \ h \in \mathbf{SO}(p)\} \tag{3.21}$$

We can introduce a Procrustean metric between shapes in Σ_p^n in a manner similar to the two-dimensional case. So for any shapes $\sigma_1 = \mathcal{O}(\tau_1)$ and $\sigma_2 = \mathcal{O}(\tau_2)$ in Σ_p^n we can set the *Procrustean metric* $d(\sigma_1, \sigma_2)$ to be

$$\inf\{\cos^{-1}(< \tau_1, \tau_2 >) \ : \ \sigma_j = \mathcal{O}(\tau_j) \ \text{ for } j = 1, 2\} \tag{3.22}$$

This is equivalent to the definition given in formula (1.18) with the appropriate change in dimension. However, as we shall see, appearances are deceiving here. The extension of the geometry of Σ_2^n to higher-dimensional settings is not as routine as this formula would suggest. One algebraic advantage is lost in the generalization: the algebra of the complex plane is not available for representing the Procrustean metric when landmarks are chosen from three or higher dimensions.

The metric of (3.22) does not in itself provide much immediate insight into the topological and differential structure of Σ_p^n. We can construct the topology directly on Σ_p^n without direct reference to the metric d. A subset U of Σ_p^n will be open if and only if $s_{pn}^{-1}(U)$ is an open subset of $\mathbf{S}_\star^{np-p-1}$. With this topology, the function s_{pn} becomes continuous. It follows immediately from this that all the shape spaces Σ_p^n are compact, because they are continuous images of the compact spheres $\mathbf{S}_\star^{np-p-1}$.

Now let us consider the space Σ_n^n for $n = 3, 4, ...$ The Euclidean space $\mathbf{R}^{n-1}$ can be canonically embedded in $\mathbf{R}^n$ so as to be an $(n-1)$-dimensional subspace of $\mathbf{R}^n$. This embedding induces a mapping from Σ_{n-1}^n to Σ_n^n that takes the shape of a set of n points in $\mathbf{R}^{n-1} \subset \mathbf{R}^n$ into the shape of the same set of points considered as lying in $\mathbf{R}^n$.

In the case where $n = 3$ we can see what this does. See Figure 3.5. The shapes of point configurations in $\mathbf{R}^2$ that are relections of each other through some line are not generally of the same shape. However, in $\mathbf{R}^3$ a plane can be reflected about some line by a rotation. Thus configurations that are mirror images in $\mathbf{R}^2$ have the same shape when embedded in $\mathbf{R}^3$. The shape space Σ_2^3 is the sphere $\mathbf{S}^2(1/2)$, and the associated mapping into Σ_3^3 identifies every triangle shape with its mirror image in $\mathbf{R}^2$. In the coordinate notation of formula (3.6) this identifies points of the form (w_1, w_2, w_3) with $(w_1, w_2, -w_3)$. Thus Σ_3^3 is topologically a hemisphere with the collinear shapes forming its boundary.

This example points out a major obstacle to the study of the shape spaces in general dimensions. On the boundary, the hemisphere is not locally homeomorphic to $\mathbf{R}^2$ as it is in its interior. So a hemisphere is not a topological manifold at all, but must be classified as a manifold with boundary. Generally, the spaces Σ_p^n will have boundaries whenever $p \geq n$. Even when Σ_p^n is a topological manifold, it need not have a natural definition as a differential manifold. Singularities in the smoothness can arise much as one can introduce a crease into a surface.

In a private communication to David Kendall, A. J. Casson proved that the shape spaces Σ_n^{n+1} are all topologically spheres for $n \geq 2$. That this is the case for $n = 2$ we have already seen. However, that it should be true for the topology of Σ_n^{n+1} for $n \geq 3$ is interesting because it is known that these spaces are not diffeomorphic to the usual spheres of equivalent dimension. The presence of singularities in the differential structure is enough to ensure this. In honor of Casson's discovery, D.G. Kendall proposed that the shape spaces Σ_n^{n+1} be called *Casson spheres.* Unfortunately, Casson's proof is not available in the literature although another proof has been published. See Le [104]. Le's proof makes use of the Riemannian geometry off the singularity sets of the Casson spheres to prove Casson's result. See also Carne [38] for an analysis of the geometry of these shape spaces.

Let us now consider the differential geometry of the general shape spaces Σ_p^n. In order to do this we shall need to define the concept of *submersion* between differential manifolds. We have the following definition:

Definition 3.3.1. *Let* $h : \mathbf{M}^p \to \mathbf{N}^q$ *be a differentiable mapping onto the manifold* $\mathbf{N}^q$, *where* $q \leq p$. *We say that* h *is a* submersion *at a point* $x \in \mathbf{M}^p$ *when the linear mapping*

$$(\mathcal{D}h)_x : T_x(\mathbf{M}^p) \to T_{h(x)}(\mathbf{N}^q) \tag{3.23}$$

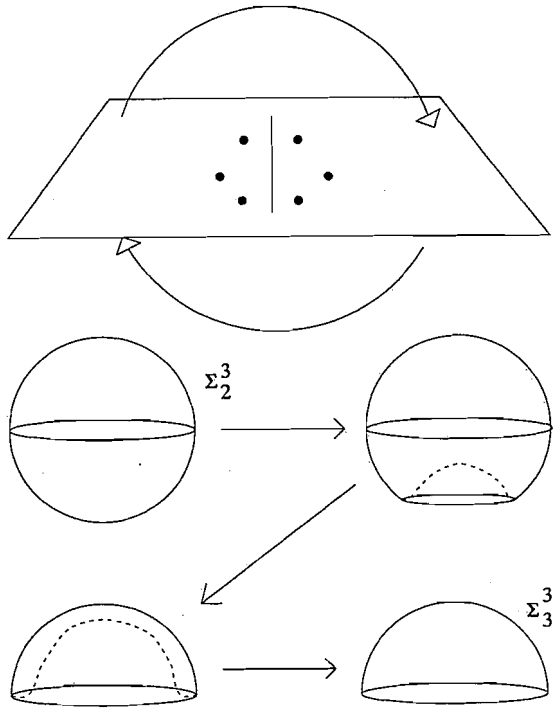

FIGURE 3.5. *The effect of embedding a configuration of three planar landmarks into three dimensions. Configurations of landmarks in* $\mathbf{R}^2$ *which are reflections of each other have different shapes, because transformations that reflect the plane are not elements of the group* $\mathbf{Sim}(2)$. *By contrast, coplanar configurations of landmarks in* $\mathbf{R}^3$ *that are reflections of each other do have the same shape. This is because the group* $\mathbf{Sim}(3)$ *includes* $180°$ *rotations of planes in* $\mathbf{R}^3$ *about an axis in the plane. The shape of three landmarks in* $\mathbf{R}^2$ *lies in* Σ_2^3, *while the shape of three landmarks in* $\mathbf{R}^3$ *lies in* Σ_3^3. *The identification of shapes that are reflections of each other in* Σ_2^3 *can be regarded as the identification of points on opposite hemispheres of the sphere of triangle shapes. Topologically, the effect of identifying points in opposite hemispheres is to fold one hemisphere into the other and to glue the two surfaces together. Thus the space of triangle shapes in three dimensions is topologically a hemisphere.*

is of full rank q *or equivalently, when* $(\mathcal{D}h)_x$ *is onto. The mapping* h *is said to be a submersion provided that it is a submersion at all points* $x \in \mathbf{M}^p$.

We have already encountered a number of examples of submersions. For example, the class of submersions includes linear projections

$$\mathbf{R}^q \times \mathbf{R}^{p-q} \rightarrow \mathbf{R}^q \tag{3.24}$$

mapping

$$(x_1, ..., x_q, ..., x_p) \rightarrow (x_1, ..., x_q) \tag{3.25}$$

A submersion between manifolds can be regarded as a differentiable mapping that is locally equivalent to a projection. From our point of view, perhaps the most important examples of submersions that we have encountered are the mappings

$$s_{2n} : \mathbf{S}_*^{2n-3} \rightarrow \Sigma_2^n \tag{3.26}$$

taking the pre-shapes of planar configurations of landmarks to their shapes. In particular, the Hopf fibration from $\mathbf{S}^3$ to $\mathbf{S}^2$ is a submersion.

The problem at hand is to make Σ_p^n into a differential manifold in such a way that its atlas is compatible with its topology and so that the mapping $s_{pn} : \mathbf{S}_*^{np-p-1} \rightarrow \Sigma_p^n$ becomes a submersion. The detailed conditions under which this is possible are given by Dieudonné [51, Section XVI.10] and will not be explained in detail here. We shall simply note that the submersion can be constructed for some pre-shapes (and their corresponding shapes) but not for others. The result is that there exists a *singularity set* within each shape space Σ_p^n such that outside this set a local smooth structure can be imposed at all the points, making s_{pn} a submersion. The particular locus of this singularity set within Σ_p^n is determined by the failure of the group $\mathbf{SO}(p)$ to act *freely* on the sphere $\mathbf{S}_*^{np-p-1}$ as defined below.

Definition 3.3.2. *Let* $\mathbf{H}$ *be a group of transformations* $h : \mathbf{M}^p \rightarrow \mathbf{M}^p$ *on a manifold* $\mathbf{M}^p$. *We say that* $\mathbf{H}$ *acts* freely *on* $\mathbf{M}^p$ *if the only transformation* $h \in \mathbf{H}$ *for which* $h(x) = x$ *for some* $x \in \mathbf{M}^p$ *is the identity transformation.*

In other words, if $\mathbf{H}$ is free, then every transformation h that moves some point of the manifold will move all points of the manifold. For example, the group $\mathbf{SO}(2)$ acts freely on $\mathbf{S}^1$, whereas the group $\mathbf{SO}(3)$ does not act freely on $\mathbf{S}^2$.

As Le and Kendall [105] have noted, the singularities in Σ_p^n arise because they are the images under s_{pn} of pre-shapes at which the action of the group $\mathbf{SO}(p)$ on $\mathbf{S}_*^{np-p-1}$ is not free. For example, let $p \geq 3$ and consider

a set of $n \geq p+1$ points $x_1, x_2, ..., x_n$ lying in $\mathbf{R}^p$. We center the location of the points by subtracting the centroid $\bar{x}$. Now suppose that there exists some $(p-2)$-dimensional subspace in which the centered points

$$x_1 - \bar{x},\ x_2 - \bar{x},\ ...,\ x_n - \bar{x} \tag{3.27}$$

all lie. Then there exists a special orthogonal transformation of $\mathbf{R}^p$ that is not the identity transformation and that leaves this $(p-2)$-dimensional subspace fixed.

To illustrate this, let us consider what happens in $p = 3$ dimensions. Put rather simply, we can say that it is possible to rotate a configuration of landmarks without changing the orientation (i.e, leaving the pre-shape fixed) provided the landmarks all lie along the axis of rotation. This is in contrast to dimension two, where a configuration cannot be left invariant under a rotation unless all the landmarks are coincident. Suppose that $x_1, ..., x_n$ are collinear landmarks in $\mathbf{R}^3$, and that $n \geq 3$. Then the centered landmarks $x_1 - \bar{x}, ..., x_n - \bar{x}$ will all lie along a line passing through the origin. Now suppose a rotation $h \in \mathbf{SO}(3)$ is chosen that has this line as its axis of rotation. Then h will leave the vector $(x_1 - \bar{x}, ..., x_n - \bar{x})$ fixed under the transformation

$$(x_1 - \bar{x},\ ...,\ x_n - \bar{x}) \ \rightarrow\ \left[h(x_1) - \overline{h(x)},\ ...,\ h(x_n) - \overline{h(x)}\right] \tag{3.28}$$

In addition, the rotation h will induce a transformation on the sphere of pre-shapes $\mathbf{S}_*^{3n-4}$, mapping the pre-shape of $x_1, ..., x_n$ to the pre-shape of $h(x_1), ..., h(x_n)$. If the transformation in (3.28) leaves centered landmarks $x_j - \bar{x}$ fixed, the same will be true of the pre-shapes. Thus $\mathbf{SO}(3)$ does not act freely on $\mathbf{S}_*^{3n-4}$. See Figure 3.6 for an illustration of this.

In general dimensions the group $\mathbf{SO}(p)$ will fail to act freely on $\mathbf{S}_*^{np-p-1}$ when $n, p \geq 3$. The singularity set in Σ_p^n will be the set of those shapes of landmarks $x_1, ..., x_n$ which lie, when recentered as in formula (3.27) above, in a $(p-2)$-dimensional subspace. In the five-dimensional Casson sphere Σ_3^4, for example, this subspace is one-dimensional. Therefore, the singularity set is the subset of collinear shapes.

3.3.2 *Riemannian Submersions*

Let us now turn to the problem of defining a metric tensor g on the open subset of Σ_p^n that is the complement of the singularity set. In order to describe this, we have to define a type of submersion, called the *Riemannian submersion*, which is specific to the theory of Riemannian manifolds. Suppose $\mathbf{M}^p$ and $\mathbf{N}^q$, for $p > q$, are Riemannian manifolds with metric tensors $g_{\mathbf{M}}$ and $g_{\mathbf{N}}$ respectively. These metrics define inner products on the tangent spaces $T(\mathbf{M}^p)$ and $T(\mathbf{N}^q)$ respectively. Now let $h : \mathbf{M}^p \rightarrow \mathbf{N}^q$ be a submersion. Then for each $x \in \mathbf{M}^p$ and $y \in \mathbf{N}^q$ such

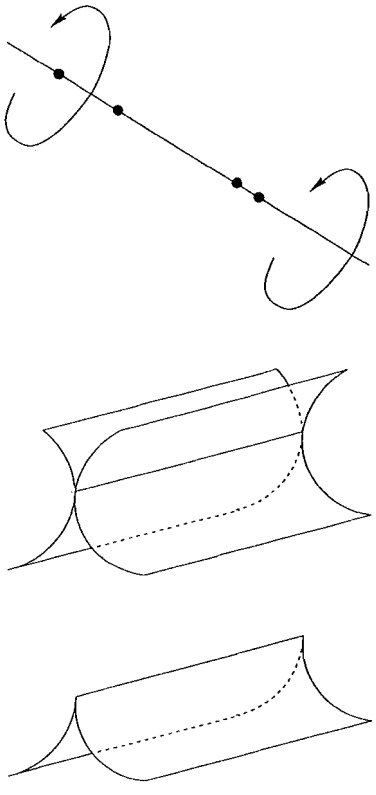

FIGURE 3.6. *Singularity sets in shape spaces. If a set of landmarks in* $\mathbf{R}^3$ *is collinear, as in the top diagram, then rotations about the line through the landmarks will leave the landmarks fixed. This is an example of the failure of the group of rotations to act freely. Singularities in the shape space* Σ_3^n *occur at points corresponding to such configurations of landmarks. Singularities in high-dimensional manifolds are difficult to understand, although singularities do appear in low enough dimensions to help us visualize them. Two types of singularities are illustrated in the middle and bottom diagrams. In the middle diagram, we see a topological singularity in a space. The singularity is the set of points where two surfaces intersect. At these points, the space fails to be locally homeomorphic to* $\mathbf{R}^2$*, and is not a topological manifold. However, if this intersection set is cut out, the remaining set does become a topological manifold. In the bottom diagram, we see another type of singularity set in a surface. In this case, the singularity is in the smoothness, or differential structure, of the manifold. Unlike the middle diagram, there is no topological singularity. The singularity set in the shape space* Σ_3^4 *is of this nature. This shape space is topologically a sphere, but contains a higher-dimensional analog of the type of singularity displayed in the bottom diagram. We cannot do differential geometry (i.e., construct tangent vectors or set up a metric tensor) at the singularity set, but we can do it elsewhere.*

that $y = h(x)$ the derivative

$$(\mathcal{D}h)_x \ : \ T_x(\mathbf{M}^p) \rightarrow T_y(\mathbf{N}^q) \tag{3.29}$$

is a linear transformation of full rank. We shall say that h is a *Riemannian submersion* if $(\mathcal{D}h)_x$ is equivalent to an orthogonal projection for all $x \in \mathbf{M}^p$. The following more precise definition can be given:

Definition 3.3.3. *Let $h : \mathbf{M}^p \rightarrow \mathbf{N}^q$ be a submersion as described above, and let $x \in \mathbf{M}^p$. We define the* vertical subspace *$V_x(\mathbf{M}^p)$ to be that subset of $T_x(\mathbf{M}^p)$ defined by*

$$V_x(\mathbf{M}^p) = \{\dot{x} \in T_x(\mathbf{M}^p) : (\mathcal{D}h)_x(\dot{x}) = 0\} \tag{3.30}$$

This is the kernel of the mapping $(\mathcal{D}h)_x$.

The vectors of the vertical subspace are called the *vertical tangent vectors* at $x \in \mathbf{M}^p$. Orthogonal to the vertical subspace is the *horizontal subspace*, which we now define.

Definition 3.3.4. *The* horizontal subspace *$V_x^{\perp}(\mathbf{M}^p)$ is defined to be the set*

$$V_x^{\perp}(\mathbf{M}^p) = \{\dot{x} \in T_x(\mathbf{M}^p) :< \dot{x}, \dot{y} >= 0 \text{ for all } \dot{y} \in V_x(\mathbf{M}^p)\} \tag{3.31}$$

where the inner product is calculated in $T_x(\mathbf{M}^p)$ using the metric tensor $g_{\mathbf{M}}$.

Similarly, the vectors of the horizontal subspace are called the *horizontal tangent vectors* at $x \in \mathbf{M}^p$. It is easy to see that any tangent vector at x can be uniquely written as a vector sum of a horizontal and a tangent vector that are orthogonal to each other with respect to the metric tensor g_M. See Figure 3.7 for an illustration of the horizontal and tangent vectors at a point $x \in \mathbf{M}^p$.

Using concepts of horizontal and vertical tangent vectors, it is now possible for us to define the concept of a *Riemannian submersion.*

Definition 3.3.5. *A submersion $h : \mathbf{M}^p \rightarrow \mathbf{N}^q$ is said to be a* Riemannian submersion *at x if*

$$(\mathcal{D}h)_x \ : \ V_x^{\perp}(\mathbf{M}^p) \rightarrow T_y(\mathbf{N}^q) \tag{3.32}$$

is a linear isometry when these spaces have metric tensors $g_{\mathbf{M}}$ and $g_{\mathbf{N}}$ respectively. We shall say that h is a Riemannian submersion if h is a Riemannian submersion at all points $x \in \mathbf{M}^p$.

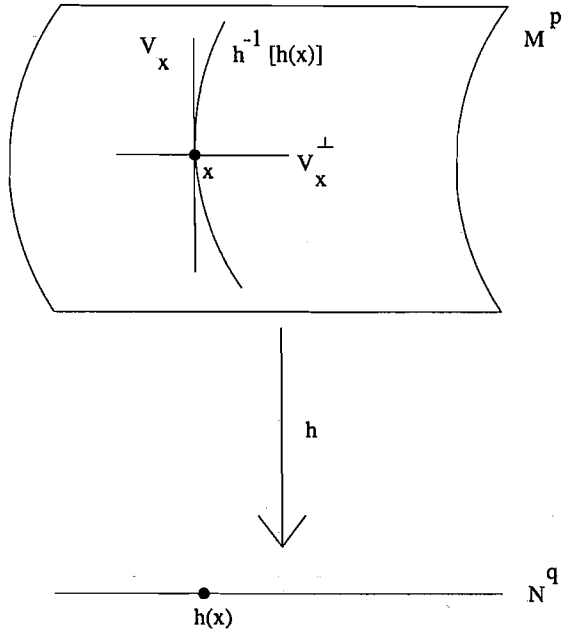

FIGURE 3.7. *Decomposition of $T_x(\mathbf{M}^p)$ into vertical and horizontal components.*

The basic principle for constructing a metric tensor on Σ_p^n off the singularity set is to define it so that s_{pn} is a Riemannian submersion at all pre-shapes $\tau \in \mathbf{S}_*^{np-p-1}$ at which s_{pn} is a submersion (i.e., pre-shapes outside the singularity set). Such a metric tensor is uniquely defined. Thus the determination of the metric tensor on Σ_p^n is equivalent to the evaluation of the metric tensor on $\mathbf{S}_*^{np-p-1}$ restricted to the horizontal tangent spaces.

D.G. Kendall and H. Le have carried out this program to evaluate the metric tensor on the shape spaces. See [104] and [105]. With this geometry, the geodesics of Σ_p^n become the images under the mapping s_{pn} of the *horizontal geodesics* of $\mathbf{S}_*^{np-p-1}$. These are the geodesics $x(t)$ of the sphere for which

$$\dot{x}(t) \ \in \ V^{\perp}(\mathbf{S}_*^{np-p-1}) \tag{3.33}$$

at all points $x(t)$ along the geodesic path.

3.4 Principal Coordinate Analysis

The Procrustean metric and the shape spaces of the previous sections provide very general tools for the representation of the shapes of landmark

configurations as points in manifolds. However, mathematically elegant as these representations are, they represent an impediment to the graphical representation for exploratory data analysis, which much be accomplished in a small number of dimensions. For example, if three landmarks are selected from each of fifty images, then the resulting landmark shapes can be displayed as a configuration of fifty points on an appropriate projection of the sphere Σ_2^3. However, more detailed descriptions of the shapes will require more landmarks from each image, and a correspondingly higher-dimensional manifold in which to portray the fifty points.

The tools for shape representation that we have been considering can be useful for the exploratory analysis of shapes when they can be coupled with dimension reduction methods that are designed to approximate the high-dimensional configuration of the points by low- (usually one or two) dimensional configurations whose interpoint distances most appropriately approximate those of the high-dimensional configuration. Such methods are called *multidimensional scaling*. There is considerable reason for optimism about the use of multidimensional scaling, because from formula (1.21) we see that the geodesic distance between two shapes can be quite simple to compute even when the complex projective spaces in which the shapes live are hard to visualize.

Suppose $x_1, x_2, ..., x_n$ are elements of some Riemannian manifold $\mathbf{M}^p$. We shall let $d_{jk} = d(x_j, x_k)$ be the geodesic distance from x_j to x_k. The $n \times n$ distance matrix (d_{jk}) is a symmetric matrix of nonnegative values. (The particular application we have in mind is that where $\mathbf{M}^p$ is a shape manifold and d is possibly the Procrustean metric given by (1.21).) The task of multidimensional scaling is to find a set of points $\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n \in \mathbf{R}^q$ (where usually $q = 1$ or 2) such that if $\tilde{d}_{jk} = d(\tilde{x}_j, \tilde{x}_k)$ then the matrix $(\tilde{d}_{jk})$ approximates (d_{jk}) in some predetermined sense. The various methods used to approximate (d_{jk}) by $(\tilde{d}_{jk})$ can be used to categorize the types of multidimensional scaling. Broadly speaking, the methods divide into two groups called *metric scaling* and *nonmetric scaling* respectively. In metric scaling, the task is to make the distance matrix $(\tilde{d}_{jk})$ match (d_{jk}) as closely as possible. In nonmetric scaling this requirement is relaxed. A typical criterion is that the distances $\tilde{d}_{jk}$ should be ordered as closely as possible to the ordering of the distances d_{jk}.

In this section we shall describe a computationally straightforward technique for metric scaling called *principal coordinate analysis* due to Gower [74]. This should not be confused with the better-known term *principal component analysis*, although the two techniques are related and rely on the common principle of an appropriate eigenvector decomposition of a positive definite matrix.

Let us begin with the following problem: Suppose that $x_1, x_2, ..., x_n$ are n points in some p-dimensional space that we can take to be Euclidean. The coordinates, or positions, of the points themselves are unknown. However, the distances $d_{jk} = d(x_j, x_k)$ between the points are given to us. As the original points are unknown, how can we construct a set of points $\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n$, which are not necessarily in p dimensions, with interpoint distances $\tilde{d}_{jk} = d(\tilde{x}_j, \tilde{x}_k)$, such that $\tilde{d}_{jk} = d_{jk}$ for all $1 \le j, k \le n$?

Let us start with any matrix (d_{jk}) of interpoint distances. The task of constructing the set of points $\tilde{x}_1, ..., \tilde{x}_n$ proceeds as follows:

Step 1. From the distance matrix (d_{jk}) we form the *association matrix* $\Upsilon = (\Upsilon_{jk})$ by defining $\Upsilon_{jk} = -d_{jk}^2/2$, for all $1 \le j, k \le n$.

Step 2. In the second step, we standardize the matrix Υ so that its rows and columns sum to zero. This is accomplished by defining

$$\bar{\Upsilon}_j = \frac{1}{n}\sum_{k=1}^{n} \Upsilon_{jk} \tag{3.34}$$

and

$$\bar{\Upsilon} = \frac{1}{n}\sum_{j=1}^{n} \bar{\Upsilon}_j \tag{3.35}$$

and then defining the matrix $\Omega = (\Omega_{jk})$ by

$$\Omega_{jk} = \Upsilon_{jk} - \bar{\Upsilon}_j - \bar{\Upsilon}_k + \bar{\Upsilon} \tag{3.36}$$

Gower [74] notes the following result, which we state without proof.

Proposition 3.4.1. *Let (d_{jk}) be a matrix of interpoint distances. Then the matrix Ω defined in Steps 1 and 2 above is nonnegative definite. That is, the eigenvalues of Ω are nonnegative.*

Step 3. In the third step, we construct an $n \times n$ matrix whose jth row is the eigenvector v_j corresponding to the jth largest eigenvalue ω_j of the matrix Ω. The eigenvector v_j is standardized so that $\omega_j = ||v_j||^2$. (This is possible because $\omega_j \ge 0$ for all j, by Proposition 3.4.1.) We can display this $n \times n$ matrix as in (3.37) below.

	$\tilde{x}_1$	$\tilde{x}_2$	...	$\tilde{x}_n$
ω_1	v_{11}	v_{12}	...	v_{1n}
ω_2	v_{21}	v_{22}	...	v_{2n}
.	.	.		.
.	.	.		.
.	.	.		.
ω_n	v_{n1}	v_{n2}	...	v_{nn}

(3.37)

For example, $v_1 = (v_{11}, v_{12}, ..., v_{1n})$. At the left of each row, the eigenvalue ω_j is listed that corresponds to the eigenvector v_j.

Step 4. Reading *across* the columns of this matrix, we obtain the eigenvectors $v_1, ..., v_n$ of the matrix Ω. However, reading *down* the rows of the matrix gives us the required vectors $\tilde{x}_1, ..., \tilde{x}_n$. For example, $\tilde{x}_1 = (v_{11}, v_{21}, ..., v_{n1})^T$.

To prove this result, we shall need the following lemma, which we state without proof.

Lemma 3.4.2. *The eigenvectors of a symmetric $n \times n$ matrix are orthogonal.*

In particular, the matrix Ω is symmetric. From this lemma, we can prove the following:

Proposition 3.4.3. *For $j = 1, ..., n$ let the row vector v_j be the jth eigenvector of Ω with corresponding eigenvalue ω_j. We suppose that v_j is standardized so that*

$$\omega_j = ||v_j||^2 \tag{3.38}$$

Then

$$\Omega = \sum_{j=1}^{n} v_j^T v_j \tag{3.39}$$

Proof. Two $n \times n$ matrices can be shown to be identical if they share common eigenvalues and eigenvectors, and the latter span $\mathbf{R}^n$. It suffices to show that $v_1, ..., v_n$ and $||v_1||^2, ..., ||v_n||^2$ are respectively the eigenvectors and eigenvalues of the matrix $\sum v_j^T v_j$. As the eigenvectors are known to be orthogonal by Lemma 3.4.2, the inner product $v_k v_j^T = 0$ when $j \neq k$. Using the fact that $v_k v_k^T = ||v_k||^2$, we have

$$v_k \left(\sum_{j=1}^{n} v_j^T v_j \right) = \sum_{j=1}^{n} (v_k v_j^T) v_j = ||v_k||^2 v_k \tag{3.40}$$

But equation (3.40) simply establishes that v_k is an eigenvector of $\sum v_j^T v_j$ as required, with eigenvalue ω_k. Q.E.D.

We can now prove our basic result.

Proposition 3.4.4. *Let $\tilde{x}_1, ..., \tilde{x}_n$ be the vectors constructed in step 4 above. Then for all $1 \leq j < k \leq n$, we have*

$$\tilde{d}_{jk} = d(\tilde{x}_j, \tilde{x}_k) = d_{jk} \tag{3.41}$$

Proof. We can write $\tilde{d}_{jk}^2 = ||\tilde{x}_j - \tilde{x}_k||^2$. Expanding this out, we get

$$\tilde{d}_{jk}^2 = \sum_{l=1}^{n} v_{lj}^2 + \sum_{l=1}^{n} v_{lk}^2 - 2 \sum_{l=1}^{n} v_{lj} v_{lk} \tag{3.42}$$

But from Proposition 3.4.3 we can write

$$\sum_{l=1}^{n} v_{lj}^2 = \Omega_{jj} \tag{3.43}$$

$$\sum_{l=1}^{n} v_{lk}^2 = \Omega_{kk} \tag{3.44}$$

and

$$\sum_{l=1}^{n} v_{lj} v_{lk} = \Omega_{jk} \tag{3.45}$$

So

$$\tilde{d}_{jk}^2 = \Omega_{jj} + \Omega_{kk} - 2\Omega_{jk} = \Upsilon_{jj} + \Upsilon_{kk} - 2\Upsilon_{jk} \tag{3.46}$$

We now use the fact that $\Upsilon_{jk} = -d_{jk}^2/2$ and that $\Upsilon_{jj} = \Upsilon_{kk} = 0$ from Step 1 to obtain the desired conclusion that

$$\tilde{d}_{jk}^2 = d_{jk}^2 \tag{3.47}$$

and complete the proof. Q.E.D.

The dimensionality of $\tilde{x}_1, ..., \tilde{x}_n$ is typically too high for convenient graphical representation. However, the principal coordinate analysis also provides a principal component analysis of these points. The eigenvalues have been ordered in decreasing size from top to bottom in the rows of the matrix (v_{jk}), thereby ordering the coordinates of $\tilde{x}_1, ..., \tilde{x}_n$ from the coordinates along the axis with highest variation (coordinates at the top) to those of lowest variation (at the bottom). So, for example, to choose a two-dimensional projection of the vectors $\tilde{x}_1, ..., \tilde{x}_n$, we can take the $2 \times n$ block consisting of the first two rows of (v_{jk}) in (3.37).

In shape analysis for planar landmarks, we will start with a matrix (d_{jk}) of interpoint geodesic distances $d(\sigma_j, \sigma_k)$ between shapes, rather than the matrix of Euclidean distances described above. In this case, $d(\sigma_j, \sigma_k)$ will be the Procrustean distance between two shapes in Σ_2^n. Now, there is no *a priori* guarantee that the matrix Ω will be nonnegative definite, as in

Proposition 3.4.1, because the interpoint geodesic distances on a manifold satisfy different inequalities from those in Euclidean space. Nevertheless, the matrix Ω can be calculated from the matrix (d_{jk}), and its eigenvalues can be checked. If the Procrustean distances d_{jk} can be approximated by Euclidean interpoint distances, then the largest eigenvalues of Ω will be positive. So, for example, if the first two principal eigenvalues are positive, then the $2 \times n$ matrix of the first two rows in (3.37) can be constructed. If we define $\tilde{x}_j \in \mathbf{R}^2$ to be the jth column of this $2 \times n$ matrix for $j = 1, ..., n$, then $\tilde{x}_1, ..., \tilde{x}_n$ will be a two-dimensional configuration whose interpoint distances approximate (d_{jk}).

More generally, with k of the eigenvalues positive, we can construct a set of points $\tilde{x}_1, ..., \tilde{x}_n$ in $\mathbf{R}^k$. (For graphical purposes, the first two dimensions, called the first two principal coordinates, are the most important.) The degree to which all the eigenvalues of Ω are nonnegative can be used as a diagnostic check on the ability to represent Procrustean distances using Euclidean approximations. This is because there exists a converse to Proposition 3.4.1, which we have effectively proved: if all the eigenvalues of Ω are positive, then the Procrustean interpoint distances can be displayed in Euclidean space.

3.5 An Application of Principal Coordinate Analysis to Brooch Data

Let us now apply the techniques of principal coordinate analysis to the Iron Age brooch data described in Chapter 1. Figure 3.8 shows the lateral and superior views of 28 brooches. In our analysis, we shall use only the lateral image. However, for a more complete shape analysis, both perspectives need to be studied. From each of the 28 brooches four landmarks are chosen according to the method described in Chapter 1 and illustrated in Figure 1.1. The Procrustean distance between the shapes of the landmarks is computed according to formula (1.21) for every pair of brooches. This gives us a 28×28 matrix of interpoint Procrustean distances for a set of 28 points in Σ_2^4. The first two principal coordinates of the principal coordinate analysis are shown in Figure 3.9. The first principal coordinate is displayed horizontally, and in broad terms appears to be measuring the degree of elongation of the brooch as seen through the lateral perspective. The second principal coordinate, measured vertically, seems in rough terms to measure the proportional size of the triangle made from the three leftmost landmarks relative to the entire configuration of four landmarks. It should also be noted that the centroid of the points has been fixed at the origin as an artifact of the procedure. The positions of brooches 1, 2, and 3 in relation to each other, as determined in Chapter 1, has been reconfirmed by this analysis. The reader can see, by inspection of Figures 3.8 and 3.9,

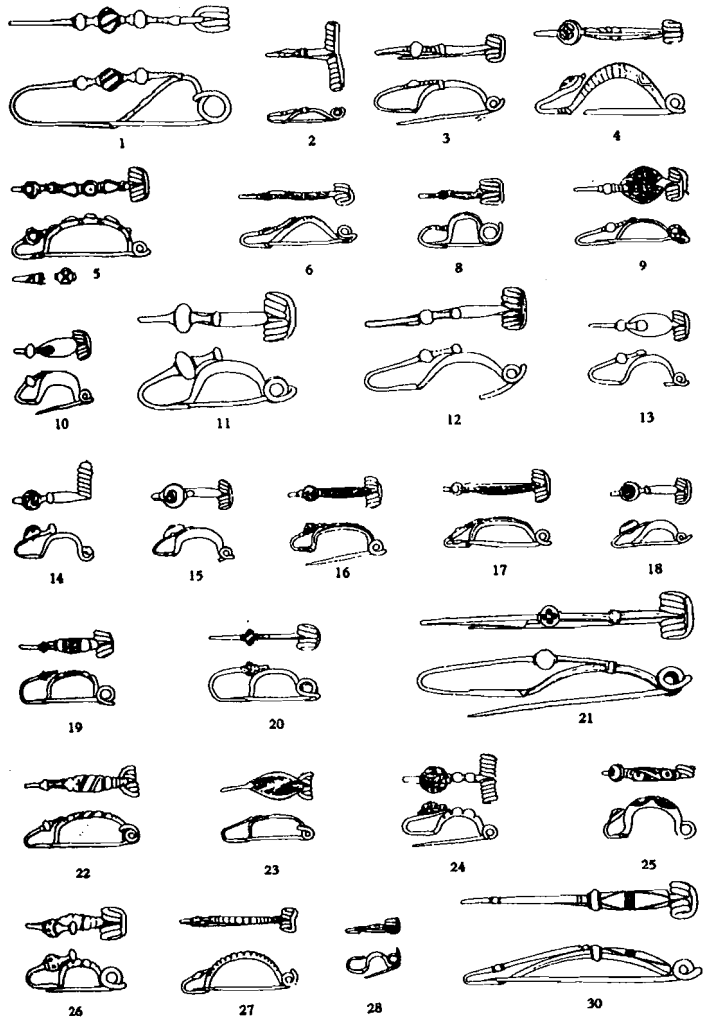

FIGURE 3.8. *Side and top views of 28 Iron Age brooches. Brooches are labeled from 1 to 30. Note that brooches 7 and 29 do not appear in the diagram. The brooches are reproduced from Hodson, Sneath, and Doran,* Biometrika *53 (1966), p. 315, by kind permission of* Biometrika Trustees.

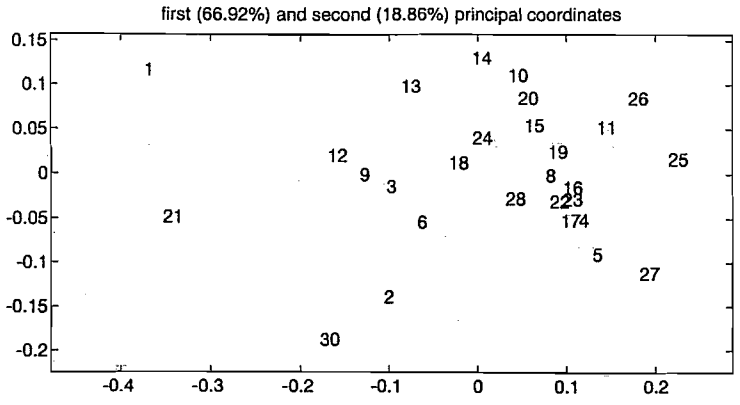

FIGURE 3.9. *Principal coordinate analysis of Iron Age brooches*

that there is a dependence between size and shape of the brooches. Those brooches with a small value for the first principal coordinate are particularly elongated, and also tend to be larger in size.

It is of greater archeological interest to investigate the relationship between the ages of the brooches and their shape. We divide the brooches into five groups from the earliest (group 1) to the latest (group 5).

Group 1: brooches 4, 5, 6, 8, 27, and 28;
Group 2: brooches 15, 18, 22, 23, and 25;
Group 3: brooches 11, 13, 16, 17, 19, 20, and 24;
Group 4: brooches 1, 3, 9, 10, 12, 14, and 26;
Group 5: brooches 2, 21, and 30.

Under these groupings a pattern becomes apparent. Most of the older brooches are to the right-hand side of Figure 3.9, while the younger brooches are to the left. The relationship is not a strict one, but the overall trend is evident. We can summarize our conclusions by saying that with the passing of time, the brooches at Münsingen became larger and more elongated.

This principal coordinate analysis suffers from the defect that it uses only a small part of the total information available from the images. In Chapter 6, we shall explore an automated homology routine that can establish a more complete correspondence between the features.

3.6 Hyperbolic Geometries for Shapes

3.6.1 Singular Values and the Poincaré Plane

We shall begin by developing a geometric theory of triangle shapes due to Bookstein [19]. In our discussion of shape differences up to this point we have assumed that shape differences can be measured by calculating the distances between points, or landmarks, that have been appropriately centered, scaled, and matched as to orientation. A rather different view of shape variation is obtained if we regard the landmarks as selected from homologous positions on bodies whose shapes themselves differ. The differences in the shapes of landmark data or point sets are then seen to be derived from the shape differences in the bodies from which they are chosen. In the biological sciences, this assumption is commonplace. Indeed, in such applications, two different sets of landmarks, or points, may be the corresponding points on a single organism differing in time. As we have argued earlier, the growing organism can undergo a steady transformation of shape that will transform landmarks through various shape changes as the organism changes.

More generally, we might suppose that two sets of distinct landmarks $x_1, ..., x_n$ and $y_1, ..., y_n$ in $\mathbf{R}^p$ are related by a transformation $h : \mathbf{R}^p \to \mathbf{R}^p$ such that $h(x_j) = y_j$ for all $j = 1, ..., n$. The degree to which h departs from the family of similarity transformations can be used as a measure of shape difference. Consider the case where h is an affine transformation of the plane $\mathbf{R}^2$. In matrix form, we can write the transformation h as

$$h(x) = \Lambda x + a \tag{3.48}$$

where x and a are 2×1 column vectors and Λ is a 2×2 square matrix. Henceforth, we shall restrict the analysis to the case where $\det(\Lambda) > 0$. Figure 3.10 shows how an affine transformation affects the shape of a two-dimensional figure. To measure the departure of h from the family of similarity transformations, consider the ellipse that is the image of the unit circle about the origin under the transformation h. We can write this ellipse as

$$\{\Lambda x + a \ : \ x \in \mathbf{R}^2, \ ||x|| = 1\} \tag{3.49}$$

The affine transformation maps a unit circle to an ellipse with semimajor axis of length α and semiminor axis of length β. The values α and β are the *singular values* of Λ, as defined in Section 2.1.5. The ratio α/β is called the *anisotropy* of Λ. It is a useful measure of shape variation induced by h because h will be a similarity transformation if and only if the anisotropy is equal to one. Therefore, the logarithm $\log(\alpha/\beta)$ of the anisotropy serves as a measure of departure of h from the family of similarity transformations. Henceforth, we shall refer to this as the *log-anisotropy*. This measure is also invariant under composition of h with a similarity transformation. See Figure 3.10.

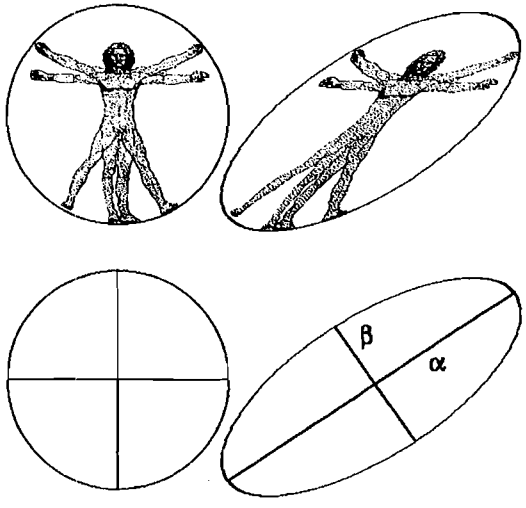

FIGURE 3.10. *Shape change induced by an affine transformation. In this picture, we see the effect of a shear on the shape of a figure. To measure the distortion in shape induced by an affine transformation* $x \to x\Lambda + a$ *we consider how* Λ *transforms a circle to an ellipse. The lengths* α *and* β *of the semimajor and semiminor axes, respectively, are the singular values of the matrix* Λ. *The ratio of* α *and* β *or the logarithm of this ratio can be used to measure the shearing effect of the affine transformation.*

Let x_1, x_2, and x_3 be three planar landmarks that are not collinear. For any other noncollinear landmarks y_1, y_2, and y_3 there exists a unique affine transformation h such that $y_j = h(x_j)$ for $j = 1, 2,$ and 3. Let us standardize the orientation of the triangles $x_1x_2x_3$ and $y_1y_2y_3$ by supposing that they are labeled in a counterclockwise direction. As we are only interested in the difference in shape between $x_1x_2x_3$ and $y_1y_2y_3$, we can map both triangles by a similarity transformation that anchors x_1 and y_1 at the point -1 in the complex plane and similarly anchors x_2 and y_2 at $+1$, as we did in Figure 3.1. The landmarks x_3 and y_3 are then mapped to the respective Bookstein coordinates for the shapes of the two triangles. Let $z = (z_1, z_2)$ and $w = (w_1, w_2)$ be the Bookstein coordinates of $x_1x_2x_3$ and $y_1y_2y_3$ respectively. As the labeling of the triangles is counterclockwise, the Bookstein coordinates z and w will lie in the upper half plane. The affine transformation that maps -1, $+1$, and z to -1, $+1$, and w, respectively, is a linear transformation. It can be represented by left multiplication of a 2×1 column vector by the upper triangular matrix

$$\Lambda = \begin{pmatrix} 1 & \frac{w_1 - z_1}{z_2} \\ 0 & \frac{w_2}{z_2} \end{pmatrix} \tag{3.50}$$

Thus we have

$$\Lambda \begin{pmatrix} \pm 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \pm 1 \\ 0 \end{pmatrix} \tag{3.51}$$

and

$$\Lambda \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \tag{3.52}$$

Now let us consider the form of Λ when w is an infinitesimal perturbation of z. We can then write $w = z + dz$ with coordinates $w_1 = z_1 + dz_1$ and $w_2 = z_2 + dz_2$. See Figure 3.11. The matrix Λ can then be written as $I + d\Lambda$, where I is the 2×2 identity matrix, and

$$d\Lambda = \frac{1}{z_2} \begin{pmatrix} 0 & dz_1 \\ 0 & dz_2 \end{pmatrix} \tag{3.53}$$

To find the singular values of Λ, we first calculate the eigenvalues of $\Lambda^T\Lambda$. Because Λ is a perturbation of the identity matrix, we can write $\Lambda^T\Lambda$ as

$$(I + d\Lambda)^T(I + d\Lambda) = I + (d\Lambda^T + d\Lambda) \tag{3.54}$$

The characteristic equation for the eigenvalues of $\Lambda^T\Lambda$ can be simplified using equation (3.54) and written as

$$\det[\lambda I - (I + d\Lambda^T + d\Lambda)] = 0 \tag{3.55}$$

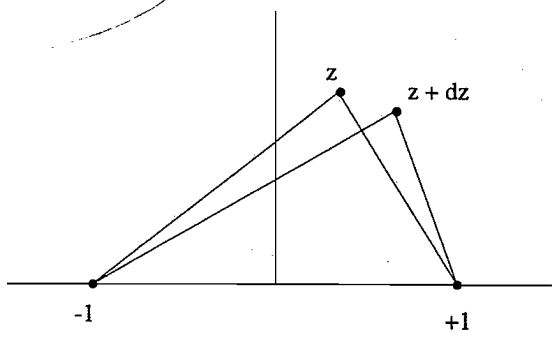

FIGURE 3.11. *An infinitesimal change in the Bookstein coordinates of triangle shape. As three landmarks x_1, x_2, and x_3 are perturbed to landmarks x_1+dx_1, x_2+dx_2, and x_3+dx_3 so the Bookstein coordinates $z=(z_1,z_2)$ are perturbed to $z+dz=(z_1+dz_1,z_2+dz_2)$. The matrix Λ is a perturbation of the identity matrix I. Therefore we can write $\Lambda = I + d\Lambda$.*

This is a quadratic equation in λ. The eigenvalues of $\Lambda^T\Lambda$ are the two roots of this equation, and can be seen to be perturbations of unity. So we can write these eigenvalues as $\lambda_1 = 1+d\lambda_1$ and $\lambda_2 = 1+d\lambda_2$. The roots of the quadratic equation in (3.55) can be found using the time-honored formula known to all high school students. We find that λ_1 and λ_2 are

$$1 + \frac{dz_2 \pm \sqrt{dz_2^2 + dz_1^2}}{z_2} \tag{3.56}$$

Let λ_1 be the larger of these two eigenvalues and λ_2 the smaller. As we are working in the upper half plane of Bookstein coordinates, the coordinate z_2 is positive. So λ_1 has the plus sign in (3.56) while λ_2 has the minus sign.

The singular values of Λ are the square roots of the eigenvalues of $\Lambda^T\Lambda$. They are also perturbations of unity, and can be written as

$$\alpha = \sqrt{1+d\lambda_1} = 1+\frac{d\lambda_1}{2} \tag{3.57}$$

and

$$\beta = \sqrt{1+d\lambda_2} = 1+\frac{d\lambda_2}{2} \tag{3.58}$$

So the log-anisotropy of Λ will be

$$\log(\alpha/\beta) = \frac{d\lambda_1}{2} - \frac{d\lambda_2}{2} \tag{3.59}$$

Plugging (3.56) into (3.59) we obtain the log-anisotropy, and thereby the infinitesimal distance between the shapes with Bookstein coordinates z and $z+dz$. This gives us the following result:

Proposition 3.6.1. *The infinitesimal distance from Bookstein coordinates z to $z+dz$ is given by*

$$ds = \frac{\sqrt{dz_1^2 + dz_2^2}}{z_2} \tag{3.60}$$

where $dz = (dz_1, dz_2)$.

This can be recognized as the distance formula for the Poincaré Plane, as given in formula (2.101). With this measure of infinitesimal distance, the upper half plane of Bookstein coordinates becomes the Poincaré Plane $\mathbf{HS}^2$.

It should be noted in passing that the infinitesimal distance ds is not dependent upon which two of the three landmarks are mapped to ± 1. It is only necessary that homologous landmarks x_j and y_j be mapped correspondingly.

3.6.2 A Generalization into Higher Dimensions

It is possible to generalize the shape manifold $\mathbf{HS}^2$ of triangle shapes to a family of manifolds of shapes of $n+1$ landmarks in n dimensions, provided the $n+1$ landmarks are in general position in $\mathbf{R}^n$. This is equivalent to requiring that the simplex that has these landmarks as its vertices has positive n-dimensional volume.

Let $x=(x_1,...,x_{n+1})$ be any set of $n+1$ landmarks in $\mathbf{R}^n$ in general position. The coordinates of the jth landmark x_j shall be denoted as $(x_{j1},...,x_{jn})$. We begin by arranging the coordinates of these landmarks into an $n\times(n+1)$ matrix whose jth column is the vector of coordinates of the jth landmark. We can eliminate the information about location in the landmarks by subtracting off the first column from all the others, yielding the $n\times n$ matrix

$$\Xi_x = \begin{pmatrix} x_{21}-x_{11} & x_{31}-x_{11} & \cdots & x_{(n+1)1}-x_{11} \\ x_{22}-x_{12} & x_{32}-x_{12} & \cdots & x_{(n+1)2}-x_{12} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ x_{2n}-x_{1n} & x_{3n}-x_{1n} & \cdots & x_{(n+1)n}-x_{1n} \end{pmatrix} \tag{3.61}$$

The matrix Ξ_x can be called the *pre-size-and-shape matrix* of the landmarks, for reasons that will be made clear below.

Next, we eliminate orientation information in the landmarks. Suppose we let ξ_j be the $n\times 1$ column vector consisting of the jth column of Ξ_x. A Gram-Schmidt orthogonalization of the vectors $\xi_1,...,\xi_n$ produces a set of orthonormal vectors $\xi_1',...,\xi_n'$ with the property that ξ_j lies in

the subspace generated by $\xi_1', ..., \xi_j'$ and such that $< \xi_j, \xi_j' >$ is positive. Let Ω be the orthogonal $n \times n$ matrix whose jth column is the vector ξ_j'. Then $\Psi_x = \Omega^{-1}\Xi_x$ can be shown to be an upper triangular matrix with positive entries down the main diagonal. For the proof of this result, see Problem 7. The matrix Ω^{-1} produces an orthogonal transformation of the column vectors of Ξ_x that standardizes the orientation information in Ξ_x. For this reason, we can call Ψ_x the *size-and shape matrix* of the landmarks.

Next, we eliminate scale information in Ψ_x by dividing every element of this $n \times n$ matrix by the element in the upper left corner. We need have no fear that this element of the matrix is zero because the elements on the main diagonal of Ψ_x are all positive. Upon dividing every element of Ψ_x by the upper leftmost element, we are left with the upper triangular matrix

$$\Pi_x = \begin{pmatrix} 1 & z_{31} & z_{41} & \cdots & z_{(n+1)1} \\ 0 & z_{32} & z_{42} & \cdots & z_{(n+1)2} \\ 0 & 0 & z_{43} & \cdots & z_{(n+1)3} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & z_{(n+1)n} \end{pmatrix} \tag{3.62}$$

which is the matrix representation of the shape of the landmarks.

The reduction to shape coordinates has proceeded via a series of reductions. First, we reduced to the pre-size-and-shape matrix Ξ_x, then to the size-and-shape matrix Ψ_x, and finally, after standardization, to the shape matrix Π_x.

The reason for the rather strange labeling of the elements of Π_x is the following. Suppose we define

$$z_1 = (0,\ 0,\ 0,\ ...,\ 0) \tag{3.63}$$

$$z_2 = (+1,\ 0,\ 0,\ ...,\ 0) \tag{3.64}$$

and for $3 \leq j \leq n+1$,

$$z_j = (z_{j1},\ z_{j2},\ ...,\ z_{j(j-1)},\ 0,\ 0,\ ...,\ 0) \tag{3.65}$$

Then the simplex with vertices $x_1,\ x_2,\ ...,\ x_{n+1}$ (or its mirror image) and the simplex with vertices $z_1,\ z_2,\ ...,\ z_{n+1}$ have the same shape. The coordinates defined by (3.63)–(3.65) encode the information about the shape of the landmarks $x_1, ..., x_{n+1}$. Thus we have the following definition.

Definition 3.6.2. *The coordinates*

$$z = (z_{jk})_{3 \leq j \leq n+1, 1 \leq k \leq j-1} \tag{3.66}$$

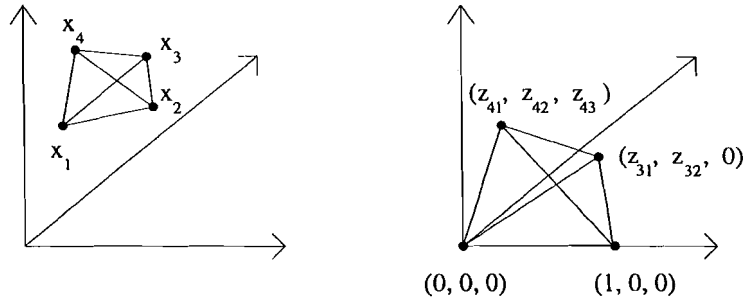

FIGURE 3.12. *Generalized Bookstein coordinates for a simplex in three dimensions. The simplex is mapped by a similarity transformation so that the landmarks x_1 and x_2 are mapped to $(0,0,0)$ and $(1,0,0)$ respectively. The simplex is rotated about the axis through $(0,0,0)$ and $(1,0,0)$ until the third landmark is of the form $(z_{31}, z_{32}, 0)$ with $z_{32} > 0$. If the coordinate z_{43} is negative, the fourth landmark is reflected through the plane of the other three landmarks to make this coordinate positive. Compare this figure with Figure 3.1.*

shall be called generalized Bookstein coordinates *of* $x_1, ..., x_{n+1}$.

See Figure 3.12. The reader should note that that these coordinates do not generalize Bookstein coordinates in the strict sense because for the case $n = 2$ the simplex with vertices at $z_1, ..., z_{n+1}$ has its first point anchored at 0 rather than at -1, as was the case in the previous section.

Definition 3.6.3. *We define* $\mathbf{UT}(n)$ *to be the set of all upper triangular* $n \times n$ *matrices* $\Pi = (\Pi_{jk})$ *for which* $\Pi_{11} = 1$ *and for which the diagonal elements* Π_{jj} *are all positive.*

We shall call the matrix Π_x of (3.62) the *upper triangular shape representation* of $x = (x_1, ..., x_{n+1})$, or the *UT-shape representation* of x for short.

It is easy to see that $\mathbf{UT}(n)$ is closed under matrix multiplication and inversion. Moreover, since the identity matrix is in $\mathbf{UT}(n)$ it follows that $\mathbf{UT}(n)$ is a group with matrix multiplication.

Our next task is to make $\mathbf{UT}(n)$ into a Riemannian manifold by constructing a metric tensor on it. Let Π_x be an element of $\mathbf{UT}(n)$. We perturb Π_x to a neighboring matrix Π_{x+dx}. To introduce a metric tensor on $\mathbf{UT}(n)$, we need to find the singular values of $\Pi_{x+dx}\Pi_x^{-1}$. The extent to which these singular values differ from each other is a measure of the shape change induced by left multiplication by the matrix $\Pi_{x+dx}\Pi_x^{-1}$. Let

$$\Lambda = I + d\Lambda = \Pi_{x+dx}\Pi_x^{-1} \tag{3.67}$$

To construct a metric tensor on $\mathbf{UT}(n)$ we need to find an appropriate

quadratic form on the coordinates of $d\Lambda$. There are n eigenvalues of the matrix $\Lambda^T\Lambda$, and these eigenvalues, as perturbed values of unity, can be written as $\lambda_j = 1 + d\lambda_j$ for $j = 1, ..., n$. Unlike the case for $n = 2$, these eigenvalues cannot generally be found with simple algebraic expressions. Fortunately this is unnecessary, as the first and second moments of the eigenvalues can be computed from the coefficients of the characteristic polynomial

$$\det[\lambda I - \Lambda^T\Lambda] = \det[\lambda I - (I + d\Lambda^T + d\Lambda)] = 0 \tag{3.68}$$

Writing this in the form

$$\lambda^n - a_1\lambda^{n-1} + a_2\lambda^{n-2} - ... + (-1)^n a_n = 0 \tag{3.69}$$

we recall that

$$a_1 = \sum_j \lambda_j \tag{3.70}$$

and

$$a_2 = \sum_{1\le j<k\le n} \lambda_j\lambda_k \tag{3.71}$$

from which the first and second moments of the eigenvalues can be computed. We shall define the metric tensor on the space of simplex shapes so that ds^2 is the variance of the eigenvalues of $\Lambda^T\Lambda$. In terms of the coefficients of the characteristic polynomial, this gives us the following definition:

Definition 3.6.3. *Let Π_x and Π_{x+dx} be the UT-shape representations of x and $x + dx$ and let $\Lambda = \Pi_{x+dx}\Pi_x^{-1}$. We define the infinitesimal distance ds from Π_x to Π_{x+dx} to be given by the formula*

$$ds^2 = \frac{\sum_{j=1}^n(\lambda_j - \bar\lambda)^2}{n} = \frac{(n-1)a_1^2}{n^2} - \frac{2a_2}{n} \tag{3.72}$$

where $\lambda_1, ..., \lambda_n$ are the eigenvalues of $\Lambda^T\Lambda$. Also, $\bar\lambda = n^{-1}\sum_j \lambda_j$, and a_1 and a_2 are the coefficients of the characteristic polynomial of $\Lambda^T\Lambda$ given by equations (3.70) and (3.71).

In particular, formula (3.60) is a special case of formula (3.72) where $n = 2$. This metric is related to a theory of norms on upper triangular matrices due to Frobenius and von Neumann. See [79]. The infinitesimal distance measure ds can be interpreted as the evaluation of a von Neumann seminorm on the upper triangular matrix leading to the infinitesimal shape change.

In order to evaluate the components of the metric tensor on $\mathbf{UT}(n)$ we need to write out the coefficients a_1 and a_2 in terms of the elements

$d\Lambda_{jk}$ of the matrix $d\Lambda$. To do this, we return to the characteristic equation given in (3.68) and evaluate the coefficients explicitly. We obtain

$$a_1 = \sum_{j=1}^n (1 + 2d\Lambda_{jj}) \tag{3.73}$$

which is the trace of $I + d\Lambda^T + d\Lambda$. The second coefficient is a sum over determinants of 2×2 minors of $I + d\Lambda^T + d\Lambda$, namely

$$a_2 = \sum_{1\le j<k\le n} [(1 + 2d\Lambda_{jj})(1 + 2d\Lambda_{kk}) - d\Lambda_{jk}^2] \tag{3.74}$$

Note that $d\Lambda_{11} = 0$ because $\Lambda_{11} = 1$. When we plug (3.73) and (3.74) into (3.72), the terms of order 1 and of order $d\Lambda$ cancel, and we are left with a quadratic form in the differentials $d\Lambda_{jk}$. The coefficients of the quadratic form are the metric tensor. Our formula for ds^2 becomes

$$\frac{4}{n^2}\left[(n-1)\sum_{j=2}^n d\Lambda_{jj}^2 + \frac{n}{2}\sum_{j<k} d\Lambda_{jk}^2 - 2\sum_{j<k} d\Lambda_{jj}d\Lambda_{kk}\right] \tag{3.75}$$

As a check on this formula, we can plug in the 2×2 matrix $d\Lambda$ given in formula (3.53). The expression in (3.75) can then be seen to reduce to formula (3.60).

At this stage, it is appropriate to ask how the metric tensor on $\mathbf{UT}(n)$ changes if we label the points x_1, x_2, ..., x_{n+1} in a different order. In the case of Kendall's Procrustean geometry, we found that a relabeling (or permutation) of the points induced an isometry on the shape manifold Σ_2^n. Similarly, this is the case here.

Proposition 3.6.4. *The shape metric ds^2 on $\mathbf{UT}(n)$ given by formula (3.72) is invariant under permutations. That is, let $j(1), ..., j(n+1)$ be a permutation of the integers $1, 2, ..., n+1$. We define a mapping p_n : $\mathbf{UT}(n) \to \mathbf{UT}(n)$ taking the UT-shape representation of $x_1, ..., x_{n+1} \in \mathbf{R}^n$ to the UT-shape representation of $x_{j(1)}, ..., x_{j(n+1)}$. Then p_n is an isometry of $\mathbf{UT}(n)$.*

Proof. It suffices to show that p_n preserves the metric tensor. Let $x = (x_1, ..., x_n)$ and

$$x' = (x_{j(1)}, ..., x_{j(n+1)}) \tag{3.76}$$

In addition, let Λ be that element of $\mathbf{UT}(n)$ such that $x + dx = \Lambda x$. Similarly, let $x' + dx' = \Lambda x'$. We can write Λ and Λ' as perturbations of the identity matrix. Thus we have $\Lambda = I + d\Lambda$ and $\Lambda' = I + d\Lambda'$. Suppose that

$$1 + d\lambda_1 \ge 1 + d\lambda_2 \ge ... \ge 1 + d\lambda_n \tag{3.77}$$

are the eigenvalues of $\Lambda^T\Lambda = I + d\Lambda^T + d\Lambda$ arranged in decreasing order. In a similar vein, suppose

$$1 + d\lambda'_1 \geq 1 + d\lambda'_2 \geq \ldots \geq 1 + d\lambda'_n \tag{3.78}$$

are the eigenvalues of $(\Lambda')^T\Lambda'$.

It is not hard to see that the eigenvalues of $\Lambda^T\Lambda$ and $(\Lambda')^T\Lambda'$ are scaled versions of each other. So there exists a positive constant c such that $1 + d\lambda'_j = c(1 + d\lambda_j)$ for all $j = 1, 2, \ldots, n$. The constant c can be written as $c = 1 + dc$, which means that we can write $1 + d\lambda'_j = 1 + d\lambda_j + dc$. So the effect of the permutation $j(1), \ldots, j(n+1)$ is to shift these eigenvalues by an amount dc. However, the variance of the eigenvalues is invariant under shifts of location, which implies that

$$\sum_{j=1}^{n} \frac{(\lambda'_j - \bar{\lambda}')^2}{n} = \sum_{j=1}^{n} \frac{(\lambda_j - \bar{\lambda})^2}{n} \tag{3.79}$$

from which the result follows. Q.E.D.

Matrix multiplication on the right in $\mathbf{UT}(n)$ can be shown to be a family of isometries of this Riemannian manifold, as the following proposition states.

Proposition 3.6.5. *The family of transformations* $\mathbf{UT}(n) \to \mathbf{UT}(n)$ *of right matrix multiplications* $\Pi \to \Pi\Lambda$ *for each* $\Lambda \in \mathbf{UT}(n)$ *is a group of isometries on* $\mathbf{UT}(n)$.

Proof. It is sufficient to prove that the metric tensor on $\mathbf{UT}(n)$ is invariant under right multiplication. This follows fairly easily from the construction of the metric, and is left to Problem 5. Q.E.D.

Problem 6 asks the reader to show that left matrix multiplications are *not* isometries of $\mathbf{UT}(2)$.

3.6.3 Geodesic Distance in $\mathbf{UT}(2)$

The metric tensor of Definition 3.6.3 can be used to measure the geodesic distance of any $\Lambda \in \mathbf{UT}(n)$ from the identity matrix I. For example, let us consider the geodesic distance from I to any Λ in $\mathbf{UT}(2)$. This geodesic distance will be a function of the singular values α and β of the matrix Λ. The orientation of the axes is quite incidental to the calculation. Therefore, we can choose the axis to be along the direction of the eigenvectors of $\Lambda^T\Lambda$. Thus we can reduce, without loss of generality, to the case where Λ is of the form

$$\Lambda = \begin{pmatrix} 1 & 0 \\ 0 & \alpha/\beta \end{pmatrix} \tag{3.80}$$

where $\alpha \geq \beta$. This linear transformation maps the Bookstein coordinates $(0, 1)$ to $(0, \alpha/\beta)$. We can find the geodesic distance from I to Λ in $\mathbf{UT}(2)$ directly from Proposition 3.6.1. Setting $dz_1 = 0$ and integrating along the vertical axis from $z_2 = 1$ to $z_2 = \alpha/\beta$ we find the geodesic distance to be

$$\int_1^{\alpha/\beta} \frac{1}{z_2}\, dz_2 = \log(\alpha/\beta) \tag{3.81}$$

Compare this formula with formula (2.105). Not surprisingly, we return to our original log-anisotropy, and now are able to interpret it as a geodesic distance between Bookstein coordinates or upper triangular matrices. Henceforth, we shall call (3.81) the *anisotropy metric* on the half space of Bookstein coordinates. See the notes at the end of the chapter for the generalization of this metric to $\mathbf{UT}(n)$.

3.6.4 The Geometry of Tetrahedral Shapes

Let us now consider the detailed geometry of $\mathbf{UT}(3)$, the space of tetrahedral shapes in three dimensions. Let x be a vector of four landmarks in $\mathbf{R}^3$ and let x be perturbed to $x + dx$. The infinitesimal distance ds between Π_x and Π_{x+dx} will be the standard deviation of the eigenvalues of $\Lambda = I + d\Lambda$, where $\Pi_{x+dx} = \Lambda\Pi_x$.

Let

$$z_{31},\ z_{32},\ z_{41},\ z_{42},\ \text{and}\ z_{43} \tag{3.82}$$

be the generalized Bookstein coordinates of x. Similarly, let $z_{jk} + dz_{jk}$ be the (j,k)th generalized Bookstein coordinate of $x + dx$. It is straightforward to check that

$$d\Lambda = \begin{pmatrix} 0 & dz_{31}/z_{32} & dz_{41}/z_{43} - (z_{42}\, dz_{31})/(z_{32}z_{43}) \\ 0 & dz_{32}/z_{32} & dz_{42}/z_{43} - (z_{42}\, dz_{32})/(z_{32}z_{43}) \\ 0 & 0 & dz_{43}/z_{43} \end{pmatrix} \tag{3.83}$$

From formula (3.75) we can write ds^2 as

$$ds^2 = dz_3^T\, g_1\, dz_3 + 2\, dz_3^T\, g_2\, dz_4 + dz_4^T\, g_3\, dz_4 \tag{3.84}$$

where

$$g_1 = \frac{1}{3z_{32}^2 z_{43}^2} \begin{pmatrix} 2(z_{42}^2 + z_{43}^2) & 0 \\ 0 & 2(3z_{42}^2 + 4z_{43}^2)/3 \end{pmatrix} \tag{3.85}$$

$$g_2 = \frac{1}{3z_{32}z_{43}^2} \begin{pmatrix} -2z_{42} & 0 & 0 \\ 0 & -2z_{42} & -4z_{43}/3 \end{pmatrix} \tag{3.86}$$

and

$$g_3 = \frac{2}{3z_{43}^2}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 4/3 \end{pmatrix} \tag{3.87}$$

and where

$$dz_3 = (dz_{31},\ dz_{32})^T \tag{3.88}$$

and

$$dz_4 = (dz_{41},\ dz_{42},\ dz_{43})^T \tag{3.89}$$

Unlike the metric tensor for $\mathbf{UT}(2)$, which was isometric to the Poincaré Plane $\mathbf{HS}^2$, the space $\mathbf{UT}(3)$ has off-diagonal elements in its metric tensor. Note, however, that g_3 is diagonal in form. Suppose that we set $dz_{31} = 0$ and $dz_{32} = 0$ in the formula for ds^2. This is equivalent to fixing the shape of the triangular base of a tetrahedron made up of the first three points. In addition, suppose that we transform the coordinate z_{43} by setting

$$z'_{43} = \frac{2\ z_{43}}{\sqrt{3}} \tag{3.90}$$

Then the coordinates z_{41}, z_{42}, z'_{43} can be seen to form a three-dimensional half space with metric

$$ds^2 = \frac{8}{9}\left[\frac{(dz_{41})^2 + (dz_{42})^2 + (dz'_{43})^2}{(z'_{43})^2}\right] \tag{3.91}$$

This is isometric to $\mathbf{HS}^3$ except for a scale factor.

3.7 Local Analysis of Shape Variation

3.7.1 Thin-Plate Splines

In the previous section, we considered shape differences due to the rather restrictive class of affine transformations of $\mathbf{R}^n$, and the Euclidean plane in particular. However, in practice, much more general transformations are necessary to explain shape differences. Let us revisit Figure 1.7 and the four skulls in profile. How can we quantify the variation in shape that is evident from the curvilinear coordinates?

The first problem we encounter is that of constructing transformations that correspond to the pictures constructed by Thompson's method of coordinates. Thompson [172] suggested a number of simple classes of transformations of biological interest, including the simple affine transformations of the previous section. As is evident from Figure 1.7, such transformations

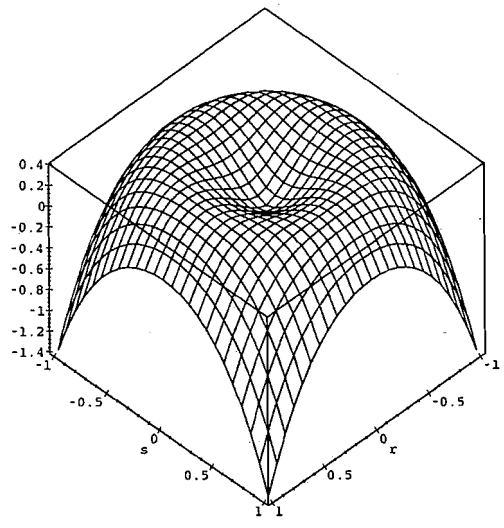

FIGURE 3.13. *A plot of the function* γ *in formula (3.92).*

are too simplistic to account for the detailed shape variation among images such as the skulls. Alternatively, we can construct a set of homologous landmarks on corresponding images and then extend the homology to the entire image by a *spatial interpolation* routine. Let $x_1, ..., x_n$ and $y_1, ..., y_n$ be two sets of planar landmarks drawn from homologous images. We shall suppose that x_j and y_j are situated at homologous positions on the corresponding images. A spatial interpolation routine is one that constructs a function $h : \mathbf{R}^2 \to \mathbf{R}^2$ subject to the constraint that $h(x_j) = y_j$ for all $j = 1, ..., n$.

Bookstein [19], [23] has proposed a family of thin-plate splines that uses the landmarks $x_1, ..., x_n$ as the knots of the spline.

To define the family of thin-plate splines, we first define a function $\gamma : \mathbf{R}^2 \to \mathbf{R}$ by

$$\gamma(r, s) = -(r^2 + s^2)\ \log(r^2 + s^2) \tag{3.92}.$$

The function γ is plotted in Figure 3.13.

Thin-plate splines can be visually interpreted as the surfaces

$$\{(r, s, \delta(r, s)) \in \mathbf{R}^3\ :\ r, s \in \mathbf{R}\} \tag{3.93}$$

corresponding to functions of the form

$$\delta(r,s) \;=\; (a_0 + a_1 r + a_2 s) + \sum_{j=1}^{n} b_j \gamma(r - x_{j1}, s - x_{j2}) \tag{3.94}$$

where $x_j = (x_{j1}, x_{j2})$ are landmarks in $\mathbf{R}^2$. The constants a_k are arbitrary real values. However, the constants b_j are not completely arbitrary, but are constrained to satisfy

$$\sum_{j=1}^{n} b_j = \sum_{j=1}^{n} b_j x_{j1} = \sum_{j=1}^{n} b_j x_{j2} = 0 \tag{3.95}$$

We should note that this class of thin-plate splines contains the class of affine transformations as a special case where $b_1 = ... = b_n = 0$.

The family of surfaces defined by (3.93) is the mathematical solution to a problem in physics. If a thin metal plate is constrained to pass through a set of points that are almost coplanar, then the plate will take a shape that minimizes the *bending energy* required to deform its shape from a flat surface. The mathematical solution to the problem of minimizing the bending energy is the family of surfaces defined by (3.93).

The class of thin-plate splines of (3.93) can be used to build interpolating functions. Suppose $x_1, ..., x_n$ are n landmarks in $\mathbf{R}^2$, possibly standardized with respect to location, scale, and orientation. Let $y_1, ..., y_n$ be another set of landmarks in $\mathbf{R}^2$ such that x_j is homologous to y_j for $j = 1, ..., n$. The thin-plate splines in (3.94) provide us with a tool for finding an interpolating spline

$$h : \mathbf{R}^2 \;\rightarrow\; \mathbf{R}^2 \tag{3.96}$$

such that $h(x_j) = y_j$ for all $j = 1, ..., n$. The *thin-plate spline* interpolant will be of the form

$$h : (r,s) \;\rightarrow\; [\delta_1(r,s),\ \delta_2(r,s)] \tag{3.97}$$

where $\delta_1(r,s)$ and $\delta_2(r,s)$ are of the form (3.94). The constants a_k and b_j for each of δ_1 and δ_2 are chosen to satisfy the constraint that $h(x_j) = y_j$, for all j. See Figure 3.14 for two examples of such thin-plate splines.

These constants can be calculated as follows: Let $P = (P_{jk})$ be the $n \times n$ matrix whose diagonal elements P_{jj} are all zero, and whose off-diagonal elements are

$$P_{jk} \;=\; \gamma(x_{j1} - x_{k1},\ x_{j2} - x_{k2}) \tag{3.98}$$

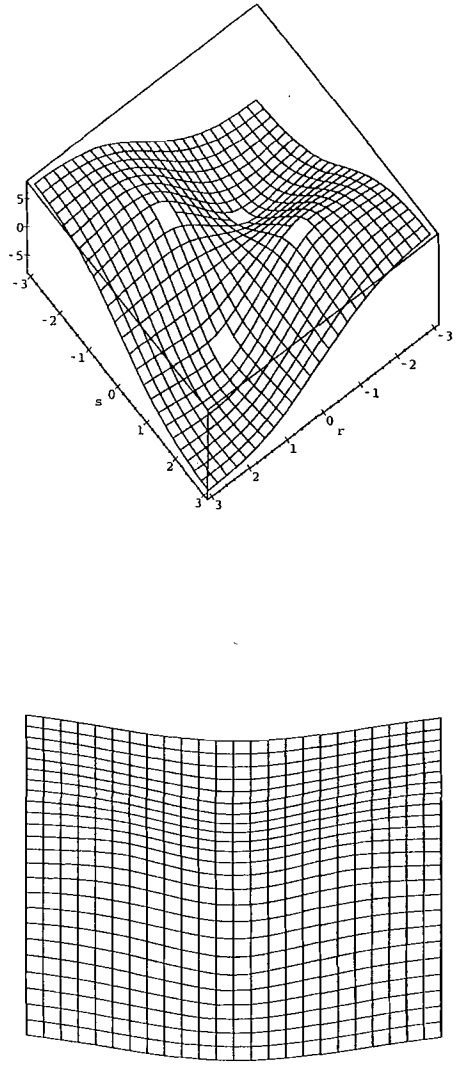

FIGURE 3.14. *Thin-plate splines. In the top figure we see an example of the function δ as defined in (3.94) using four landmarks $x_1, ..., x_4$ at the vertices of a square. The coefficients a_j have been set to zero. The bottom figure illustrates how a thin-plate spline h, as defined in (3.97), can warp a cartesian coordinate system in the plane.*

Next, we define the matrix

$$Q = \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ . & . & . \\ . & . & . \\ . & . & . \\ 1 & x_{n1} & x_{n2} \end{pmatrix} \tag{3.99}$$

The coefficients in (3.97) can then be determined by solving

$$\left(\begin{array}{c|c} P & Q \\ \hline Q^T & 0 \end{array}\right)^{-1} \begin{pmatrix} y_{11} & y_{12} \\ . & . \\ . & . \\ y_{n1} & y_{n2} \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} b_{11} & b_{12} \\ . & . \\ . & . \\ b_{n1} & b_{n2} \\ a_{01} & a_{02} \\ a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \tag{3.100}$$

The coefficients of δ_1 and δ_2 can be read from the first and second columns respectively on the right-hand side of (3.100).

It is worth making a few comments about Bookstein's thin-plate splines. First of all, we should note that the family of thin-plate splines in (3.97) is not invariant under function inversion. In general, the function h^{-1} is not a thin-plate spline when h is. This implies that the landmarks x_j and y_j cannot be treated symmetrically. It is customary to consider the configuration of landmarks $x_1, ..., x_n$ as selected from a canonical (textbook) image against which one or more other images are to be compared.

Secondly, the thin-plate spline $h : \mathbf{R}^2 \to \mathbf{R}^2$ need not be 1-1. In other words, h can fold the plane so that two distinct points are mapped onto the same point. In applications where a thin-plate spline produces a fold, the researcher should consider whether this is physically meaningful.

Finally, the family of thin-plate splines in (3.97) is closed under similarity transformations of the plane. That the family is closed under translation and rotation of the domain or range is obvious. Also obvious is its closure under scale changes to the range of the splines. So only the closure of the family of thin-plate splines under scale changes in the domain needs to be checked with any care. If $h(r, s)$ is a thin-plate spline, it can be seen that under rescaling by a factor a the function $h(ar, as)$ is also a thin-plate spline. This fact is partly a consequence of the linear constraints of (3.95).

3.7.2 Local Anisotropy of Nonlinear Transformations

Let $h : U \to V$ be a smooth transformation between two open subsets of the plane $\mathbf{R}^2$. Let us suppose that the function h establishes a homology

between two images that are subsets of U and V respectively. In this section, we shall be interested in quantifying the differences in shape between the two images by assessing the degree to which h varies from a similarity transformation in $\mathbf{Sim}(2)$. However, we shall not assume that h comes from a restricted class of functions such as the affine transformations of the plane.

The method that we shall consider is based in part upon the geometry of Section 3.6, and can be used to investigate the *local* shape variation caused by h. Suppose h maps the point $(r, s) \in \mathbf{R}^2$ to the point (u, v). To first order, the local properties of h around $x \in U$ can be determined by the Jacobian matrix $\Lambda(r, s)$. We have

$$\Lambda(r, s) = \begin{pmatrix} \partial u/\partial r & \partial u/\partial s \\ \partial v/\partial r & \partial v/\partial s \end{pmatrix} \tag{3.101}$$

Let us restrict to the case where h preserves the orientation of the plane. This is equivalent to requiring that the determinant of $\Lambda(r, s)$ be positive for all (r, s).

Let $\alpha = \alpha(r, s)$ and $\beta = \beta(r, s)$ be the two singular values of Λ, with $\alpha \geq \beta$. To measure the local shape change caused by the stretching effect of h at (r, s) we can use the anisotropy metric

$$\mathcal{M}_2(r, s) = \log(\alpha/\beta) \tag{3.102}$$

Now suppose that $\mathcal{M}_2(r, s) = 0$ for all (r, s). Then the matrix $\Lambda(r, s)$ is a scalar multiple of an orthogonal matrix for all (r, s). Moreover, because h is restricted in our discussion to those transformations that preserve the orientation of the plane, it follows that Λ is a scalar multiple of a special orthogonal matrix, which is a rotation of $\mathbf{R}^2$. This implies that

$$\frac{\partial u}{\partial r} = \frac{\partial v}{\partial s} \qquad \frac{\partial u}{\partial s} = -\frac{\partial v}{\partial r} \tag{3.103}$$

These are seen to be the *Cauchy-Riemann equations*, used to verify that h is a complex analytic function. So, the function $\mathcal{M}_2$ provides a measure of departure of h from the family of complex analytic functions. Complex analytic functions are *conformal*, and we see that conformal mappings, which preserve local angles, are local similarity transformations.

In a general number of dimensions, we can also extend the definition of $\mathcal{M}_2$ to functions $h : \mathbf{R}^n \to \mathbf{R}^n$ using the Riemannian geometry of $\mathbf{UT}(n)$. If Λ is the Jacobian matrix of h we can standardize it via a similarity transformation of $\mathbf{R}^n$ to be an element of $\mathbf{UT}(n)$. The geodesic distance of this standardized matrix in $\mathbf{UT}(n)$ from the identity matrix I is a measure of local shape change induced by the stretching effect of Λ. The resulting function $\mathcal{M}_n$ measures the degree of departure of h from a conformal transformation of $\mathbf{R}^n$. For dimensions $n \geq 3$, the

conformal transformations of Euclidean space $\mathbf{R}^n$ are more restricted than the conformal transformations of $\mathbf{R}^2$. It can be shown that any conformal transformation of $\mathbf{R}^n$, $n \geq 3$, is necessarily a Moebius transformation, characterized in dimension n as a diffeomorphism that maps $(n-1)$-spheres to $(n-1)$-spheres. Although these transformations are restricted, they can still change shapes, as they include inversion transformations. Thus it is useful to try to supplement $\mathcal{M}_n$ by a function that measures shape variation a different way.

3.7.3 Another Measure of Local Shape Variation

While $\mathcal{M}_n$ is a useful measure of local shape variation, it does not provide a complete description of the changes in shape induced by a transformation h. As we have seen, conformal transformations can be markedly different from the similarity transformations that preserve shape, despite the fact that they are locally similarity transformations themselves.

At this stage, it is helpful to appeal to the allometric model of Section 1.2, in which biological shape changes are modeled as occurring when different parts of an organism grow at different rates. This suggests that we look for two components to shape change, measured by $\mathcal{M}_n$ and a new measure that we shall call $\mathcal{N}_n$.

Conformal mappings (for which $\mathcal{M}_n$ vanishes) have nonconstant Jacobians, and therefore have heterogeneous scale changes at the local level. In the case of similarity transformations, the Jacobian of the transformation is a constant function equal to the nth power of the scale factor induced by h. This suggests that we try to measure the variability of the Jacobian as a measure of shape change. In dimension two, let us define

$$\mathcal{N}_2 = \sqrt{\left\{\frac{\partial}{\partial r}[\log(\mathcal{J}h)]\right\}^2 + \left\{\frac{\partial}{\partial s}[\log(\mathcal{J}h)]\right\}^2} \tag{3.104}$$

In general dimensions, using more compact notation, we can define

$$\mathcal{N}_n = ||\nabla \log(\mathcal{J}h)|| \tag{3.105}$$

where ∇ is the gradient operator for real valued functions on $\mathbf{R}^n$. The transformation of the Jacobian by the logarithm ensures that the function $\mathcal{N}_n$ is invariant under homogeneous scale changes of $\mathbf{R}^n$, measuring only the heterogeneity in scale changes by h.

Earlier, we found that the transformations for which $\mathcal{M}_n \equiv 0$ were the conformal transformations of $\mathbf{R}^n$. It is also useful to examine the class of smooth transformations of $\mathbf{R}^n$ for which $\mathcal{N}_n \equiv 0$. Any such transformation h must have a constant Jacobian $\mathcal{J}h$. Integrating volume elements over subsets of $\mathbf{R}^n$, we find that h must be equivalent, except for an arbitrary similarity transformation, to a *volume-preserving* transformation of $\mathbf{R}^n$.

Another way of saying this is that h must preserve the ratios of volumes: if A and B are subsets of $\mathbf{R}^n$, then the ratio of the volume of A to that of B is the same as the ratio of the volume of $h(A)$ to the volume of $h(B)$. This property of h is reminiscent of certain aspects of the allometric approach to shape analysis. The log-volumes of A and B can be regarded as size variables in an image, and their difference as a shape variable. When $\mathcal{N}_n \equiv 0$, these allometric shape variables are left invariant by h.

As we shall see in the following proposition, the functions $\mathcal{M}_n$ and $\mathcal{N}_n$ together describe the total variation in shape due to the transformation h.

Proposition 3.7.1. *Suppose that $h : \mathbf{R}^n \to \mathbf{R}^n$ is a diffeomorphism such that $\mathcal{M}_n \equiv 0$ and $\mathcal{N}_n \equiv 0$ throughout $\mathbf{R}^n$. Then $h \in \mathbf{Sim}(n)$.*

Proof. As $\mathcal{M}_n \equiv 0$ it follows that the Jacobian matrix of h is locally a rescaling of an orthogonal matrix. That is, there exists a positive function $\lambda : \mathbf{R}^n \to \mathbf{R}$ such that $\lambda\Lambda$ is an orthogonal matrix at all points in $\mathbf{R}^n$. But because $\mathcal{N}_n \equiv 0$ we also observe that $\mathcal{J}h$ is a constant function throughout $\mathbf{R}^n$. Moreover, $\mathcal{J}h \equiv \lambda^{-n}$. Therefore, λ is constant function on $\mathbf{R}^n$. Thus we can say that $h = \lambda h_0$, where λ is a positive scalar, and Λ_0, the Jacobian matrix of h_0, is an orthogonal matrix at all points in $\mathbf{R}^n$.

However, any function h_0 must be an isometry of $\mathbf{R}^n$ if Λ_0 is everywhere an orthogonal matrix. To prove this, consider two points $x, y \in \mathbf{R}^n$, and consider the line segment L with endpoints x and y. Let $h_0(L)$ be that arc in $\mathbf{R}^n$ from $h_0(x)$ to $h_0(y)$ consisting of the image under h_0 of all points on the line segment L. Because Λ_0 is an orthogonal matrix, it follows that h_0 is locally an isometry, so that the length of the path $h_0(L)$ is equal to the length of the line segment L. But L has length $||x - y||$, and the length of $h_0(L)$ is greater than or equal to $||h_0(x) - h_0(y)||$, the length of the arc being at least as great as the distance between its endpoints. Therefore $||x - y|| \geq ||h_0(x) - h_0(y)||$. However, a similar argument using h_0^{-1} rather than h_0 shows that $||x - y|| \leq ||h_0(x) - h_0(y)||$. Thus $||x - y|| = ||h_0(x) - h_0(y)||$, and h_0 is an isometry of $\mathbf{R}^p$. From this fact, we conclude that h is a similarity transformation of $\mathbf{R}^p$. Q.E.D.

We can see the effects of these two types of local shape variation by considering the curvilinear coordinates of Figure 1.7. The coordinate system for the modern human skull was chosen to be a standard Cartesian coordinate system, with intersecting lines meeting at orthogonal and equally spaced parallel lines in each direction. This divides the region into squares. In the curvilinear coordinate systems below this, the images of the squares in the first coordinate system are approximately parallelograms except for those regions where shape variation is occurring too rapidly for the coarseness of

the grid. The function $\mathcal{M}_2$ measures those shape changes that stretch the squares of the top grid into the parallelograms of the lower grids. Another type of effect that we can observe is that while the squares in the top grid are all of the same area, the parallelograms of the lower grids have varying areas. This effect is measured by $\mathcal{N}_2$.

3.8 Notes

Bookstein's approach to shape analysis leads to a manifold of constant negative curvature for the representation of triangle shapes. In contrast to this, Kendall's approach leads to a sphere, which is a manifold of constant positive curvature. This discrepancy need not confuse us nor lead us to consider one geometry superior to the other. In each case, the Riemannian geometry of triangle shape space is motivated by consideration of the mechanisms that give rise to shape variation. Our generalization of Fred Bookstein's triangle shape geometry to the family of shape spaces $\mathbf{UT}(n)$ provides an alternative to the family of spaces Σ_n^{n+1} introduced by David Kendall.

Huiling Le has recently computed the anisotropy metric for $\mathbf{UT}(n)$. This is the generalization of formula (3.81) from $\mathbf{UT}(2)$ to the higher-dimensional simplex shape spaces $\mathbf{UT}(n)$, where $n > 2$. The following proposition can be compared with Proposition 3.6.4, which is a special case for infinitesimal distances.

Proposition 3.8.1. *Let Π_x and Π_y be the UT-shape representations of x and y respectively. Let $\Lambda = \Pi_y \Pi_x^{-1}$. The square of the geodesic distance from Π_x to Π_y in $\mathbf{UT}(n)$ is given by*

$$\sum_{j=2}^{n} \frac{1}{j(j-1)n} \left[(j-1)\ \log(\lambda_j/\lambda_1) - \sum_{k=1}^{j-1} \log(\lambda_k/\lambda_1) \right]^2 \tag{3.106}$$

where $\lambda_1, ..., \lambda_n$ are the eigenvalues of $\Lambda^T \Lambda$.

It can be checked that this reduces to formula (3.81) when $n = 2$ and to formula (3.72) when $y = x + dx$.

For $n > 2$ the simplex shape spaces are not spaces of constant curvature.

3.9 Problems

1. Consider two similar triangles $x_1x_2x_3$ and $y_1y_2y_3$ in the plane. We define the average of the two triangles to be $z_1z_2z_3$ where z_1, z_2, and z_3 are the midpoints of x_1y_1, x_2y_2, and x_3y_3, respectively. Show that the average of $x_1x_2x_3$ and $y_1y_2y_3$ is similar to these triangles. Does this result hold if the triangles are not constrained to lie in the same plane?

2. Show that the 1-1 correspondence between Σ_2^3 and $\mathbf{S}^2(1/2)$ established in formula (3.9) of Section 3.1 is a Riemannian isometry. More specifically, show that formula (1.21) for the distance between shape points on the sphere is equivalent to formula (3.9) for shape distance on Σ_2^3. Hint: as both formulas are invariant under similarity transformations, it is sufficient to consider two triangles, $-1, +1, z_1$ and $-1, +1, z_2$, of complex landmarks and to compute the distance between their shapes by the two methods. First find the coordinates of their pre-shapes and plug into formula (1.21). Then find the coordinates of shape on the sphere from formula (3.7) and plug into formula (3.9). Do you get the same answer?

3. Find all points on the sphere $\mathbf{S}^2(1/2) \cong \Sigma_2^3$ that correspond to right triangles. What does this region look like?

4. Find all points on the sphere $\mathbf{S}^2(1/2) \cong \Sigma_2^3$ that correspond to isosceles triangles. What does this region look like?

5. Prove Proposition 3.6.5.

6. Show that left matrix multiplications in $\mathbf{UT}(2)$ are not isometries by considering what happens to the matrices

$$\begin{pmatrix} 1 & x \\ 0 & y \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & x+dx \\ 0 & y+dy \end{pmatrix} \tag{3.107}$$

under left multiplication by the matrix

$$\begin{pmatrix} 1 & a \\ 0 & b \end{pmatrix} \tag{3.108}$$

In addition, confirm the results of Proposition 3.6.5 as applied to $\mathbf{UT}(2)$ by performing the same calculation using right multiplication.

7. Let Ω be the orthogonal $n \times n$ matrix defined in Section 3.6.2. Let ξ_j be the jth column of Ξ_x.

(a) Using the fact that ξ_j lies in the subspace generated by $\xi_1', ..., \xi_j'$, show that $\Omega^{-1}\xi_j$ has only its first j elements nonzero. (Hint: what does $\Omega^{-1}\xi_j'$ look like?) Conclude that Ψ_x is an upper triangular matrix.

(b) Using the fact that $< \xi_j, \xi_j' >$ is positive, show that the entries down the main diagonal of $\Psi_x = \Omega^{-1}\Xi_x$ are also positive.

8. At the end of Section 3.4, we noted in passing that it is possible to

arrange a set of points on a Riemannian manifold so that their interpoint geodesic distances do not match the interpoint Euclidean distances of any configuration of points in any dimension. In this problem we shall verify this. Let x_j, $j = 1, ..., 4$, be four points spaced at equal intervals around the unit circle $\mathbf{S}^1$.

(a) Find the 6×6 matrix of interpoint geodesic distances between the points x_j using arc length to measure distance.

(b) Show that this 6×6 matrix cannot be an interpoint Euclidean distance matrix for any set of four points in any Euclidean space $\mathbf{R}^n$.

4

Some Stochastic Geometry

4.1 Probability Theory on Manifolds

4.1.1 Sample Spaces and Sigma-Fields

We begin with a review of some basic definitions and ideas from probability theory. The reader wishing a more detailed description of the tools that will be necessary can consult [43].

By a *sample space* we shall mean a set $\mathcal{S}$ whose elements s shall be called *outcomes* or *points*. Within $\mathcal{S}$ we shall suppose that we have a particular class $\mathcal{F}$ of subsets $A \subset \mathcal{S}$ that shall be called *events*. This class has to be sufficiently rich to allow us to do probability calculations. To do this we shall require that $\mathcal{F}$ be a *sigma-field* of subsets, which we now define.

Definition 4.1.1. *A class $\mathcal{F}$ of subsets of $\mathcal{S}$ is said to be a* sigma-field *on $\mathcal{S}$ if the following three properties are satisfied.*
1. $\mathcal{S} \in \mathcal{F}$.
2. If $A \in \mathcal{F}$ then its complement $A^c \in \mathcal{F}$.
3. For any sequence $A_1, A_2, A_3, ...$ of elements of $\mathcal{F}$ the union

$$\bigcup_{j=1}^{\infty} A_j = \{s \in \mathcal{S} : s \in A_j \text{ for some } j\} \tag{4.1}$$

is an element of $\mathcal{F}$.

Henceforth, we shall assume that the class $\mathcal{F}$ of events is a sigma-field. From Definition 4.1.1, it is possible to show that any subset of $\mathcal{S}$ constructed as a countable Boolean combination of events in $\mathcal{F}$ is itself an event in $\mathcal{F}$.

4.1.2 Probabilities

We can now define a probability on a sample space.

Definition 4.1.2. *By a* probability, *we mean a function*

$$\mathcal{P} : \mathcal{F} \to \mathbf{R} \tag{4.2}$$

such that $0 \leq \mathcal{P}(A) \leq 1$ *for all* $A \in \mathcal{F}$ *satisfying the properties*

$$\mathcal{P}(\mathcal{S}) = 1 \tag{4.3}$$

and

$$\mathcal{P}\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} \mathcal{P}(A_j) \tag{4.4}$$

whenever $A_j \cap A_k = \emptyset$ *for all* $j \neq k$. *The set* $\mathcal{S}$ *when endowed with a sigma-field* $\mathcal{F}$ *of events and with a probability* $\mathcal{P}$ *is said to be a* probability space.

If $\mathcal{G}$ is any class of subsets of a set $\mathcal{S}$ we can also define the *sigma-field generated by* $\mathcal{G}$. This is the intersection of all sigma-fields $\mathcal{G}'$ such that $\mathcal{G} \subset \mathcal{G}'$. It is clearly the smallest sigma-field containing the subsets $A \in \mathcal{G}$.

4.1.3 Statistics on Manifolds

Of particular interest are functions from a sample space into a manifold called *statistics.*

Definition 4.1.3. *Let* $\mathbf{M}^p$ *be a differential manifold. By a* statistic X *on* $\mathbf{M}^p$ *we shall mean a function*

$$X : \mathcal{S} \to \mathbf{M}^p \tag{4.5}$$

with the property that

$$X^{-1}(U) = \{s \in \mathcal{S} : X(s) \in U\} \tag{4.6}$$

is an event (i.e., an element of $\mathcal{F}$*) for every open set* $U \subset \mathbf{M}^p$.

In the special case where $\mathbf{M}^p = \mathbf{R}$ we also refer to a statistic X on $\mathbf{R}$ as a *random variable.* More generally, a statistic on $\mathbf{R}^p$ is called a *random vector.*

We can also define a class $\mathcal{B}$ of subsets on $\mathbf{M}^p$ called the *Borel subsets.*

Definition 4.1.4. *The class* $\mathcal{B}$ *of* Borel subsets *of* $\mathbf{M}^p$ *is defined as the sigma-field of subsets of* $\mathbf{M}^p$ *generated by the class* $\mathcal{U}$ *of open subsets of* $\mathbf{M}^p$.

Thus all the open sets of $\mathbf{M}^p$ are Borel sets including $\mathbf{M}^p$ itself. In addition, the countable intersection of open subsets of $\mathbf{M}^p$ is also Borel, although it is, in general, not open. Closed subsets are also Borel, because they are the complements of open sets in $\mathbf{M}^p$. The possible types of Borel sets are not exhausted by this list as the possible types of sets generated by countably taking intersections, unions, and complements in any order is very rich indeed.

For any Borel set $B \subset \mathbf{M}^p$ and for all statistics $X : \mathcal{S} \to \mathbf{M}^p$, the set

$$X^{-1}(B) = \{s \in \mathcal{S} : X(s) \in B\} \tag{4.7}$$

is an event in $\mathcal{S}$. That is, if B is a Borel set then $X^{-1}(B) \in \mathcal{F}$. This property indicates the importance of Borel sets in $\mathbf{M}^p$. They form a natural class of subsets B for which we can assign a probability that a statistic X lies in B.

4.1.4 Induced Distributions on Manifolds

Because the class of Borel sets $\mathbf{B}$ is a sigma-field, we can regard the manifold $\mathbf{M}^p$ as a sample space in its own right, with $\mathcal{B}$ as its class of events. A statistic X then induces a probability on $\mathcal{B}$ in the same way that the original probability $\mathcal{P}$ is defined on $\mathcal{F}$. We can define the *induced probability distribution* $\mathcal{P}X^{-1}$ on the Borel sets of $\mathbf{M}^p$ to be

$$\mathcal{P}X^{-1}(B) = \mathcal{P}\left[X^{-1}(B)\right] \tag{4.8}$$

It can be checked that $\mathcal{P}X^{-1}$ satisfies the properties of a probability on $\mathbf{M}^p$ given in Definition 4.1.2 above. In much of probability theory, the fact that statistics are functions on sample spaces tends to be suppressed in the notation. Thus we shall henceforth write $X^{-1}(B)$ as $(X \in B)$, both being equivalent to the event

$$\{s \in \mathcal{S} : X(s) \in B\} \tag{4.9}$$

So we shall typically write $\mathcal{P}(X \in B)$ for the probability in equation (4.8). Other notations are similar. For example, if $\mathbf{M}^p = \mathbf{R}$, we write $(X \leq t)$ for the set $(X \in (-\infty, t])$, etc. Another abbreviation is to use a comma to

stand for the logical operation "and" and the corresponding set operation of intersection. Thus we shall write

$$(X_1 \in B_1,\ X_2 \in B_2) \tag{4.10}$$

to stand for $(X_1 \in B_1) \cap (X_2 \in B_2)$.

We can make new statistics out of old. One way to do this is through the use of Cartesian products. For example, if X_1 is a statistic on $\mathbf{M}^p$ and X_2 is a statistic on $\mathbf{N}^q$ then $X = (X_1, X_2)$ is a statistic on $\mathbf{M}^p \times \mathbf{N}^q$. Correspondingly, any statistic X on $\mathbf{M}^p \times \mathbf{N}^q$ defines X_1 and X_2 uniquely on $\mathbf{M}^p$ and $\mathbf{N}^q$ respectively. Another way to build new statistics out of old ones is through composition of functions. For example, if X is a statistic on $\mathbf{M}^p$ and $h : \mathbf{M}^p \to \mathbf{N}^q$ is continuous, then

$$h(X) : \mathcal{S} \to \mathbf{N}^q \tag{4.11}$$

is a statistic on $\mathbf{N}^q$. This follows from the fact that the continuous preimage $h^{-1}(B)$ of a Borel set $B \subset \mathbf{N}^q$ is a Borel set of $\mathbf{M}^p$.

4.1.5 Random Vectors and Distribution Functions

Suppose X is a random vector taking values in Euclidean space $\mathbf{R}^p$. If we write X in terms of its coordinates $X_1, ..., X_p$ then $X_1, ..., X_p$ are random variables. We define the *distribution function* of X to be a real valued function $F_X : \mathbf{R}^p \to \mathbf{R}$ such that

$$F_X(t_1,\ ...,\ t_p) = \mathcal{P}(X_1 \le t_1,\ ...,\ X_p \le t_p) \tag{4.12}$$

It is important and nontrivial to show that the induced distribution $\mathcal{P}X^{-1}$ on the Borel sets of $\mathbf{R}^p$ is determined by its joint distribution function F_X. We say that a random variable X is *absolutely continuous*, or simply *continuous*, if there exists a nonnegative function $f : \mathbf{R} \to \mathbf{R}$ such that

$$\mathcal{P}(X \in U) = \int_{x \in U} f(x)\, dx \tag{4.13}$$

for all open sets $U \subset \mathbf{R}$. Equation (4.13) holds true if U is replaced by any Borel subset of $\mathbf{R}$. A random vector X taking values in $\mathbf{R}^p$ is said to be absolutely continuous if the higher-dimensional analog of (4.13) holds for some nonnegative function $f : \mathbf{R}^p \to \mathbf{R}$ and all open $U \subset \mathbf{R}^p$. Many probability distributions can be constructed on $\mathbf{R}^p$ that are not continuous, although many of the models in this book will be of the continuous type. Another important class of probability distributions are the *discrete* probability distributions, which assign unit probability to some countable set.

4.1.6 Stochastic Independence

We now briefly review some basic definitions and properties related to stochastic independence. Let $X_1, X_2, ..., X_n$ be statistics taking values in a differential manifold $\mathbf{M}^p$. These statistics are said to be *mutually stochastically independent*, or simply *independent*, if for all Borel sets $B_1, B_2, ..., B_n \subset \mathbf{M}^p$ we have

$$\mathcal{P}(X_1 \in B_1, ..., X_n \in B_n) = \prod_{j=1}^{n} \mathcal{P}(X_j \in B_j) \tag{4.14}$$

If the induced distributions $\mathcal{P}X_j^{-1}$ on $\mathbf{M}^p$ are equal in the sense that

$$\mathcal{P}(X_1 \in B) = \mathcal{P}(X_2 \in B) = ... = \mathcal{P}(X_n \in B) \tag{4.15}$$

for all Borel $B \subset \mathbf{M}^p$ then we shall say that $X_1, ..., X_n$ are *identically distributed*. The condition that random variables are both independent and identically distributed is often abbreviated by saying that they are IID.

4.1.7 Mathematical Expectation

If X is a random variable with distribution function F, then we can define the *mathematical expectation*, also known as the *mean* or *expected value* of X, to be

$$\mathcal{E}(X) = \int_{-\infty}^{+\infty} x\, dF(x) \tag{4.16}$$

for those random variables for which the integral is finite. The expected value of a random vector $X = (X_1, ..., X_n)$ we shall define as the vector of expected values

$$\mathcal{E}(X) = [\mathcal{E}(X_1), ..., \mathcal{E}(X_n)] \tag{4.17}$$

In a similar way, the expected value of a matrix can be defined as the matrix of expected values of its elements.

We cannot do justice in this brief survey to the full range of definitions and results on expectation, independence, conditional probability, and marginalization of distributions. The reader is referred to standard sources for the results needed throughout the remainder of this book.

4.2 The Geometric Measure

We now seek to generalize the concept of a p-dimensional content, or volume, in $\mathbf{R}^p$ to a Riemannian manifold. We have already seen that the metric tensor is instrumental in defining the lengths of paths in a Riemannian manifold. One would naturally expect the metric tensor to be essential

to the definition of content as well. We begin by considering the relationship between the metric tensor and p-dimensional content in $\mathbf{R}^p$. Let

$$x_j = (x_{j1}, x_{j2}, ..., x_{jp}) \tag{4.18}$$

for $j = 1, 2, ..., p$ be a set of p linearly independent vectors in $\mathbf{R}^p$. The basis vectors $x_1, x_2, ..., x_p$ define the edges of a parallelepiped in $\mathbf{R}^p$ given by

$$\left\{ \sum_{j=1}^{p} a_j x_j \ : \ 0 \le a_j \le 1 \ \text{ for all } \ \text{j} \right\} \tag{4.19}$$

A standard result in linear algebra tells us that the volume, or content, of this parallelepiped is the absolute value of the determinant of the matrix

$$\begin{pmatrix} x_{11} & x_{12} & \cdot & \cdot & \cdot & x_{1p} \\ x_{21} & x_{22} & \cdot & \cdot & \cdot & x_{2p} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{p1} & x_{p2} & \cdot & \cdot & \cdot & x_{pp} \end{pmatrix} \tag{4.20}$$

Now, Problem 4 asks the reader to show that

$$|\det(x_{jk})| = \sqrt{\det(g_{jk})} \tag{4.21}$$

where

$$g_{jk} = < x_j, x_k > = \sum_{l=1}^{p} x_{jl} x_{kl} \tag{4.22}$$

is the metric tensor for $\mathbf{R}^p$ endowed with a coordinate system based on $x_1, x_2, ..., x_p$ rather than the usual orthonormal set of vectors. Formula (4.21) tells us that p-dimensional volume is characterized by the metric tensor.

To calculate p-dimensional volume on a general Riemannian manifold we use an approach that is similar to the calculation of arc length using the metric tensor. The metric tensor on a Riemannian manifold $\mathbf{M}^p$ permits us to define the volume of a parallelepiped in $T_x(\mathbf{M}^p)$, the tangent space at $x \in \mathbf{M}^p$. From this we define the volume $d\mathcal{V}_p(x)$ of a small region whose coordinates are

$$\{y = (y_1, y_2, ..., y_p) \in \mathbf{M}^p \ : \ x_j \le y_j \le x_j + dx_j \ \text{ for all } \ \text{j}\} \tag{4.23}$$

to be

$$d\mathcal{V}_p(x) = \sqrt{\det(g_{jk}\, dx_j dx_k)} \tag{4.24}$$

The volume of an open set U in the manifold is then found to be

$$\mathcal{V}_p(U) = \int_{x \in U} d\mathcal{V}_p(x) \tag{4.25}$$

The volume function $\mathcal{V}_p$, when extended to the Borel sets of the manifold, is called the *geometric measure.*

Now suppose S is a probability space endowed with a probability $\mathcal{P}$, and suppose $X : S \to \mathbf{M}^p$ is a statistic on the Riemannian manifold. The statistic X will be said to be *absolutely continuous*, or simply *continuous*, provided that there exists a nonnegative function $f : \mathbf{M}^p \to \mathbf{R}$ such that

$$\mathcal{P}(X \in U) = \int_{x \in U} f(x)\, d\mathcal{V}_p(x) \tag{4.26}$$

for all open sets U on the manifold. If this is the case, we shall call f the *density function* of X. An important special case occurs when $\mathcal{V}_p(\mathbf{M}^p) < \infty$ and

$$f \equiv \frac{1}{\mathcal{V}_p(\mathbf{M}^p)} \tag{4.27}$$

In this case, we say that X is *uniformly distributed* on the manifold, or that the induced distribution on $\mathbf{M}^p$ is uniform. Note that the density can never be constant on a manifold for which $\mathcal{V}_p(\mathbf{M}^p) = \infty$.

4.2.1 Example: Surface Area on Spheres

To illustrate the idea of volume and content, consider the 2-sphere $\mathbf{S}^2(r)$ of radius r from the example in Section 2.2.14. Let θ_1 be the longitude and θ_2 the colatitude of that example as defined in formula (2.63). Then applying formula (4.24) above, we see that

$$d\mathcal{V}_2(\theta_1, \theta_2) = r^2 \sin(\theta_2)\, d\theta_1\, d\theta_2 \tag{4.28}$$

This formula is quite commonly derived from heuristics in multivariate calculus courses. Thus the surface area of the sphere is

$$\mathcal{V}_2\left[\mathbf{S}^2(r)\right] = \int_0^{2\pi} \int_0^{\pi} r^2 \sin(\theta_2)\, d\theta_2\, d\theta_1 = 4\pi r^2 \tag{4.29}$$

as is well known.

4.2.2 Example: Volume in Hyperbolic Half Spaces

Consider the hyperbolic half spaces of Section 2.2.17. For these spaces

$$d\mathcal{V}_p(x_1, x_2, ..., x_p) = x_p^{-p}\, dx_1\, dx_2 \ ... \ dx_p \tag{4.30}$$

Close to the horizon at infinity where $x_p = 0$, when measured in Euclidean coordinates the volume element $d\mathcal{V}_p$ goes to infinity. Unlike their positive curvature counterparts, the spheres $\mathbf{S}^p$, the hyperbolic half spaces $\mathbf{HS}^p$ have infinite volume.

4.3 Transformations of Statistics

4.3.1 Jacobians of Diffeomorphisms

Consider two manifolds $\mathbf{M}^p$ and $\mathbf{N}^p$ of the same dimension, and let $h : \mathbf{M}^p \to \mathbf{N}^p$ be a differentiable function. Suppose $\mathbf{M}^p$ and $\mathbf{N}^p$ have coordinate systems $x = (x_1, ..., x_p)$ and $y = (y_1, ..., y_p)$ respectively. Then we can express h in terms of these coordinates as a function

$$(x_1, ..., x_p) \to (y_1, ..., y_p) \tag{4.31}$$

Then the Jacobian matrix of h can be defined in terms of these coordinates as the matrix of partial derivatives $\Lambda = (\partial y_j / \partial x_k)$ as in formula (2.9). Similarly, we can define the Jacobian of h to be

$$(\mathcal{J}h)_x = \det(\Lambda) \tag{4.32}$$

Note that the Jacobian is *dependent* on the particular coordinate system used on each manifold. That is, the Jacobian is extrinsic to the manifolds. However, it appears in calculations to compensate for changes of coordinates, and therefore can be used to build intrinsic quantities that are independent of the coordinate system. If h is a diffeomorphism between manifolds then

$$(\mathcal{J}h^{-1})_{h(x)} = [(\mathcal{J}h)_x]^{-1} \tag{4.33}$$

Moreover, both quantities will be bounded away from zero and infinity. If h is a local diffeomorphism, the identity (4.33) remains true because of the Jacobian's local nature. But in this case, the inverse transformation has to be interpreted locally.

4.3.2 Change of Variables Formulas

Now suppose X is a statistic on $\mathbf{M}^p$ with density function f and suppose that $h : \mathbf{M}^p \to \mathbf{N}^p$ is a diffeomorphism. We define the statistic $Y = h(X)$, and consider the problem of calculating the density function of Y on $\mathbf{N}^p$. Let $g_{\mathbf{M}}$ be the metric tensor on $\mathbf{M}^p$, and let $g_{\mathbf{N}}$ be the metric tensor on $\mathbf{N}^p$. Then it can be shown that the density function of Y on $\mathbf{N}^p$ can be written in terms of the Jacobian $\mathcal{J}h^{-1}$ and the metric tensors as

$$f[h^{-1}(y)] \frac{d\mathcal{V}_p[h^{-1}(y)]}{d\mathcal{V}_p(y)} \tag{4.34}$$

where the ratio of differentials of the geometric measure is a coordinate-free notation for the expression

$$|(\mathcal{J}h^{-1})_y| \sqrt{\frac{\det g_{\mathbf{M}}[h^{-1}(y)]}{\det g_{\mathbf{N}}(y)}} \tag{4.35}$$

using formula (4.24). In measure-theoretic language, we can also call the ratio of differentials a *Radon-Nikodym derivative.*

An extension of (4.34) allows us to calculate the distribution of $Y = h(X)$ for transformations $h : \mathbf{M}^p \to \mathbf{N}^q$ where $q < p$. Suppose we can find a manifold $\mathbf{N}^{p-q}$ and a transformation

$$h' : \mathbf{M}^p \to \mathbf{N}^{p-q} \tag{4.36}$$

such that

$$(h, h') : \mathbf{M}^p \to \mathbf{N}^q \times \mathbf{N}^{p-q} \tag{4.37}$$

is a diffeomorphism. Let $h_1 = (h, h')$ and $Y' = h'(X)$. The density function of

$$Y_1 = h_1(X) = (Y, Y') \tag{4.38}$$

can be calculated from formula (4.34) above. We then find the *marginal* density of $Y = h(X)$ by integrating this density over its second variable, leading to the formula

$$\int_{\mathbf{N}^{p-q}} f[h_1^{-1}(y_1)] \frac{d\mathcal{V}_p[h_1^{-1}(y_1)]}{d\mathcal{V}_p(y_1)} d\mathcal{V}_{p-q}(y') \tag{4.39}$$

4.4 Invariance and Isometries

In order to prove that a particular induced distribution on $\mathbf{M}^p$ is uniform, it is possible to check directly that its density is constant. Often, however, there is another way based upon the concept of *invariance.* Suppose that for any two points $x, y \in \mathbf{M}^p$ there is a geodesic from x to y. Then there is a well defined concept of geodesic distance between points. For such manifolds, we can use $\mathbf{Iso}(\mathbf{M}^p)$, the group of isometries on $\mathbf{M}^p$, to investigate whether a statistic has a uniform distribution on $\mathbf{M}^p$. Now the volume measure on $\mathbf{M}^p$ is invariant under isometries in the sense that

$$\mathcal{V}_p[h(B)] = \mathcal{V}_p(B) \tag{4.40}$$

for all Borel sets B and for all $h \in \mathbf{Iso}(\mathbf{M}^p)$. Similarly, if X has a uniform distribution on $\mathbf{M}^p$ then the probability distribution is invariant under the group in the sense that

$$\mathcal{P}[X \in h(B)] = \mathcal{P}(X \in B) \tag{4.41}$$

This invariance property of (4.41) is illustrated in Figure 4.1. We shall be concerned with the converse of this result. Namely, if a continuous statistic X has this invariance property with respect to $\mathbf{Iso}(\mathbf{M}^p)$ does it follow that X is uniformly distributed on $\mathbf{M}^p$? The answer, in general, is no.

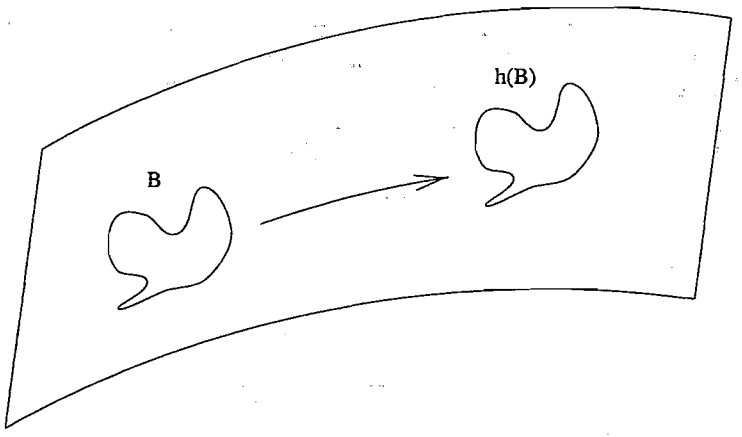

FIGURE 4.1. *Invariance of the uniform distribution with respect to isometries of the manifold. A Borel set B is shifted by an isometry h of the manifold to a set $h(B)$. If X is a uniformly distributed statistic on the manifold, then X will lie with equal probability in B and $h(B)$.*

However, for an important special case, the converse is true. The following definition provides the necessary class of isometries for which a converse can be obtained.

Definition 4.4.1. *Let $\mathbf{H}$ be any group of transformations on $\mathbf{M}^p$. Then $\mathbf{H}$ is said to be* transitive *if for every x and y in $\mathbf{M}^p$, there is an $h \in \mathbf{H}$ such that $h(x) = y$.*

Using the concept of transitivity of the group action on $\mathbf{M}^p$, we can obtain our converse.

Proposition 4.4.2. *Let X be a continuous statistic on the manifold $\mathbf{M}^p$, with density function $f(x)$. Suppose that $\mathbf{H}$ is any transitive subgroup of $\mathbf{Iso}(\mathbf{M}^p)$, and that the distribution of X is invariant under $\mathbf{H}$ in the sense of equation (4.41) above. Then f is a constant density on $\mathbf{M}^p$. That is, X is uniformly distributed on $\mathbf{M}^p$.*

Proof. This result follows from a special case of formula (4.34), in which $\mathbf{M}^p = \mathbf{N}^p$ and $g_M = g_N$. To prove that X is uniform, we must show that $f(x) = f(y)$ for all $x, y \in \mathbf{M}^p$. As $\mathbf{H}$ is a transitive group, it follows that there exists an $h \in \mathbf{H}$ such that $h(x) = y$. We can set up coordinate systems around x and y so that $g_{\mathbf{M}}(x)$ and $g_{\mathbf{N}}(y)$ are the identity matrices. (This can be achieved at specific points in a manifold, but will only hold over an open set when the manifold is flat on that set.) Because

h is an isometry, it follows that the Jacobian matrices of h and h^{-1} will be orthogonal at the points x and y respectively. Therefore,

$$(\mathcal{J}h^{-1})_y = \pm 1 \tag{4.42}$$

Furthermore, because X has a distribution that is invariant for all $h \in \mathbf{Iso}(\mathbf{M}^p)$, the statistics X and $h(X)$ have the same distribution and the same density functions. Therefore, formula (4.34) gives us

$$f(y) = f[h^{-1}(y)] \tag{4.43}$$

for all y. An immediate consequence of this is that f must be a constant function when $\mathbf{Iso}(\mathbf{M}^p)$ is transitive. Thus X must have a uniform distribution. Q.E.D.

4.4.1 Example: Isometries of Spheres

In 4.2.1 we wrote the differential $d\mathcal{V}_2$ of surface area in terms of the coordinate system. We now characterize surface area through invariance. Both $\mathbf{O}(p)$ and $\mathbf{U}(p)$ are isometry groups on $\mathbf{R}^p$ and $\mathbf{C}^p$ respectively. Now $\mathbf{O}(p)$ maps the unit sphere $\mathbf{S}^{p-1}$ to itself, and is therefore an isometry group for $\mathbf{S}^{p-1}$. It is a simple exercise to show that this group acts transitively on $\mathbf{S}^{p-1}$. See Problem 6 at the end of the chapter. The uniform distribution is the unique distribution on $\mathbf{S}^2$ that is invariant under the action of $\mathbf{O}(3)$. More generally, the uniform distribution on $\mathbf{S}^{p-1}$ is the unique invariant distribution under the action of $\mathbf{O}(p)$.

Now, the unitary group $\mathbf{U}(q)$ is identifiable as a subgroup of $\mathbf{O}(2q)$. Thus $\mathbf{U}(q)$ is a group of isometries of $\mathbf{S}^{2q-1}$. With a bit of work, we can also show that $\mathbf{U}(q)$ acts transitively on $\mathbf{S}^{2q-1}$. See Problem 7. Thus the uniform distribution on $\mathbf{S}^{2q-1}$ is also characterized as the unique distribution that is invariant under $\mathbf{U}(q)$.

4.4.2 Example: Isometries of Real Projective Spaces

Let

$$\mathcal{A} : \mathbf{S}^p \to \mathbf{RP}^p \tag{4.44}$$

denote the covering mapping taking each $x \in \mathbf{S}^p$ to the pair of antipodal points $\{x, -x\} \in \mathbf{RP}^p$. Suppose X is a statistic that is uniformly distributed on $\mathbf{S}^p$. As $\mathcal{A}$ is continuous, it follows that $\mathcal{A}(X)$ is a statistic on $\mathbf{RP}^p$. As will be seen, $\mathcal{A}(X)$ is uniformly distributed on $\mathbf{RP}^p$.

The fact that this is true can be shown by calculating the density function of $\mathcal{A}(X)$ directly. If X has density function f on the sphere $\mathbf{S}^p$, then it can be shown that the density function of $\mathcal{A}(X)$ at the point $\mathcal{A}(x)$ is $f(x) + f(-x)$. The constancy of f on $\mathbf{S}^p$ implies the constancy of $f(x) + f(-x)$, and thereby the uniformity of the distribution of $\mathcal{A}(X)$.

However, an alternative proof using invariance is useful. To prove the result, we first note that an orthogonal transformation of the sphere preserves the property that points are antipodal. That is,

$$h(-x) = -h(x) \tag{4.45}$$

for all transformations $h \in \mathbf{O}(p+1)$. This means that the group $\mathbf{O}(p+1)$ acts on $\mathbf{RP}^p$ as well, the element $h \in \mathbf{O}(p+1)$ taking any pair of antipodal points $\{x, -x\}$ to another pair of antipodal points $\{h(x), -h(x)\}$. At the risk of some confusion, we write

$$h : \mathbf{RP}^p \to \mathbf{RP}^p \tag{4.46}$$

as well as

$$h : \mathbf{S}^p \to \mathbf{S}^p \tag{4.47}$$

letting the context decide the transformation under consideration. With this understanding, we can show that

$$\mathcal{A}[h(x)] = h[\mathcal{A}(x)] \tag{4.48}$$

for every $h \in \mathbf{O}(p+1)$ and every $x \in \mathbf{S}^p$. More compactly, we can write $\mathcal{A} \circ h = h \circ \mathcal{A}$. Equivalently, the functions $\mathcal{A}$ and h can be said to *commute*. The diagram of this looks as follows:

$$\begin{array}{ccc} \mathbf{S}^p & \to & \mathbf{S}^p \\ \downarrow & & \downarrow \\ \mathbf{RP}^p & \to & \mathbf{RP}^p \end{array} \tag{4.49}$$

Now $\mathbf{O}(p+1)$ can be checked to be a group of isometries of $\mathbf{RP}^p$, because the mapping

$$\{x, -x\} \to \{h(x), -h(x)\} \tag{4.50}$$

preserves geodesic distance in $\mathbf{RP}^p$. That $\mathbf{O}(p+1)$ acts transitively on $\mathbf{RP}^p$ follows easily from the fact that it acts transitively on $\mathbf{S}^p$. See Problem 6 at the end of the chapter. In addition, for any Borel set $B \subset \mathbf{RP}^p$,

$$\begin{aligned} \mathcal{P}[\mathcal{A}(X) \in h(B)] &= \mathcal{P}\{X \in \mathcal{A}^{-1}[h(B)]\} \\ &= \mathcal{P}\{X \in h[\mathcal{A}^{-1}(B)]\} \end{aligned} \tag{4.51}$$

The second equality follows from the fact that h and $\mathcal{A}$ commute. However,

$$\begin{aligned} \mathcal{P}\{X \in h[\mathcal{A}^{-1}(B)]\} &= \mathcal{P}[X \in \mathcal{A}^{-1}(B)] \\ &= \mathcal{P}[\mathcal{A}(X) \in B] \end{aligned} \tag{4.52}$$

the first equality following from the invariance of the uniform distribution on the sphere $\mathbf{S}^p$. This demonstrates the invariance. So $\mathcal{A}(X)$ is uniformly distributed on $\mathbf{RP}^p$.

The reader should note that the group $\mathbf{O}(p+1)$ is a little bigger than the group of isometries it induces on $\mathbf{RP}^p$. The transformation $x \to -x$ is an element of $\mathbf{O}(p+1)$ that maps $\{x, -x\}$ back onto itself. Thus it induces the identity transformation on $\mathbf{RP}^p$. This transformation, together with the identity transformation, forms a subgroup of $\mathbf{O}(p+1)$ that is the center of $\mathbf{O}(p+1)$. It is isomorphic to the group $\mathbf{O}(1)$. Thus in group-theoretic terms we can write the group of isometries induced on $\mathbf{RP}^p$ as the factor group $\mathbf{O}(p+1)/\mathbf{O}(1)$.

4.4.3 Example: Isometries of Complex Projective Spaces

Our next example is an extension of the previous case to include the class of complex projective spaces. Suppose that Z is a statistic that is uniformly distributed on the sphere $\mathbf{S}^{2q+1}$, this time understood as the unit sphere about the origin in $\mathbf{C}^{q+1}$. Let

$$\mathcal{O} : \mathbf{S}^{2q+1} \to \mathbf{CP}^q \tag{4.53}$$

be the mapping taking each point z of the sphere $\mathbf{S}^{2q+1}$ into its orbit $\mathcal{O}(z)$, as in Section 2.2.16. We claim that $\mathcal{O}(Z)$ is uniformly distributed on $\mathbf{CP}^q$. The proof of this result parallels the case for $\mathbf{RP}^p$ above, with the exception that the group $\mathbf{U}(q+1)$ must be used to provide the invariance rather than $\mathbf{O}(2q+2)$. The proof goes through in a similar way to that of Section 4.4.2, above. In this case, our commutative diagram becomes

$$\begin{array}{ccc} \mathbf{S}^{2q+1} & \to & \mathbf{S}^{2q+1} \\ \downarrow & & \downarrow \\ \mathbf{CP}^q & \to & \mathbf{CP}^q \end{array} \tag{4.54}$$

The transformations on $\mathbf{CP}^q$ induced by $U(q+1)$ are isometries, the group as a whole acting transitively on $\mathbf{CP}^q$. As above, this follows easily from the fact that $U(q+1)$ acts transitively on $\mathbf{S}^{2q+1}$. See Problem 7.

The reader should note that as in the previous example, the group $\mathbf{U}(q+1)$ is bigger than the group of isometries it induces on $\mathbf{CP}^q$. A subgroup of $\mathbf{U}(q+1)$ determined as its *center* maps orbits $\mathcal{O}(z)$ back to themselves. Such transformations induce the identity transformation on $\mathbf{CP}^q$, and together form a group that is isomorphic to $\mathbf{U}(1)$. In a manner similar to the previous example, in group-theoretic terms we can write the group of isometries induced on $\mathbf{CP}^q$ as the factor group $\mathbf{U}(q+1)/\mathbf{U}(1)$.

4.5 Normal Statistics on Manifolds

4.5.1 Multivariate Normal Distributions

In this section, we give a brief summary of some definitions and results from multivariate normal theory for Euclidean spaces and spheres that we shall need for shape modeling. In keeping with the spirit of Section 4.4, we shall consider these models from the perspective of invariance.

Definition 4.5.1. *Let* $X = (X_1, X_2, ..., X_n)^T$ *be a column vector of random variables. Then the random vector* X *is said to have a* multivariate normal *distribution if it has a density function of the form*

$$f(x) = [2\pi \det(\Gamma)]^{-n/2} \exp\left[-\frac{1}{2}(x-\mu)^T\Gamma^{-1}(x-\mu)\right] \tag{4.55}$$

where $\mu = (\mu_1, \mu_2, ..., \mu_n)^T$ *is a column vector, called the* mean vector, *and* Γ *is an* $(n \times n)$*-dimensional positive definite symmetric matrix, called the* covariance matrix. *In the special case where* $n = 1$, *we simply say that* X *has a normal distribution with mean parameter* μ *and variance parameter* $\Gamma > 0$, *understanding* X, μ, *and* Γ *to be scalars.*

In fact, $\mu = \mathcal{E}(X)$ and $\Gamma = \mathcal{E}(XX^T)$. The entries X_j of a multivariate normal vector X can be shown to be normally distributed random variables.

Suppose that $X \to \Lambda X + a$ is an affine transformation of full rank, where a is an $n \times 1$ column vector. Then it can be shown that $Y = \Lambda X + a$ also has a multivariate normal distribution, with mean vector $\Lambda\mu + a$ and covariance matrix $\Lambda\Gamma\Lambda^T$.

Of particular interest to us here will be the special case where $\Gamma = cI$. In this case, we say that X has a *spherical normal* distribution. The random vector X can be shown to have a spherical normal distribution if and only if the random variables $X_1, X_2, ..., X_n$ are independent normal random variables with common variance parameter. The spherical normal density function is preserved under similarity transformations of $\mathbf{R}^n$. Suppose X is spherical normal and $Y = b\Lambda X + a$, where Λ is an $n \times n$ orthogonal matrix, a is an $n \times 1$ column vector, and b is a positive scalar. Then Y is also spherical normal.

4.5.2 Helmert Transformations

The following class of orthogonal transformations and their matrix representations will be of interest for shape theory. By a *Helmert matrix* of order n we shall understand an $n \times n$ matrix whose first row is a row vector of entries equal to $1/\sqrt{n}$. For $j = 2, ..., n$, the jth row is a row vector whose first $j-1$ entries equal $1/\sqrt{j(j-1)}$, the jth entry being $-\sqrt{(j-1)/j}$, and whose remaining $n-j$ entries are zero. Thus, for example, the Helmert matrix of order 4 is

$$\begin{pmatrix} 1/\sqrt{4} & 1/\sqrt{4} & 1/\sqrt{4} & 1/\sqrt{4} \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 & 0 \\ 1/\sqrt{6} & 1/\sqrt{6} & -2/\sqrt{6} & 0 \\ 1/\sqrt{12} & 1/\sqrt{12} & 1/\sqrt{12} & -3/\sqrt{12} \end{pmatrix} \tag{4.56}$$

The Helmert matrices can be shown to be orthogonal. Now suppose that $X_1, ..., X_n$ are independent identically distributed random variables. Then $X = (X_1, ..., X_n)^T$ has a spherical normal distribution. Suppose also that Λ is a Helmert matrix of order n. As Λ is orthogonal, it follows that $Y = \Lambda X$ is also spherical normal. This implies that $Y_1, Y_2, ..., Y_n$ are independent normal random variables with common variance. With the exception of Y_1, which generally has nozero mean, the other random variables $Y_2, ..., Y_n$ have zero mean, and are therefore identically distributed. The random vector $(Y_2, ..., Y_n)^T$ can be placed in one-to-one correspondence with the vector of residuals $(X_1 - \bar{X}, ..., X_n - \bar{X})$ used to eliminate location information from landmarks in Chapter 1. The former vector can be regarded as an orthogonalized form of the latter, which has a linear constraint on its components.

4.5.3 Projected Normal Statistics on Spheres

As we saw in Chapter 1, projection onto a sphere arises in shape analysis from the removal of scale variables in the reduction to the pre-shape of the data.

Definition 4.5.2. *Suppose that* $X \in \mathbf{R}^n$ *has a spherical normal distribution with mean vector* μ. *With probability one,* X *will be nonzero. Let* $||X||$ *be the norm of* X. *The scaled vector*

$$\theta(X) = \frac{X}{||X||} \tag{4.57}$$

is a point on the unit sphere $\mathbf{S}^{n-1}$ *centered about the origin. Thus* $\theta(X)$ *is the projection of a normally distributed vector onto* $\mathbf{S}^{n-1}$ *and is said to have a* projected normal distribution.

To compute the density function for the projected normal distribution on $\mathbf{S}^{n-1}$, we transform variables, writing any nonzero point $x \in \mathbf{R}^n$ in polar form as (r, θ), where $r = ||x||$ and $\theta = x/||x||$. The pair (r, θ) naturally lies in the product manifold $\mathbf{R}^+ \times \mathbf{S}^{n-1}$. Under the identification of x with (r, θ), the volume element on $\mathbf{R}^n$ decomposes as

$$\begin{aligned} d\mathcal{V}_n(x) &= r^{n-1}\, d\mathcal{V}_{n-1}(\theta)\, d\mathcal{V}_1(r) \\ &= r^{n-1}\, d\mathcal{V}_{n-1}(\theta)\, dr \end{aligned} \tag{4.58}$$

Without loss of generality, we consider the standardized case where the covariance matrix of X is the identity. If this is not the case, then X can be rescaled before projection onto $\mathbf{S}^{n-1}$. Let θ_j be the angle between θ and the x_j-axis. So $x_j = r\, \cos(\theta_j)$. Then the density function with respect to the element $d\mathcal{V}_{n-1}(\theta)\, dr$ is

$$(2\pi)^{-n/2} \exp\left\{ -\frac{1}{2} \sum_{j=1}^{n} [r_j\, \cos(\theta_j) - \mu_j]^2 \right\} r^{n-1} \tag{4.59}$$

Thus the density function of θ with respect to the volume element $d\mathcal{V}_{n-1}(\theta)$ has integral representation

$$(2\pi)^{-n/2} \int_0^\infty \exp\left\{ -\frac{1}{2} \sum_{j=1}^{n} [r\, \cos(\theta_j) - \mu_j]^2 \right\} r^{n-1}\, dr \tag{4.60}$$

For simplicity, let us assume a coordinate system where the components of the mean vector of X have the form $\mu_1 = \nu$ and $\mu_2 = ... = \mu_n = 0$. Then the density function reduces to

$$(2\pi)^{-n/2} \exp(-\nu^2/2) \int_0^\infty \exp\left\{ -\frac{1}{2} \left[r^2 - 2\nu r\, \cos(\theta_1) \right] \right\} r^{n-1}\, dr \tag{4.61}$$

Following the notation of [70], we define

$$\mathcal{I}_k(u) = \int_0^\infty r^k \exp\left(-\frac{r^2}{2} \right) e^{ru}\, dr \tag{4.62}$$

for $k = 0, 1, 2, ...$ Then the density function can be written as

$$(2\pi)^{-n/2} \exp(-\nu^2/2)\, \mathcal{I}_{n-1}[\nu\, \cos(\theta_1)] \tag{4.63}$$

The special functions can be computed recursively using the formulas

$$\mathcal{I}_0(u) = \Phi(u)/\phi(u) \tag{4.64}$$

$$\mathcal{I}_1(u) = 1 + u\Phi(u)/\phi(u) \tag{4.65}$$

and

$$\mathcal{I}_{k+1}(u) = u\mathcal{I}_k(u) + k\mathcal{I}_{k-1}(u) \tag{4.66}$$

where ϕ is the density function for a standard normal random variable and Φ is the distribution function of a standard normal random variable. For a more general sequence $\mu = (\mu_1, \mu_2, ..., \mu_n)^T$ we replace θ_1 in formula (4.60) by the angle made between the vector θ and the vector μ. Problem 9 asks the reader to investigate the density function when $\nu = 0$ and $\nu \to \infty$. It can be seen that when $\nu = 0$, the distribution on $\mathbf{S}^{n-1}$ reduces to the uniform distribution, the density having a value that is the reciprocal of the volume of the sphere.

To see the relevance of this distribution to shape analysis, and the distribution of pre-shape statistics in particular, consider a set of independent random vectors

$$X_1,\ X_2,\ ...,\ X_n \in \mathbf{R}^p \tag{4.67}$$

each of which has a spherical normal distribution. Suppose also that the random vectors have common $p \times p$ covariance matrix cI, say, but possibly different mean vectors

$$\mu_1,\ \mu_2,\ ...,\ \mu_n \in \mathbf{R}^p \tag{4.68}$$

We construct a $p \times n$ matrix X whose jth column is X_j. The matrix X can also be regarded as a vector in $\mathbf{R}^{pn}$ whose components are the entries X_{jk}. Let Λ be the $(n-1) \times n$ matrix made by deleting the first row of the Helmert matrix of order n. Then the $p \times (n-1)$ matrix

$$Y = X\Lambda^T \tag{4.69}$$

has a spherical normal distribution in $\mathbf{R}^{p(n-1)}$. The columns of the matrix Y are independent and identically distributed normal random variables containing the information from X with location information removed. We find the pre-shape of X by rescaling Y so that the sum of squares of its $p(n-1)$ components equals one. This is equivalent to projecting Y as a vector in $\mathbf{R}^{p(n-1)}$ onto the unit sphere $\mathbf{S}^{np-p-1}$ around the origin in $\mathbf{R}^{p(n-1)}$. Let

$$\tau = \frac{Y}{||Y||} \in \mathbf{S}^{np-p-1} \tag{4.70}$$

be the pre-shape of X. Then τ is seen to have a projected normal distribution on $\mathbf{S}^{np-p-1}$.

The reader should note that although for the case $p = 2$ the coordinate representation of pre-shape information given here differs from the representation of Chapter 1, the two representations are isometric. In the earlier chapters, the sphere $\mathbf{S}_*^{2n-3}$ was embedded as the unit sphere in a $(2n-2)$-dimensional subspace of $\mathbf{R}^{2n}$. However, here we use the sphere $\mathbf{S}^{2n-3}$, which is the unit sphere in $\mathbf{R}^{2n-2}$. The reader can demonstrate

the isometry by checking that the formulas for geodesic distances between pre-shapes are identical for the two representations.

To go from pre-shape statistics to shape statistics involves one additional integration. In view of the complexity of the projected normal density function, it might be questioned as to whether the associated shape density can be written in manageable form. It was the conclusion of Mardia and Dryden [116] that for $p = 2$, this distribution is not only simple in form but can be written in terms of the geodesic distance between shapes on the shape manifolds $\Sigma_2^n \cong \mathbf{CP}^{n-2}$. We will consider this in detail in the next chapter.

4.6 Binomial and Poisson Processes

We now turn to some examples of *point processes* that will be important for our development of the statistics of shape.

4.6.1 Uniform Distributions on Open Sets

We have noted that if $\mathbf{M}^p$ has infinite volume, then no uniform distribution exists for it. However, it is possible to impose a uniform distribution on regions of the manifold and to associate with these uniform distributions a limited form of invariance. Let B be an open set of $\mathbf{M}^p$ for which $\mathcal{V}_p(B) < \infty$. A continuous statistic $X \in \mathbf{M}^p$ is said to be *uniformly distributed* on B if its density function has the form

$$f(x) = \begin{cases} \frac{1}{\mathcal{V}_p(B)} & x \in B \\ 0 & x \notin B \end{cases} \tag{4.71}$$

A particular case of interest to us will be that for which $\mathbf{M}^p = \mathbf{R}^p$ and B is a convex subset of $\mathbf{R}^p$. We shall explore this in greater detail in the next chapter, where the formulas of integral geometry will be directly applicable to the shapes of points uniformly and independently generated in convex sets.

4.6.2 Binomial Processes

For any event A the *indicator random variable*

$$1_A : \mathcal{S} \to \mathbf{R} \tag{4.72}$$

for the event A is defined by

$$1_A(s) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A \end{cases} \tag{4.73}$$

Let B be an open subset of a Riemannian manifold $\mathbf{M}^p$ such that $\mathcal{V}_p(B) < \infty$. Now, suppose that B_0 is an open subset of B. Let $X_1, X_2, ..., X_n$ be n independent statistics in $\mathbf{M}^p$ that are uniformly distributed over the open set B. We define the nonnegative integer-valued random variable N to be

$$N = \sum_{j=1}^{n} 1_{(X_j \in B_0)} \tag{4.74}$$

The random variable N is well known to have a *binomial* distribution, with the property that

$$\mathcal{P}(N = k) = \binom{n}{k} \left[\frac{\mathcal{V}_p(B_0)}{\mathcal{V}_p(B)}\right]^k \left[1 - \frac{\mathcal{V}_p(B_0)}{\mathcal{V}_p(B)}\right]^{n-k} \tag{4.75}$$

For this reason, an indepedent and uniform scattering of a fixed number of statistics over an open set B is often called a *binomial process.* For a binomial process, the *random set of points* in $\mathbf{M}^p$ so generated is the principal object of interest, the ordering of the points being of secondary interest.

4.6.3 Example: Binomial Processes of Lines

Line processes provide a mathematical model for the random scattering of straight lines. Examples include the cracking of surfaces and the tracks left by small particles scattered randomly and homogeneously through a region with random orientations to their velocities.

Suppose X is a line in the plane that passes through a bounded convex set $A \subset \mathbf{R}^2$. Let us suppose that X is directed. We can think of this as providing a rule as to which is the left side of the line. The rule is arbitrary but must be consistently imposed all the way along the line. The set of all directed lines in the plane can be placed in a natural 1-1 correspondence with the points of the cylinder $\mathbf{R} \times \mathbf{S}^1$, as shown in Figure 4.2.

The group $\mathbf{Euc}(2)$ of Euclidean motions of $\mathbf{R}^2$ maps lines to lines. Thus $\mathbf{Euc}(2)$ can be said to act upon the manifold $\mathbf{R} \times \mathbf{S}^1$. A Euclidean motion takes a line with coordinates (r, θ) to its image with coordinates (r', θ'). Among these transformations are the translations on $\mathbf{R} \times \mathbf{S}^1$ mapping

$$(r,\ \theta) \to (r + s,\ \theta) \tag{4.76}$$

and

$$(r,\ \theta) \to (r,\ \theta + \phi) \tag{4.77}$$

where angle addition is performed modulo 2π. So any measure on the space $\mathbf{R} \times \mathbf{S}^1$ of directed lines that is invariant under the Euclidean motions of $\mathbf{R}^2$ will be invariant under the translations (4.76) and (4.77) in particular. Now

$$d\mathcal{V}_2(r, \theta) = dr\ d\theta \tag{4.78}$$

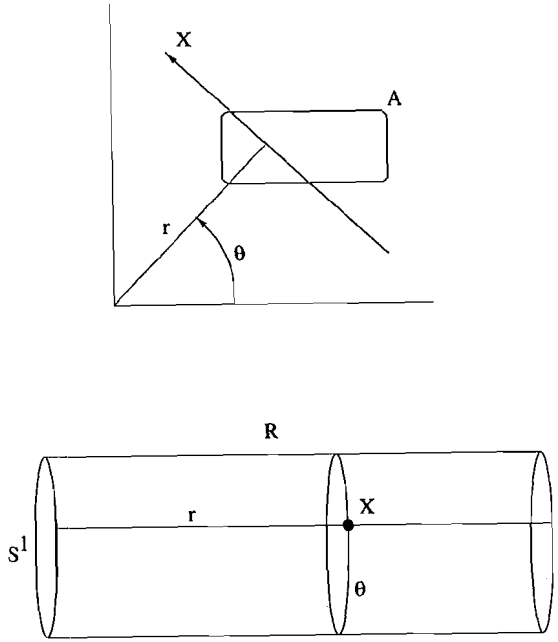

FIGURE 4.2. *A directed line X in the plane represented as a point on a cylinder. The angle θ is defined so that $\theta + \pi$ is the counterclockwise angle from the horizontal axis to the line X. The real number r is defined as the signed distance from the origin to X, the sign being chosen so that $r > 0$ when the origin is on the side of the line as shown and $r < 0$ when the origin is on the other side.*

is invariant under these translations. Moreover, the group of translations generated by (4.76) and (4.77) acts transitively on $\mathbf{R} \times \mathbf{S}^1$. Therefore, we can show in a manner similar to Proposition 4.4.2 that the measure in (4.78) is the unique invariant measure up to an arbitrary scalar multiple. See, for example, Santalo [147, pp. 27-30] for the details of this.

Using $d\mathcal{V}_2(r,\theta)$ we can construct a binomial process of directed lines passing through A. Let B be the set of all (r,θ) such that the line with coordinates (r,θ) passes through A. As $\mathcal{V}_2(B) < \infty$, we can construct a collection of independent random lines $X_1, ..., X_n$ that are uniformly distributed in B. Having constructed such a binomial process of lines, we can make the lines undirected by erasing the arrows.

We can also consider the *shape* of any configuration of lines generated by a binomial process. Like $\mathbf{Euc}(2)$, the group $\mathbf{Sim}(2)$ maps lines to lines, and therefore can be regarded as acting upon the manifold $\mathbf{R} \times \mathbf{S}^1$. The shape of any configuration of lines can be defined as the total information in the set $X_1, ..., X_n$ that is invariant under the action of this group. The situation is analogous to our definition of the shape of a set of landmarks in Chapter 1. However, instead of the direct action of $\mathbf{Sim}(2)$ on $\mathbf{R}^2$, the action of $\mathbf{Sim}(2)$ on the space of lines is induced from the action on the Euclidean space in which the lines reside. In this and other contexts, the study of the shapes of configurations of geometric objects becomes the study of invariants of configurations of points in manifolds. See Carne [38].

The assignment of directions to lines is a mathematical convenience that allows us to put a simple coordinate system on the space of lines. Undirected lines are more physically natural for the purposes of modeling many physical processes involving line data. We obtain an undirected line by throwing away the direction, so to speak. More formally, we can define an undirected line as a *pair of directed lines* having coordinates of the form $\{(r,\theta),(-r,\theta+\pi)\}$. The manifold of undirected lines in the plane is thereby seen to be the space of such pairs of points in the manifold $\mathbf{R} \times \mathbf{S}^1$. Wilfrid Kendall has noted, in a private communication, that the space of undirected lines is homeomorphic to the Moebius strip defined in Problem 3 of Chapter 2. Problem 11 at the end of this chapter sketches the steps necessary to prove this fact.

4.6.4 Poisson Processes

Let us return to the general binomial process in some open set $B \subset \mathbf{M}^p$. Suppose that $\mathcal{V}_p(\mathbf{M}^p) = \infty$. A limiting case is obtained when B expands to encompass all of $\mathbf{M}^p$. As noted earlier, there is no uniform distribution on all $\mathbf{M}^p$ in this case. However, there is a nondegenerate limiting form for the binomial process. Consider a nested sequence of open sets

$$B = B_1 \subset B_2 \subset B_3 \subset \ldots \tag{4.79}$$

so that $\mathbf{M}^p = \cup_{j=1}^{\infty} B_j$. Suppose that as $B_j \nearrow \mathbf{R}^p$ we select a sequence of positive integers

$$n = n_1 < n_2 < n_3 < \ldots \tag{4.80}$$

such that

$$\frac{n_j}{\mathcal{V}_p(B_j)} \rightarrow \rho > 0 \tag{4.81}$$

as $j \rightarrow \infty$. For each j we construct a binomial process of n_j points in B_j, and for each we let N_j be the number of such points falling into B_0. Then

$$\mathcal{P}(N_j = k) \rightarrow \frac{[\rho\mathcal{V}_p(B_0)]^k \exp[-\rho\mathcal{V}_p(B_0)]}{k!} \tag{4.82}$$

for $k = 0, 1, 2, \ldots$, which is the well known formula for the Poisson distribution. The limiting form of the binomial process is called the *Poisson process of intensity* ρ.

See Problem 10 for the derivation of the Poisson formula. We can think of the Poisson process intuitively as a uniform scattering of infinitely many points throughout the entire space, so that on average, ρ points fall into a region of unit volume.

We shall need to refer to given points of $\mathbf{M}^p$ that are distinct from the random set of points of the Poisson process. Some terminology helps to keep these distinct. We shall henceforth refer to the random points of a point process as *particles* and shall reserve the term *points* for given elements of $\mathbf{M}^p$ that have a fixed location. However, following traditional terminology we shall continue to refer to a random scattering of particles as a point process.

The following definition helps to formalize the construction of the Poisson process by characterizing it in terms of its properties.

Definition 4.6.1. *We define a* point process *on* $\mathbf{M}^p$ *to be a random countable set of particles* $C \subset \mathbf{M}^p$. *A point process that has finite intersection with any bounded subset of* $\mathbf{M}^p$ *is said to be* locally finite.

Henceforth, we shall assume that all point processes under consideration are locally finite. For any bounded Borel subset B, let $N(B)$ be the cardinality of the set $B \cap C$. Among the class of locally finite point processes are those satisfying certain uniformity conditions as given in the next definition.

Definition 4.6.2. *A point process is said to be* volume-preserving *if the* $N(B_1)$ *and* $N(B_2)$ *are identically distributed whenever* $\mathcal{V}_p(B_1) = \mathcal{V}_p(B_2)$.

Among the class of volume-preserving point processes are the Poisson point processes that we described above as the limit of binomial point processes. We have the following definition:

Definition 4.6.3. *A volume-preserving point process on* $\mathbf{M}^p$ *is said to be a* homogeneous Poisson process *if it satisfies the following postulates.*

Postulate 1. *If* $0 < \mathcal{V}_p(B) < \infty$ *then* $0 < \mathcal{P}[N(B) = 0] < 1$. *Moreover, as* $\mathcal{V}_p(B) \rightarrow 0$ *we have* $\mathcal{P}[N(B) = 0] \rightarrow 1$.

Postulate 2. *If the sets* $B_1, B_2, \ldots, B_m$, $m \geq 2$, *are disjoint subsets of* $\mathbf{R}^p$ *then* $N(B_1), N(B_2), \ldots, N(B_m)$ *are independent.*

Postulate 3. *As* $\mathcal{V}_p(B) \rightarrow 0$ *we have*

$$\frac{\mathcal{P}[N(B) > 1]}{\mathcal{P}[N(B) = 1]} \rightarrow 0 \tag{4.83}$$

For a Poisson process (i.e., a point process satisfying the above), it can be shown that there exists a unique intensity parameter $\rho > 0$ such that for every open set $B \subset \mathbf{M}^p$ the random variable $N(B)$ has a Poisson distribution with parameter $\rho\mathcal{V}_p(B)$ as given in formula (4.82) above.

4.7 Poisson Processes in Euclidean Spaces

In this section, we will summarize some of the properties of Poisson processes in p-dimensional Euclidean space.

The strong invariance properties of the Poisson process make it a particularly useful model for generating random shapes. The invariance of shape statistics under Euclidean motions is compatible with the motion-invariance of the Poisson process, making probability calculations easier. Although the assumptions of the Poisson process will not typically be realized in their exact form in applications, the model has been found to be useful for simulating a variety of phenomena involving random scatterings of particles.

4.7.1 Nearest Neighbors in a Poisson Process

A number of geometric properties of the particles of a Poisson process (PP) hold with probability one. For example, a *nearest neighbor* of a point $x \in \mathbf{R}^p$ is a particle X of the PP that has minimum distance from x among all such particles of the PP. It can be seen that with probability one every particle of the PP has a unique nearest neighbor. This can be generalized to the second nearest neighbor, and so on. In general, the kth nearest neighbor of a point x is a particle X such that there are exactly $k-1$ particles of the PP strictly closer to x than X. Again, with

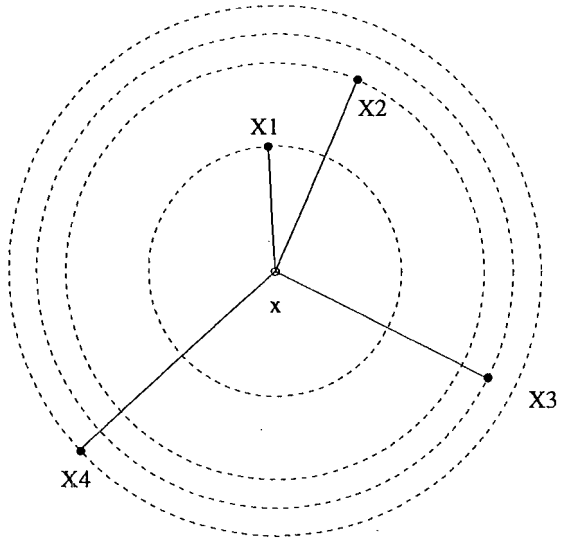

FIGURE 4.3. *The nearest neighbors in a point process. Given any fixed point x, with probability one no two particles of a Poisson process will be at the same distance from x. The closest particle to x, labeled X_1 in the diagram, is called the nearest neighbor of x, and is on a sphere centered about x that has no particles in its interior. In general, the kth nearest neighbor of x, labeled X_k, is on a sphere centered about x with $k-1$ particles in its interior. This fact implies that the distance to the kth nearest neighbor of x will be greater than $s > 0$ if and only if there are $k-1$ or fewer particles in the interior of the sphere of radius s centered at x. This property can be used to calculate the distribution function of the distance to the kth nearest neighbor of x.*

probability one the kth nearest neighbor is unique. Note that the point x can itself be a particle of the PP if desired, with the understanding that it is not its own nearest neighbor. See Figure 4.3.

4.7.2 The Nonsphericity Property of the PP

A set of $p+1$ particles of a PP in $\mathbf{R}^p$ are said to be in *general position* if the convex hull of the particles has an nonempty interior (or equivalently, contains an open subset of $\mathbf{R}^p$). Thus three particles are in general position in $\mathbf{R}^2$ provided they are not collinear. Four particles are in general position in $\mathbf{R}^3$ provided no three are collinear and the four particles are not coplanar, etc. It can be shown that with probability one for a PP in $\mathbf{R}^p$, all sets of $p+1$ particles of the PP are simultaneously in general position.

A property related to this is the *nonsphericity property*. Through any set of $p+1$ particles, which with probability one are in general position, a unique $(p-1)$-dimensional sphere may be drawn, for $p \geq 2$. The nonsphericity property of the PP in $\mathbf{R}^p$ states that with probability one such a sphere passing through $p+1$ particles will meet no other particles of the PP. Particles may be found in the p-dimensional ball bounded by the sphere, but not on the spherical boundary itself.

4.7.3 The Delaunay Tessellation

There are many mechanisms for selecting finitely many points from a Poisson process for the purpose of generating shape distributions. One of the most important is the *Delaunay tessellation*, which decomposes $\mathbf{R}^p$ into p-dimensional simplexes (i.e., triangles, tetrahedra, etc.) that are nonoverlapping in the sense that any two simplexes can share at most a common $(p-1)$-dimensional face. The vertices of these simplexes are the points of the Poisson process itself.

To construct the Delaunay tessellation, we take advantage of the nonsphericity property of the Poisson process. We have the following definition:

Definition 4.7.1. *Let $X_1, ..., X_{p+1}$ be a set of $p+1$ particles from some PP in $\mathbf{R}^p$. Let*

$$\Delta = \Delta(X_1, X_2, ..., X_{p+1}) \tag{4.84}$$

be the p-dimensional simplex whose vertices are these $p+1$ particles. We say that Δ is a Delaunay simplex of the PP *provided that the $(p-1)$-dimensional sphere passing through $X_1, X_2, ..., X_{p+1}$ encloses no particle of the process within its interior.*

For the planar case the simplexes are triangles, and so we speak of the Delaunay triangles. Figure 4.4 shows the Delaunay triangles associated with a particular arrangement of particles in $\mathbf{R}^2$.

We now have the following:

Definition 4.7.2. *A collection $\{\Delta_j\}$ of countably many p-dimensional simplexes in $\mathbf{R}^p$ is said to be a* tessellation *if $\bigcup_j \Delta_j = \mathbf{R}^p$ and if in addition, the interiors of the sets Δ_j and Δ_k have empty intersection whenever $j \neq k$.*

We state the following proposition without proof. The reader who is interested in the details of the proof should consult [128].

Proposition 4.7.3. *With probability one the Delaunay simplexes of a PP in $\mathbf{R}^p$ form a tessellation.*

In fact, a stronger statement can be made about the tessellation. Any two Delaunay simplexes with nonempty intersection will share a face in

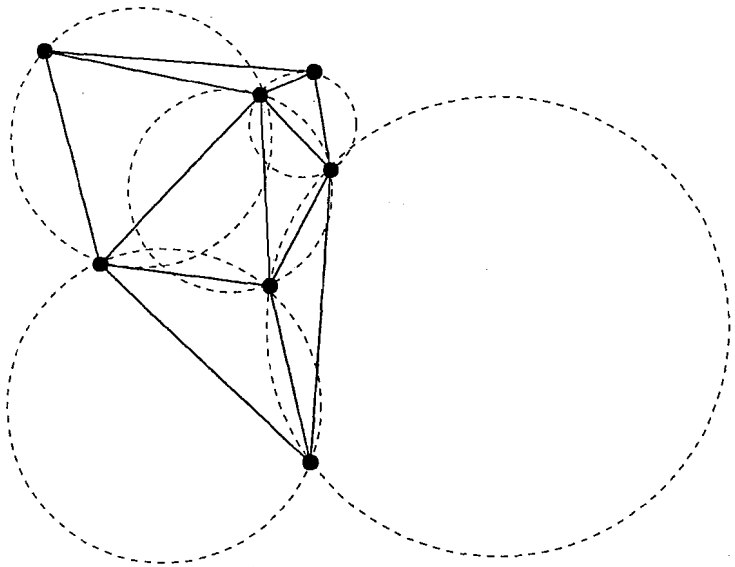

FIGURE 4.4. *Delaunay triangles in the plane. With probability one the particles of a PP satisfy the nonsphericity property. Therefore the circumcircle through any three particles will meet no other particle of the PP. However, many such circumcircles will have particles in their interiors. Those triangles whose circumcircles do not have any particles within their interiors are called Delaunay triangles. If the PP is generated to fill the plane, then the Delaunay triangles form a tessellation of the plane.*

common, the face being itself a simplex, of dimension $p-1$.

The Delaunay tessellation of a PP provides a stochastic mechanism for generating *simplex shapes*. We shall consider this shape distribution in the next chapter. At this stage, we consider the distribution of the geometric characteristics of a typical random Delaunay simplex. The precise nature of this distribution depends of course on what is meant by a typical simplex. Along with the randomness of the PP that generates the tessellation, the idea of a random Delaunay simplex presupposes a stochastic mechanism for selecting a cell from the tessellation.

4.7.4 Pre-Size-and-Shape Distribution of Delaunay Simplexes

Consider a binomial process of $n \geq p+1$ independent particles $X_1, ..., X_n$ that are uniformly distributed in an open convex subset A of $\mathbf{R}^p$. Given that the simplex Δ with vertices $X_1, ..., X_{p+1}$, say, forms a Delaunay simplex, what is the distribution of the joint geometric characteristics of the simplex? For Δ to be a Delaunay simplex, we require that none of the points $X_{p+2}, ..., X_n$ fall inside the circumsphere through $X_1, ..., X_{p+1}$. Let

$$r = r(X_1, X_2, ..., X_{p+1}) \tag{4.85}$$

be the radius of the circumsphere through $X_1, ..., X_{p+1}$. For large n, the case of particular importance here, the volume enclosed by the circumsphere will be small compared to the volume of A, because for other cases a particle will lie with high probability in the ball enclosed by the circumsphere. It follows from this that in the conditional case that Δ is a Delaunay simplex, with high probability as $n \to \infty$ the circumsphere will lie entirely within A. For the probability calculations that follow we shall restrict to those cases where the circumsphere lies entirely within the interior of A. Given Δ, the expected number N of particles among $X_{p+1}, ..., X_n$ that fall into the interior of the circumsphere through Δ is

$$\begin{aligned} \mathcal{E}(N) &= [n-(p+1)]\, \mathcal{P}(X \in \Delta^\circ) \\ &= [n-(p+1)]\, r^p\, \kappa_p / \mathcal{V}_p(A) \end{aligned} \tag{4.86}$$

where X is any of the particles $X_{p+1}, ..., X_n$ and κ_p is the volume of the unit ball in $\mathbf{R}^p$. Given Δ, the number of particles $X_{p+2}, ..., X_n$ that fall inside this circumsphere has a binomial distribution. So the Poïsson approximation for the probability that no particles among $X_{p+1}, ..., X_n$ fall inside the circumsphere is

$$\exp[-\mathcal{E}(N)] = \exp\left\{-\frac{[n-(p+1)]\, r^p\, \kappa_p}{\mathcal{V}_p(A)}\right\} \tag{4.87}$$

The volume κ_p evaluates by integration to

$$\kappa_p = \frac{2\,\pi^{p/2}}{p\,\mathrm{Gam}(p/2)} \tag{4.88}$$

where Gam(.) is the usual Gamma function. Letting n and $\mathcal{V}_p(A)$ go to infinity, so that

$$\frac{n}{\mathcal{V}_p(A)} \rightarrow \rho \tag{4.89}$$

we obtain the limiting form of this probability for the PP of intensity ρ, namely

$$\mathcal{P}(N=0) \rightarrow \exp\left[-\rho\,\kappa_p\,r^p(x_1,x_2,...,x_{p+1})\right] \tag{4.90}$$

Now, the density function for $X_1,...,X_{p+1}$ uniformly distributed in A is given by

$$[\mathcal{V}_p(A)]^{-(p+1)} \prod_{j=1}^{p+1} 1_{(x_j \in A)} \tag{4.91}$$

Let us introduce a transformation of variables

$$X_1,...,X_{p+1} \leftrightarrow Y_1,...,Y_p,X_{p+1} \tag{4.92}$$

where

$$Y_j = X_j - X_{p+1} \tag{4.93}$$

Using transformation of variables techniques, we find the density function of $Y_1,...,Y_p,X_{p+1}$ to be

$$\left[\frac{1_{(x_{p+1}\in A)}}{\mathcal{V}_p(A)}\right] \times [\mathcal{V}_p(A)]^{-p} \prod_{j=1}^{p} 1_{(y_j+x_{p+1}\in A)} \tag{4.94}$$

the Jacobian being one. Multiplying formula (4.87) and formula (4.94) yields an expression *proportional to* the density function for the vertices, conditionally on Δ being a Delaunay simplex. So the joint density of the coordinates $y_1,...,y_p$ and x_{p+1}, given that Δ is a Delaunay simplex, is

$$c_p[\mathcal{V}_p(A)]^{-p} \times \left[\frac{1_{(x_{p+1}\in A)}}{\mathcal{V}_p(A)}\right] \times \exp\left[-\rho\,\kappa_p\,r^p\right] \prod_{j=1}^{p} 1_{(y_j+x_{p+1}\in A)} \tag{4.95}$$

where c_p is the constant of proportionality.

The constant of proportionality can be found by integrating this expression over all its variables. If c_p is omitted, (4.95) integrates to the unconditional probability that $X_1,X_2,...,X_{p+1}$ form a Delaunay simplex. Thus to make this expression into a density function in its variables, we must divide by this probability. It is not hard to see that in the limit as n and $\mathcal{V}_p(A)$ go to infinity, the number of Delaunay simplexes in A will be proportional to the number of particles in A, which in turn is proportional to $\rho\mathcal{V}_p(A)$. However, the total number of subsets of $p+1$ particles is proportional to $[\rho\mathcal{V}_p(A)]^{p+1}$. Therefore, the probability that Δ is a Delaunay simplex is proportional to the ratio of these, which is $[\rho\mathcal{V}_p(A)]^{-p}$. In turn, c^p will be proportional to the reciprocal of this quantity. So we can write

$$c_p = a_p\,[\rho\mathcal{V}_p(A)]^p \tag{4.96}$$

Note that the radius of the circumcircle through Δ is not a function of the location of Δ. Therefore, we can write

$$r = r(y_1,\ y_2,\ ...,\ y_p) \tag{4.97}$$

In the limit, as A expands to encompass $\mathbf{R}^p$, the effect of the boundary from the indicator functions

$$1_{(y_j+x_{p+1}\in A)} \tag{4.98}$$

becomes negligible except when x_{p+1} is close to the boundary of A, an event of small probability. Thus we obtain the limiting form for the density of the pre-size-and-shape coordinates of Δ to be

$$\left[\frac{1_{(x_{p+1}\in A)}}{\mathcal{V}_p(A)}\right] \times a_p\,\rho^p \exp\left[-\rho\,\kappa_p\,r^p(y_1,...,y_p)\right] \tag{4.99}$$

This density can be seen to factorize in its limiting form into two components, the first being the density for X_{p+1} and the second *a fortiori* being the density for $Y_1,Y_2,...,Y_p$. Thus the marginal density of $Y_1,Y_2,...,Y_p$ has the form

$$f(y_1,y_2,...,y_p) = a_p\,\rho^p \exp\left[-\rho\,\kappa_p\,r^p(y_1,...,y_p)\right] \tag{4.100}$$

We shall examine this distribution in greater detail in the next chapter, where it will be seen to factorize into scale, orientation, and shape components.

4.8 Notes

For background and details on measure-theoretic probability, the reader is referred to the book by Chung [43] and particularly to the first three chapters. Many of the basic methods of multivariable calculus applied to probability theory can be found in the standard textbook treatments. See [82], for example.

The theory of invariance can be developed in terms of invariant measures on compact groups. When the group of isometries on a manifold is

compact, as will typically be the case for the examples under consideration, the left and right invariant measures on the group will coincide. If an invariant measure can be normalized to a probability measure, the action of the group can be interpreted as inducing random isometries on the manifold. For more on the theory of probability on groups, see the books by Parthasarathy [132] and Heyer [80]. Invariance also plays an important role in geometric probability and stochastic geometry. See [147] for results in this area, including some stochastic geometry on manifolds of constant curvature. The theory of geometric measures on manifolds and minimal surfaces has an extensive literature as well. An excellent introduction to this subject is the book by Morgan [121].

For more on the theory and applications of directional statistics, see [55], [56], and [113]. Directional data commonly arise in applications in the geosciences and meteorology, to name two areas. The real projective plane can be considered the natural space for *axial data*, that is, data in which an individual observation consists of an axis, or line, through the origin in $\mathbf{R}^3$. Data of this kind arise in astronomy, for example, where the orientation of orbital planes of comets or other bodies can be represented by axes normal to the plane.

The Delaunay tessellation has its origins in the work of Voronoi and Delone. See [119] and [128]. The Delaunay tessellation can be regarded as the dual concept to the *Voronoi tessellation*. Suppose we divide up space as follows: Given the positions of particles of a point process, we assign each point in space to the nearest particle. This tessellates space into polytopes called the cells of the Voronoi tessellation.

The Delaunay tessellation bears the following relation to the Voronoi tessellation: Two particles of the point process are vertices of a common Delaunay simplex if and only if their corresponding Voronoi cells share a common face. The Delaunay simplexes can then be put into one-to-one correspondence with the points that are vertices of a Voronoi cell. Voronoi tessellations and the dual Delaunay tessellations have been applied in a number of fields including crystallography (using lattices of particles) and geology, where Voronoi cells are interpreted as area of influence polygons. For mathematical purposes, these methods have also provided tools for spatial interpolation of real valued functions defined on some general-dimensional Euclidean space. For more on these and other applications, see the references in [128].

The Poisson model for Delaunay simplexes is a simple stochastic model for the haphazard simplicial decomposition of space. By contrast, particles that lie, or tend to lie, on a lattice will form more regular simplexes, with less variation in their internal angles. Thus the statistics of the geometric characteristics of Delaunay simplexes can be used as test statistics for point process hypotheses. An example of this is the central place hypothesis in geography, for which see [118].

4.9 Problems

1. Let $\mathcal{S}$ be the set of natural numbers. For any subset $A \subset \mathcal{S}$ and for any $n \in \mathcal{S}$ we define a function $N_n(A)$ to be the number of elements of A that are $\leq n$. Then define

$$\mathcal{P}(A) = \lim_{n\to\infty} N_n(A)/n \tag{4.101}$$

wherever this limit exists. Let $\mathcal{F}$ be the set of all $A \subset \mathcal{S}$ such that $\mathcal{P}(A)$ exists. Is $(\mathcal{S}, \mathcal{F}, \mathcal{P})$ a probability space? Justify your answer.

2. Let $\mathcal{G}$ be the class of all subsets of $\mathbf{R}$ of the form $(-\infty, x]$ where $x \in \mathbf{R}$. Show that the sigma-field generated by $\mathcal{G}$ is the Borel sigma-field of $\mathbf{R}$. Hint: to prove this, show that every set of the form $(-\infty, x]$ is a Borel set and that every open set of $\mathbf{R}$ is in the sigma-field generated by $\mathcal{G}$. From this the result follows. Why?

3. Let $X_1, X_2, ..., X_n$ be independent continuous random variables with density functions $f_1, f_2, ..., f_n$ and distribution functions $F_1, F_2, ..., F_n$, respectively. Let $Y = \max(X_1, ..., X_n)$. Find the density function of Y in terms of $f_1, f_2, ..., f_n, F_1, F_2, ..., F_n$. Hint: find the distribution function of Y first.

4. Verify formula (4.21) from Section 4.2.

5. Using polar coordinates r, θ for the plane, find the formula for the metric tensor g assuming the usual inner product between vectors. Use this formula for g to show that in terms of polar coordinates we can write $d\mathcal{V}_2 = r\ dr\ d\theta$.

6. Prove that $\mathbf{O}(n)$ acts transitively on the sphere $\mathbf{S}^{n-1}$. Hint: Suppose x and y are elements of $\mathbf{S}^{n-1}$ and we wish to find an orthogonal transformation mapping x to y. Let $x = x_1$, and find points $x_2, ..., x_n$ such that the matrix (x_{jk}) whose jth row is x_j is an orthogonal matrix. This can be accomplished by choosing $x_1, ..., x_n$ to be orthogonal vectors of unit length, which is equivalent to lying on $\mathbf{S}^{n-1}$. Let (y_{jk}) be a similar orthogonal matrix for $y = y_1$. Construct the required orthogonal transformation from (x_{jk}) and (y_{jk}).

7. Prove that $\mathbf{U}(m)$ acts transitively on the sphere $\mathbf{S}^{2m-1}$.

8. Find the formula for the geometric measure $d\mathcal{V}_2$ for the Poincaré Disk of Section 2.2.17 in terms of polar coordinates. Justify your answer.

9. Prove that the formula for the density of the projected normal distribution reduces to the uniform density when $\nu = 0$. Evaluate this density. What happens when $\nu \to \infty$.

10. Let $0 < a < 1$ and let n be a positive integer. Show that as $a \to 0$ and $n \to \infty$ such that $na \to \mu$ the binomial probability

$$\binom{n}{x} a^x (1-a)^{n-x} \tag{4.102}$$

converges to the Poisson probability

$$\frac{\mu^x \exp(-\mu)}{x!} \tag{4.103}$$

11. In Section 4.6.3, we noted that the space of undirected lines in the plane is homeomorphic to the Moebius strip. In this problem, we shall go through the steps to prove this fact.

(a) Any undirected line can be directed in two possible ways. If one of those directed lines has coordinates (r, θ), show that the other line has coordinates $(-r, \theta + \pi)$, where summation of angles is performed modulo 2π.

(b) We can embed the cylinder of directed lines in $\mathbf{R}^3$ as

$$\{(x, y, z) : y^2 + z^2 = 1\} \tag{4.104}$$

which is homeomorphic to $\mathbf{R} \times \mathbf{S}^1$. In terms of these three-dimensional coordinates, plot the points

$$\{(r, \cos(\theta), \sin(\theta)),\ (-r, \cos(\theta + \pi), \sin(\theta + \pi))\} \tag{4.105}$$

and note that the line in $\mathbf{R}^3$ that passes through these two points passes through the origin.

(c) Argue that the space of undirected lines in the plane is homeomorphic to the space of lines constructed in part (b). This space is *not* the projective plane $\mathbf{RP}^2$ because there is a line through the origin missing. Which one is it?

(d) From part (c) above, we conclude that the space of undirected lines is homeomorphic to the projective plane $\mathbf{RP}^2$ with one point removed. Use Problem 4 from Chapter 2 to argue that such a space is homeomorphic to the Moebius strip. (Hint: removing a point from a manifold is topologically equivalent to removing a closed disk. Why is this?)

5

Distributions of Random Shapes

5.1 Landmarks from the Spherical Normal: IID Case

We are now in a position to state and prove a central result, due to Kendall [90], for the induced distribution of the shapes of planar landmarks generated by an IID spherical normal model.

Proposition 5.1.1. *Let $X_1, X_2, ..., X_n$, $n \geq 3$, be independent and identically distributed spherical normal variables in $\mathbf{R}^2$. Let*

$$\sigma = \sigma(x_1, ..., x_n) \tag{5.1}$$

be the shape representation of the points as an element of $\Sigma_2^n \cong \mathbf{CP}^{n-2}$. Then σ has a uniform distribution on Σ_2^n.

Proof. This result now follows directly from two results in Chapter 4. From the remarks at the end of Section 4.5.3, we note that the pre-shape $\tau \in \mathbf{S}^{2n-3}$ has a projected normal distribution. The density function for this distribution in given in formula (4.63) with $\nu = 0$. From Problem 9 of Chapter 4, we conclude that this distribution is uniform on $\mathbf{S}^{2n-3}$. The result then follows immediately by applying Section 4.4.3 using $q = n - 2$. Q.E.D.

As a special case, we note that the shape of a random triangle of spherical

normal variables is uniformly distributed on the sphere $\mathbf{S}^2(1/2)$. A simple geometric application is the following:

Corollary 5.1.2. *Under the conditions of Proposition 5.1.1 above, for* $n = 3$ *the probability that* $X_1X_2X_3$ *forms an acute triangle is* $1/4$.

Proof. Let us use the representation of $\mathbf{S}^2(1/2)$ in $\mathbf{R}^3$ given in formula (3.6). We begin by calculating the probability that the angle at vertex X_3 is greater that $\pi/2$. This corresponds to the region on the sphere where $w_1 > 1/4$. However, from Problem 8 at the end of this chapter, we note that the surface area of a sphere cut off by parallel planes is proportional to the distance between the planes. Now, since the sphere $\mathbf{S}^2(1/2)$ has unit diameter, it follows that the probability that the angle at X_3 is greater than $\pi/2$ is $1/4$. Similar results hold for X_1 and X_2 by symmetry. As a triangle can have at most one internal angle greater than $\pi/2$ it follows that $X_1X_2X_3$ is obtuse with probability $3/4$. The result then follows immediately. Q.E.D.

As complex projective spaces are hard to visualize, it is useful to write out the density function in terms of Bookstein coordinates. Once again, the arithmetic of the complex plane is useful. Let $n \geq 3$. We introduce a transformation of complex variables

$$(X_1,\ X_2,\ ...,\ X_n) \leftrightarrow (U,\ V,\ Z_1,\ ...,\ Z_{n-2}) \tag{5.2}$$

where

$$U = \frac{X_1 + X_2}{2} \tag{5.3}$$

$$V = \frac{X_2 - X_1}{2} \tag{5.4}$$

and $Z_1, ..., Z_{n-2}$ are the Bookstein coordinates of the shape of $X_1, ..., X_n$. Let us suppose for the moment that the landmarks $X_1, ..., X_n$ are independent, and have absolutely continuous distributions in the complex plane $\mathbf{C}$ with common density $f(x)$. The joint density of $X_1, ..., X_n$ is

$$\prod_{j=1}^{n} [f(x_j)\ d\mathcal{V}_2(x_j)] \tag{5.5}$$

Under the change of variables in (5.2), the volume elements transform as

$$\prod_{j=1}^{n} d\mathcal{V}_2(x_j) = 4|v|^{2n-4}\ d\mathcal{V}_2(u)\ d\mathcal{V}_2(v) \prod_{j=1}^{n-2} d\mathcal{V}_2(z_j) \tag{5.6}$$

So the joint density of $U, V, Z_1, ..., Z_{n-2}$ with respect to the volume element $d\mathcal{V}_2(u)d\mathcal{V}_2(v)\prod_{j=1}^{n-2} d\mathcal{V}_2(z_j)$ is

$$4\ |v|^{2n-4}\ f(u-v)f(u+v) \prod_{j=1}^{n-2} f(u+vz_j) \tag{5.7}$$

Next, we integrate over the variables u and v to get the density function for the Bookstein coordinates. This has general representation as

$$4 \int_{\mathbf{C}} \int_{\mathbf{C}} |v|^{2n-4}\ f(u-v)f(u+v) \prod_{j=1}^{n-2} f(u+vz_j)\ d\mathcal{V}_2(u)\ d\mathcal{V}_2(v) \tag{5.8}$$

When the density under consideration is spherical normal, then this iterated integral can be computed exactly as follows:

Proposition 5.1.3. *Let* f *be the density function for a spherical normal distribution in the complex plane* $\mathbf{C}$ *centered at the origin. We define*

$$\mathcal{K}(z_1,\ ...,\ z_{n-2}) = 2 + \sum_{j=1}^{n-2} |z_j|^2 - \frac{|\sum_{j=1}^{n-2} z_j|^2}{n} \tag{5.9}$$

Let $f^\sharp$ *be the density of* $Z_1, ..., Z_{n-2}$ *in formula (5.8). Then*

$$f^\sharp(z_1,\ ...,\ z_{n-2}) = \frac{4\ (n-2)!}{n\ \pi^{n-2}\ \mathcal{K}^{n-1}(z_1,\ ...,\ z_{n-2})} \tag{5.10}$$

Proof. Without loss of generality, we can scale the distribution of the landmarks so that the covariance matrix of all X_j's is the identity matrix. Plugging the normal density into formula (5.8) we obtain

$$\frac{4}{(2\pi)^n} \int_{\mathbf{C}} \int_{\mathbf{C}} |v|^{2n-4} \exp[\ --\ ]\ d\mathcal{V}_2(u)\ d\mathcal{V}_2(v) \tag{5.11}$$

where

$$[\ --\ ] = -\frac{1}{2}(|u-v|^2 + |u+v|^2 + \sum_{j=1}^{n-2} |u+z_jv|^2) \tag{5.12}$$

Some simplification is obtained by expanding the absolute values in formula (5.12) using

$$|u-v|^2 + |u+v|^2 = 2|u|^2 + 2|v|^2 \tag{5.13}$$

We can also write

$$|u+z_jv|^2 = |u|^2 + |z_j|^2|v|^2 + 2 < u, z_jv > \tag{5.14}$$

Next, we complete the square in the exponent with respect to the variable u. After some reorganization, we find that (5.11) becomes

$$\frac{4}{(2\pi)^n}\int_{\mathbf{C}}|v|^{2n-4}\exp\left(-\frac{\mathcal{K}|v|^2}{2}\right)\int_{\mathbf{C}}\exp\left(-\frac{n}{2}|u-\frac{v\sum z_j}{n}|^2\right)\,d\mathcal{V}_2(u)\,d\mathcal{V}_2(v) \tag{5.15}$$

The inner integral can be computed by changing variables, expressing $u - v\sum z_j/n$ in polar coordinates (ρ,θ). In polar coordinates, we can express the volume element as $d\mathcal{V}_2(u) = \rho\,d\rho\,d\theta$. Computing the inner integral, we find that (5.15) reduces to

$$\frac{4}{n(2\pi)^{n-1}}\int_{\mathbf{C}}|v|^{2n-4}\exp\left(-\frac{\mathcal{K}|v|^2}{2}\right)\,d\mathcal{V}_2(v) \tag{5.16}$$

Next, we change v to polar coordinates to compute the outer integral. After a routine integration over the two polar coordinates of v we obtain the formula given in (5.10). Q.E.D.

For $n = 3$ we obtain the density for the Bookstein coordinate $Z = Z_1$ to be

$$f^\sharp(z) = \frac{3}{\pi\,(3+|z|^2)^2} \tag{5.17}$$

When $n = 4$ the density of (Z_1, Z_2) reduces to

$$f^\sharp(z_1,z_2) = \frac{128}{\pi^2\,(8+2|z_1|^2+2|z_2|^2+|z_1-z_2|^2)^3} \tag{5.18}$$

Formula (5.10) is a special case of the shape density derived by Mardia and Dryden [116], in which their parameter τ (not to be confused with our notation for pre-shape) goes to infinity.

5.2 Shape Densities under Affine Transformations

5.2.1 Introduction

In this section we shall use formula (5.8) to study the transformation of shape densities when landmark variables are themselves transformed by an affine transformation of the plane. As shape distributions are unaffected by translations and scale changes, it is sufficient to study the effect on shape distributions of linear transformations of the plane that preserve area. We follow the development given in Small [155].

Suppose

$$h : \mathbf{R}^2 \to \mathbf{R}^2 \tag{5.19}$$

is a linear transformation of the plane that is area preserving, so that $\mathcal{J}h \equiv 1$. For $n \geq 3$, let

$$X_1,\ X_2,\ \ldots,\ X_n \in \mathbf{R}^2 \tag{5.20}$$

be IID and continuous with density f.

Now suppose that we let the landmarks $X_1,\ldots,X_n$ be jointly transformed by the common linear transformation h. Define

$$Y_j = h(X_j) \tag{5.21}$$

for $j = 1,2,\ldots,n$. A simple transformation of variables argument shows that the density of Y_j is $f\circ h^{-1}$. Let $Z_1,\ldots,Z_{n-2}$ be the Bookstein coordinates for the shape of the landmarks $X_1,\ldots,X_n$, and let $W_1,\ldots,W_{n-2}$ be the Bookstein coordinates for the shape of the transformed variables $Y_1,\ldots,Y_n$. We will represent the density of $Z_1,\ldots,Z_{n-2}$ by

$$f^\sharp(z_1,\ z_2,\ \ldots,\ z_{n-2}) \tag{5.22}$$

and those of $W_1,\ldots,W_{n-2}$ by

$$(f\circ h^{-1})^\sharp(w_1,\ w_2,\ \ldots,w_{n-2}) \tag{5.23}$$

In this section, we consider how the shape densities $f^\sharp$ and $(f\circ h^{-1})^\sharp$ are related.

Replacing f by $f\circ h^{-1}$ in formula (5.8), we see that $(f\circ h^{-1})^\sharp$ has integral formula

$$4\iint|v|^{2n-4}f[u'-v']f[u'+v']\prod_{j=1}^{n-2}f[u'+(vz_j)']\,d\mathcal{V}_2(u)\,d\mathcal{V}_2(v) \tag{5.24}$$

where $u' = h^{-1}(u)$, $v' = h^{-1}(v)$, and $(vz_j)' = h^{-1}(vz_j)$. For convenience, we recycle notation a bit, replacing u' by u and v' by v. Note that h has unit Jacobian. So the integral reduces to

$$4\iint|h(v)|^{2n-4}f(u-v)f(u+v)\prod_{j=1}^{n-2}f\left\{u+h^{-1}[z_jh(v)]\right\}\,d\mathcal{V}_2(u)\,d\mathcal{V}_2(v) \tag{5.25}$$

Now let us write v in polar coordinates as (ρ,α). Writing the area element $d\mathcal{V}_2(v)$ as $\rho\,d\rho d\alpha$ the integral expression becomes

$$4\int_0^{2\pi}\int_0^\infty\int_{\mathbf{C}}|h(v)|^{2n-4}f(u-v)f(u+v)\prod_{j=1}^{n-2}f\left\{u+h^{-1}[z_jh(v)]\right\}\,d\mathcal{V}_2(u)\,\rho\,d\rho\,d\alpha \tag{5.26}$$

Suppose we now define

$$z_{j\alpha} = e^{-i\alpha}h^{-1}[z_jh(e^{i\alpha})] \tag{5.27}$$

Then our integral becomes

$$\int |h(e^{i\alpha})|^{2n-4} \left\{ 4 \iint \rho^{2n-3} f(u-v) f(u+v) \prod_{j=1}^{n-2} f(u+v z_{j\alpha}) d\mathcal{V}_2(u) d\rho \right\} d\alpha \tag{5.28}$$

Let us now restrict the class of densities f under consideration to those that are *circularly symmetric* about the origin in $\mathbf{C}$. By this we mean that the level curves of the density f are circles centered about the origin, or equivalently, that the distribution is invariant under rotations of the plane about the origin. Then exploiting this symmetry, we see that the expression $\{-\}$ in (5.28) is equal to

$$\frac{1}{2\pi} f^{\sharp}(z_{1\alpha}, z_{2\alpha}, ..., z_{(n-2)\alpha}) \tag{5.29}$$

Thus we obtain the following proposition:

Proposition 5.2.1. *Let f be circularly symmetric about the origin in $\mathbf{C}$. Then*

$$(f \circ h^{-1})^{\sharp}(z_1, ..., z_{n-2}) = \frac{1}{2\pi} \int_0^{2\pi} |h(e^{i\alpha})|^{2n-4} \, f^{\sharp}(z_{1\alpha}, ..., z_{(n-2)\alpha}) \, d\alpha \tag{5.30}$$

where $z_{j\alpha}$ is as defined in formula (5.27).

It should be noted that formula (5.30) makes no reference to the evaluation of densities in the original space of landmarks $\mathbf{C}^n$. So under the symmetry assumption, $(f \circ h^{-1})^{\sharp}$ can be computed directly from $f^{\sharp}$ without reference to f.

5.2.2 *Shape Density for the Elliptical Normal Distribution*

To illustrate Proposition 5.2.1, consider the case where f is spherical normal and $n = 3$. Suppose we consider a linear transformation

$$\Re(z) + i\Im(z) \;\rightarrow\; s^{-1/2}\Re(z) + is^{1/2}\Im(z) \tag{5.31}$$

that *stretches* the plane, taking the circle

$$\Re^2(z) + \Im^2(z) \;=\; 1 \tag{5.32}$$

into the ellipse

$$s\Re^2(z) + s^{-1}\Im^2(z) \;=\; 1 \tag{5.33}$$

The density function $f \circ h$ will be that of an elliptical normal distribution with covariance matrix Γ, where $\Gamma_{11} = s^{-1}$, $\Gamma_{22} = s$, and $\Gamma_{12} = \Gamma_{21} = 0$.

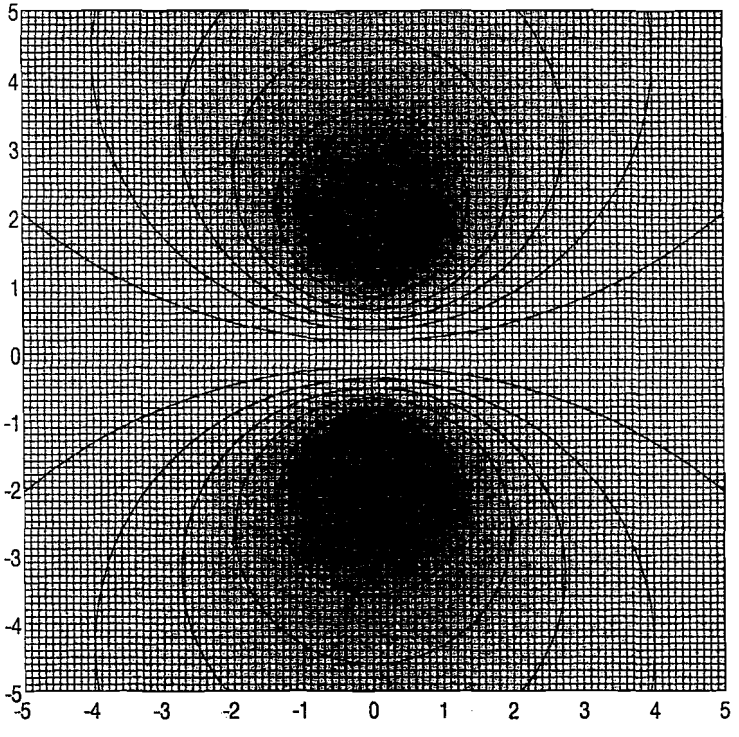

FIGURE 5.1. *The shape density for the elliptical normal. The shape density has been plotted in Bookstein coordinates for graphical convenience. The density displayed, however, is relative to the uniform probability distribution on the sphere of shapes Σ_2^3.*

Plugging formula (5.17) into (5.30) and grinding out the integral, we get

$$(f \circ h^{-1})^{\sharp}(z) = \frac{3(s + s^{-1})}{2\pi \, (3 + |z|^2)^2 \, \{1 + 3[(s - s^{-1})\Im(z)/(3 + |z|^2)]^2\}^{3/2}} \tag{5.34}$$

Figure 5.1 shows a contour plot of the ratio $(f \circ h^{-1})^{\sharp}/f^{\sharp}$, using the stretch factor $s = 2$. The function is symmetrical about the real axis, is maximized on the axis, and is minimized above and below at Bookstein coordinates corresponding to equilateral triangles. The level curves are circles. These circles become a little easier to understand if plotted on the shape sphere $\mathbf{S}^2(1/2)$. Using the coordinates of formula (3.6), we see that the level curves on the sphere are of the form

$$w_3 \;=\; \text{constant} \tag{5.35}$$

As the spherical normal induces a uniform distribution on $\mathbf{S}^2(1/2)$ it follows that these curves are the level curves of the induced density from the

elliptical normal distribution. The interpretation is then clear: the density induced by the elliptical normal is uniformly *squashed* towards the great circle of collinearities corresponding to $w_3 = 0$. The density is maximized on the great circle $w_3 = 0$ and minimized at $w_3 = \pm 1/2$.

The reader should note that formula (5.34) becomes very simple if we restrict ourselves to evaluating the shape density for aligned sets of triangles. These are those for which $\Im(z) = 0$. In such cases, the formula becomes

$$(f \circ h^{-1})^\sharp(z) = \left(\frac{s + s^{-1}}{2}\right) f^\sharp(z) \tag{5.36}$$

5.2.3 *Broadbent Factors and Collinear Shapes*

The simplification in (5.36) for aligned landmarks is not unique to the normal distribution nor to the case $n = 3$. Let us return to the general case of formula (5.30). A general simplification in formula (5.30) is introduced if we restrict attention to those shapes that correspond to aligned sets of landmarks. These are shapes whose Bookstein coordinates $(z_1, z_2, ..., z_{n-2})$ are real, so that $\Im(z_j) = 0$. When z_j is real, then $z_{j\alpha} = z_j$ for all $0 \leq \alpha < 2\pi$. Thus our formula (5.30) reduces to

$$(f \circ h^{-1})^\sharp(z_1, ..., z_{n-2}) = \left[\frac{1}{2\pi}\int_0^{2\pi} |h(e^{i\alpha})|^{2n-4}\, d\alpha\right] f^\sharp(z_1, ..., z_{n-2}) \tag{5.37}$$

The factor

$$\frac{1}{2\pi}\int_0^{2\pi} |h(e^{i\alpha})|^{2n-4}\, d\alpha \tag{5.38}$$

is known as the *Broadbent factor*, named after Simon Broadbent, who proposed its use and calculated some approximate values in [33]. The interpretation of these factors is straightforward. If we suppose that landmarks are initially generated by some circularly symmetric distribution, then, broadly speaking, shapes of landmarks will also tend to be rounded. If the distribution is then stretched by a linear transformation, much as a circle is stretched into an ellipse, then we naturally expect shapes of landmarks to be correspondingly elongated. This means that the shape density on the region corresponding to aligned landmarks undergoes an increase by the Broadbent factor. It can easily be checked that the Broadbent factor is always greater than or equal to one.

For example, consider again the linear transformation

$$h : \Re(z) + i\Im(z) \to s^{-1/2}\Re(z) + is^{1/2}\Im(z) \tag{5.39}$$

as in (5.31). The corresponding Broadbent factor reduces to

$$\frac{1}{2\pi}\int_0^{2\pi} |h(e^{i\alpha})|^{2n-4}\, d\alpha = \mathcal{L}_{n-2}\left(\frac{s + s^{-1}}{2}\right) \tag{5.40}$$

where $\mathcal{L}_m$ is the mth order Legendre polynomial. In particular, the first order Legendre polynomial is the identity function. Thus when $n = 3$, formula (5.37) reduces to (5.36) for all densities f that satisfy the circular symmetry condition of Proposition 5.2.1. In addition, we have

$$\mathcal{L}_2\left(\frac{s + s^{-1}}{2}\right) = \frac{3s^2 + 2 + 3s^{-2}}{8} \tag{5.41}$$

The circular symmetry used to obtain formula (5.37) is stronger than necessary. The following proposition weakens the symmetry assumption used to derive the Broadbent factor.

Proposition 5.2.2. *Let m be the least common multiple of the integers $2k-4$, with $k = 3, 4, ..., n$. Let f be a density for planar distributions that is invariant with respect to rotations by π/m about the origin. Suppose h is an area-preserving linear transformation of the plane. Then on the collinearity set where $z_1, z_2, ..., z_{n-2}$ are real, we have*

$$(f \circ h^{-1})^\sharp(z_1, ..., z_{n-2}) = \left[\frac{1}{2\pi}\int_0^{2\pi} |h(e^{i\alpha})|^{2n-4}\, d\alpha\right] f^\sharp(z_1, ..., z_{n-2}) \tag{5.42}$$

Proof. To prove this, we return to formula (5.28), and let $a(\alpha)$ be the expression $\{-\}$. Furthermore, let $b(\alpha) = |h(e^{i\alpha})|^{2n-4}$. Let us write out $a(\alpha)$ and $b(\alpha)$ in trigonometric series in the variable α. The function $b(\alpha)$ is a trigonometric polynomial of the form

$$\frac{1}{2\pi}\int_0^{2\pi} |h(e^{i\alpha})|^{2n-4}\, d\alpha + \sum_{j=1}^{n-2} c_{1j}\cos(2j\alpha) + \sum_{j=1}^{n-2} c_{2j}\sin(2j\alpha) \tag{5.43}$$

From the rotational symmetry, we see that $a(\alpha)$ has a trigonometric series for which the coefficients of $\cos(2j\alpha)$ and $\sin(2j\alpha)$ are zero for $j = 1, 2, ..., n-2$. The result then follows from the orthogonality of the trigonometric terms. All terms in the integrated product

$$\int_0^{2\pi} a(\alpha)b(\alpha)\, d\alpha \tag{5.44}$$

vanish with the exception of the products of the leading constant terms in the series. Q.E.D.

Of course, in the limiting form as $n \to \infty$ this is simply the circular symmetry assumed earlier. However, for $n = 3$ the assumption is only that f is invariant under rotations by $\pi/2$, a much weaker assumption.

5.3 Tools for the Ley Hunter

To illustrate the methods developed in the last two sections, let us consider a statistical problem that provided some of the impetus for the development of the Kendall school of shape analysis.

In 1925, Alfred Watkins published *The Old Straight Track* [177], which proposed the imaginative hypothesis that a variety of megalithic sites in Britain were, in fact, curiously aligned along tracks he called *leys*. Watkins was an amateur archeologist with a fascination for folklore and mysticism, and his writings drew deeply upon the latter. In addition to sites marked by standing stones and burial chambers, Watkins also included the locations of churches, river fords, and certain place names, on the assumption that although the present-day marker is relatively recent, the site was chosen for its importance as part of the system of ley lines. Watkins' hypothesis is not to be confused with the alignment hypotheses of Alexander Thom and his investigation of megalithic sites as ancient observatories.

The ley hypothesis is unlikely to be settled by statistical argument, because the validity of folklore is not subject to direct statistical analysis. Those who find the arguments from folklore convincing may consider the statistical arguments irrelevant. On the other hand, the hardened empiricist may dismiss the issue out of hand.

However, statistical problems of this nature are commonplace in archeology and deserve consideration as a family of similar questions. In many cases, the presence of patterns in such data can be interpreted as the consequence either of design or of chance, the latter interpretation usually based upon the large number of combinatorial possibilities that the data provide.

The ley line hypothesis is a case in point. For example, consider the coordinates of the 52 megalithic monuments in Cornwall, England known as the *Old Stones of Land's End*. These coordinates are displayed in Figure 5.2. While there are indeed many megalithic sites that can be connected by straight lines to a high degree of precision, we would normally expect a reasonable number of nearly perfect alignments by chance among such a large number. For example, among 52 landmarks, there are 22,100 triangles that can be formed with vertices among the landmarks, and 270,725 quadrilaterals of landmarks. In standard stochastic models, the probability that three or four landmarks are approximately collinear is small. Nevertheless, balancing this is the large number of subsets of triangles and quadrilaterals that can be formed. So we would expect a reasonable number of such collinearities purely by chance.

Among the megalithic data sets, the Old Stones of Land's End have received considerable attention. Broadbent [33] proposed a statistical study of the alignments among these 52 sites, which are plotted in Figure 5.2. The reader can find the data set in [33]. The 52 sites are scattered irregularly across Land's End. Alignments of the sites can be drawn through the points. However, it is difficult to tell *a priori* whether these alignments are

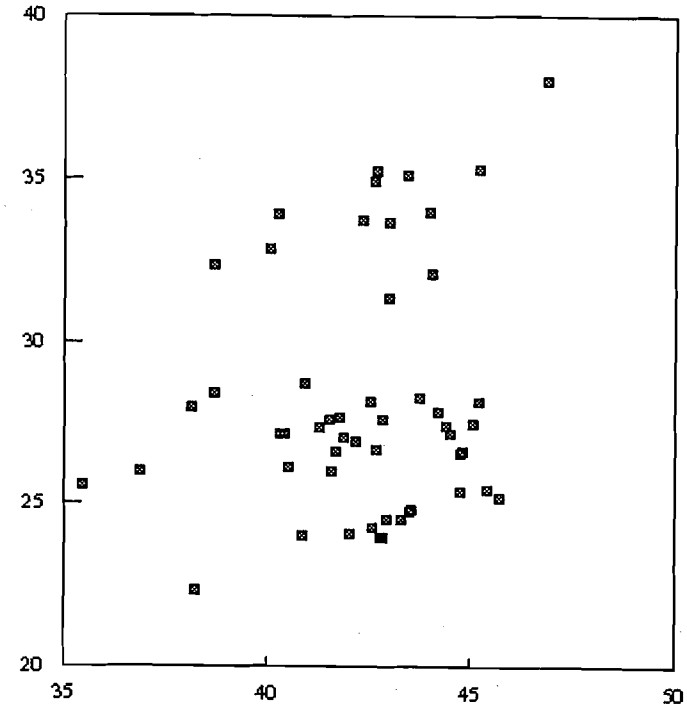

FIGURE 5.2. *The Old Stones of Land's End in Cornwall, England. The 52 plotted points are based upon measurements by John Michell, Chris Hutton-Squire, and Pat Gadsby. The horizontal axis marks the coordinates of the stones in an east-west direction and the vertical axis the coordinates from north to south.*

coincidental.

The first requirement is a definition of approximate alignment of sites or landmarks. A variety of definitions are possible. These are summarized in [33] and examined for their strengths and weaknesses in testing the ley line hypothesis. Overall, the angular criterion, used in [33] and [95], provides the best guarantee of accepting configurations of landmarks that might be intentionally aligned. Following [33] and [95], we adopt such an angular criterion.

Definition 5.3.1. *Three landmarks* X_1, X_2, X_3 *will be said to be* aligned to within tolerance ϵ *if the maximum internal angle of the triangle with vertices at* $X_1X_2X_3$ *is* $\geq \pi - \epsilon$ *radians. We shall also say that the triangle* $X_1X_2X_3$ *is* ϵ-blunt *when this condition is satisfied.*

As there are 52 such sites or landmarks, in a random scattering of 52 points in the plane the expected number of such ϵ-blunt triangles will be 22,100 times the probability that any given triangle $X_jX_kX_l$ is ϵ-blunt.

For the purposes of the analysis that follows, we shall assume that ϵ is sufficiently small that the approximations that follow are reasonable. This will involve discarding higher-order terms in ϵ, which is acceptable provided ϵ is about one degree and certainly less than 5 degrees. Such values would be realistic given the technology available to megalithic architects.

Suppose we model the landmarks as having an elliptical normal distribution in the plane. Let X_1, X_2, and X_3 be IID elliptical normal random vectors in $\mathbf{C}$. We shall assume that the eccentricity of the distribution is governed by the stretch factor s as in Section 5.2.2. For $\epsilon < \pi/2$ the probability that $X_1X_2X_3$ is ϵ-blunt is three times the probability that the internal angle at X_3 is greater than $\pi - \epsilon$. In Bookstein coordinates, the region of shapes where the triangle $X_1X_2X_3$ is ϵ-blunt at X_3 is a lens bounded by the circular arcs that meet the real axis at ± 1 making an angle ϵ. See Figure 5.3. For small ϵ, these circular arcs can be approximated by the parabolas

$$\Im(z) = \pm\frac{\epsilon[1 - \Re^2(z)]}{2}, \quad |\Re(z)| \leq 1 \tag{5.45}$$

When X_j, $j = 1, 2, 3$, are IID elliptical normal with stretch factor s, the shape density for the triangle will be as given by formula (5.34). In particular, we are interested in this shape density close to the set of aligned triangles. Thus we may assume that

$$\Im(z) \approx 0 \tag{5.46}$$

and

$$|z| \approx \Re(z) \tag{5.47}$$

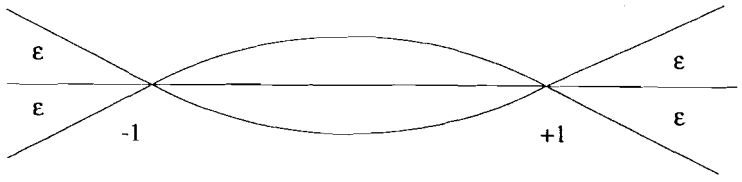

FIGURE 5.3. *Lens of blunt triangles in Bookstein coordinates. In Bookstein coordinates, the region of* ϵ*-blunt triangle shapes is the union of three sets, each set corresponding to a vertex at which the internal angle is* $\geq \pi - \epsilon$*. The set corresponding to* ϵ*-blunt angles at* X_3 *is the lens-shaped region in the middle of the figure. The wedge-shaped region on the left of the lens corresponds to triangles where the* X_1 *is* ϵ*-blunt. Similarly, the wedge-shaped region on the right corresponds to an* ϵ*-blunt angle at* X_2*. If these three sets are plotted on the sphere of triangle shapes, they are seen to be congruent to each other and of equal probability under an IID model for* $X_1X_2X_3$*.*

Applying these approximations in (5.34), we obtain

$$\frac{3(s + s^{-1})}{2\pi[3 + \Re^2(z)]^2} \tag{5.48}$$

as the approximate shape density close to alignment. Integrating (5.48) over the region between the parabolas in (5.45) gives the probability that $X_1X_2X_3$ is ϵ-blunt at the vertex X_3. This must be multiplied by three to allow for blunt angles at the other two vertices. So the probability that $X_1X_2X_3$ is ϵ-blunt is approximately

$$\left(\frac{s + s^{-1}}{2}\right)\frac{(9 - \pi\sqrt{3})\epsilon}{3\pi} \tag{5.49}$$

Note that this formula clearly breaks down when s is large, because the approximation to the probability becomes greater than one. As $s \to \infty$, the shape distribution becomes squashed down onto the real axis in Bookstein coordinates, and the density is no longer approximately constant over the lens in the imaginary coordinate.

Fitting an elliptical normal distribution to the scatterplot in Figure 5.2 gives an estimate of $s = 1.6612$. Thus we would expect on average 164.8 triangles that are blunt to within a tolerance ϵ of one degree. In fact, there are 142 such triangles in the data set, which is within chance variation. Silverman and Brown [153] have shown that under the null hypothesis that the points are IID and continuously distributed in the plane, the distribution of the number of ϵ-blunt triangles is approximately Poisson for small values of ϵ. Thus the number observed is about 1.77 standard deviations below the estimated mean under the null hypothesis model.

This analysis is preliminary at best. Several questions remain. Has the value of ϵ been chosen appropriately? Does the normal model represent

a valid null hypothesis? Should we be looking at alignments of more than three points? Finally, is the criterion of ϵ-bluntness an appropriate one for searching for leys? We cannot take the time to give satisfactory answers here. However, each of these problems can be dealt with briefly.

1. The value of ϵ can be treated as a nuisance parameter of the problem. This analysis leads to the pontogram technique of Kendall and Kendall [95].

2. The normal model is not the only mechanism that can serve as a null hypothesis of random alignment. Uniform scatterings in rectangles have been investigated in [33] and [95]. Uniform scatterings in ellipses have been investigated in [95], [154], and [155]. The expected number of alignments in the uniform elliptical model is less than the elliptical normal model. However, the observed number of alignments is still not statistically significant.

3. Alignments of four points can be investigated. However, the evidence for ley lines does not appear to be much more convincing in this case either. See [33] for some simulations. There are so few alignments of five or more points that it is difficult to draw conclusions of any statistical significance.

4. The use of the maximum angle of the triangle as a measure of alignment is only one of several ways of defining approximate alignment of points. An alternative definition is the *strip definition*, under which a set of points is aligned if it falls entirely within a strip of given width, the width defining the tolerance much as ϵ did for the angular criterion. Again, we refer the reader to [33].

5.4 Independent Uniformly Distributed Landmarks

Another model for IID landmarks that has attracted some attention is that for which the landmarks are uniformly distributed in some bounded convex region of the plane. Let A be such a bounded convex region in the plane with positive area, and suppose that $X_1, X_2, ..., X_n$ are IID uniform in A with $n \geq 3$. Then

$$f(x) = \begin{cases} 1/\mathcal{V}_2(A) & x \in A \\ 0 & x \notin A \end{cases} \tag{5.50}$$

For this case, we can rewrite formula (5.8) for $f^\sharp$ as

$$f^\sharp(z_1, ..., z_{n-2}) = \frac{1}{2^{2n-4}[\mathcal{V}_2(A)]^n} \int\int_B |x_2 - x_1|^{2n-4} \; d\mathcal{V}_2(x_1) \; d\mathcal{V}_2(x_2) \tag{5.51}$$

where

$$B = \{(x_1, x_2) \in A^2 : x_1(1 - z_j)/2 + x_2(1 + z_j)/2 \in A \text{ for all } j\} \tag{5.52}$$

Note that when $\Im(z_j) = 0$ and $-1 \leq \Re(z_j) \leq +1$ then

$$x_1(1 - z_j)/2 \; + \; x_2(1 + z_j)/2 \tag{5.53}$$

is a convex combination of x_1 and x_2 and lies on the line segment from x_1 to x_2. As A is assumed to be a convex set, this convex combination will then lie in A, and the indicator function will equal one. If this is true for all $j = 1, 2, ..., n-2$, then the shape density $f^\sharp$ reduces to

$$\frac{1}{2^{2n-4}[\mathcal{V}_2(A)]^n} \int_{\mathbf{A}} \int_{\mathbf{A}} |x_2 - x_1|^{2n-4} \; d\mathcal{V}_2(x_1) \; d\mathcal{V}_2(x_2) \tag{5.54}$$

These Bookstein coordinates correspond to aligned sets of landmarks with $X_3, ..., X_n$ falling on the line segment from X_1 to X_2. Under a permutation of the labels of the landmarks, any aligned set of landmarks can be written in this form. Thus it is possible to find the density function for shape of any aligned set or the approximate probability of an approximate alignment of landmarks. It is the latter that was useful in studying the Land's End data of the previous section. Thus the uniform model provides an alternative to the normal model used earlier. Problems 2 and 3 invite the reader to analyze the Land's End data using the uniform model of points scattered in an ellipse or rectangle.

The integral in (5.54) is closely related to the well-known Blaschke constants of the convex set A. See [147, pp. 46–49]. For example, for a circular disk A of unit radius, we find that (5.54) becomes

$$\frac{1}{2^{2n-3}\pi^{n-2}(n-1)(2n-1)} \binom{2n}{n} \tag{5.55}$$

from which the value of $f^\sharp$ can be evaluated when $\Im(z_j) = 0$ and $-1 \leq \Re(z_j) \leq +1$ for $j = 1, 2, ..., n-2$. Multiplying by a Broadbent factor gives us

$$\frac{\mathcal{L}_{n-2}[(s + s^{-1})/2]}{2^{2n-3}\pi^{n-2}(n-1)(2n-1)} \binom{2n}{n} \tag{5.56}$$

which provides the corresponding density for an ellipse with stretch factor s. See Problems 2 and 3 for an application to the collinearity calculation of the Old Stones of Land's End.

For values of the shape density off the aligned region in Bookstein coordinates, the integral is harder to evaluate. The function $f^\sharp$ is typically complicated but can be found in closed form. See Le [101, 102] for some excellent work on this difficult problem.

5.5 Landmarks from the Spherical Normal: Non-IID Case

The main intention of this section is to prove a beautiful result due to Mardia [114] based on the work of Mardia and Dryden [116, 117], that the density function for the shape of non-IID spherical normal landmarks in

the plane has a particularly elegant form. We follow the derivation given by Goodall and Mardia [70].

First of all, we will need a lemma.

Lemma 5.5.1. *Let* $\mathcal{I}_3$ *be the function defined by equation (4.62) with* $k = 2$. *Then*

$$\int_0^{2\pi} \mathcal{I}_3[t\ \cos(\theta)]\ d\theta\ =\ (2\pi)(2+t^2)e^{t^2/2} \tag{5.57}$$

Proof. Returning to formula (4.63), and setting $\nu = t$ and $n = 2$, we see that

$$\int_0^{2\pi} \exp(-t^2/2)\ \mathcal{I}_1[t\ \cos(\theta)]\ =\ 2\pi \tag{5.58}$$

because the density in (4.63) must integrate to one. Rearranging (5.58) we get

$$\int_0^{2\pi} \cos(\theta)\exp[t^2\cos^2(\theta)/2]\Phi[t\cos(\theta)]\ d\theta\ =\ \frac{\sqrt{2\pi}[\exp(t^2/2)-1]}{t} \tag{5.59}$$

Differentiating both sides with respect to t we obtain

$$\begin{aligned}\int_0^{2\pi} &\cos^3(\theta)\exp[t^2\cos^2(\theta)/2]\Phi[t\cos(\theta)]\ d\theta \\ &= \sqrt{2\pi}\ t^{-3}[1-t^2/2+t^2\exp(t^2/2)-\exp(t^2/2)]\end{aligned} \tag{5.60}$$

Using formulas (4.64–4.66) we see that

$$\mathcal{I}_3(t) = (2+t^2) + \psi(t)(3t+t^3) \tag{5.61}$$

where $\psi(t) = \Phi(t)/\phi(t)$. Plugging (5.61) into the left-hand side of (5.57), we can expand the integral into four terms. Applying the identities in (5.60) and (5.61) gives the required result. Q.E.D.

Now, suppose X_1, X_2, X_3 are independent landmarks, with X_j having a normal distribution centered at mean point $\mu_j \in \mathbf{R}^2$. Suppose also that the three landmarks have a common covariance matrix, which is some multiple of the identity matrix. Without loss of generality, we can scale the problem so that this covariance matrix is the identity. Let $\sigma \in \mathbf{S}^2(1/2)$ be the shape of the random triangle $X_1X_2X_3$ as expressed in Kendall's shape space $\Sigma_2^3 \cong \mathbf{S}^2(1/2)$. Then we have the following proposition:

Proposition 5.5.2. *The density of* σ *with respect to the area element* $d\mathcal{V}_2(\sigma)$ *is given by*

$$\pi^{-1}(1+\beta<\sigma+\sigma_\mu,\sigma_\mu>)\exp(\beta<\sigma-\sigma_\mu,\sigma_\mu>) \tag{5.62}$$

where σ_μ *is the shape of the triangle* $\mu_1\mu_2\mu_3$ *and* $\beta = \sum||\mu_j - \bar{\mu}||^2$, *with* $\bar{\mu} = (1/3)\sum\mu_j$.

The inner product used in this density formula is that of $\mathbf{R}^3$ with the sphere $\mathbf{S}^2(1/2)$ embedded in $\mathbf{R}^3$ as a sphere of radius $1/2$ centered at the origin. Note that this formula differs slightly from that given in [70]. We have chosen to construct the density function on the shape space $\mathbf{S}^2(1/2)$ rather than to renormalize the radius of the sphere to one. The result is that our β is four times the concentration parameter used in [70].

Before we derive this formula, some observations need to be made. This density function is unimodal, with maximum value at σ_μ and minimum value at the antipodal point $-\sigma_\mu$. If we think of σ_μ as a north pole then the density function is seen to be constant along lines of latitude on the sphere. The constant β acts as a measure of concentration of the density about σ_μ. High values of β produce distributions that are closely concentrated about σ_μ while low values produce distributions that are more diffuse on the sphere. In fact, as μ_1, μ_2, μ_3 converge to some common point in $\mathbf{R}^2$ the constant β goes to zero and the density function converges to $\pi^{-1} = 1/\mathcal{V}_2[\mathbf{S}^2(1/2)]$. This is the uniform distribution on $\mathbf{S}^2(1/2)$ and is seen to be in agreement with the result in Section 5.1.

Proof of Proposition 5.5.2. Let X be the 2×3 matrix of the coordinates of X_1, X_2, and X_3, such that the jth column is X_j. As in equation (4.69), let Y be the 2×2 matrix of coordinates of the centered landmarks found by right multiplying X by Λ^T, where Λ is the 2×3 row-deleted Helmert matrix of order 3. Then we can regard Y as a spherical normal random vector in $\mathbf{R}^4$, so that

$$\tau\ =\ \frac{Y}{||Y||} \tag{5.63}$$

is the pre-shape of the landmarks. As we observed in Section 4.5.3, τ has a projected normal distribution on $\mathbf{S}^3 \subset \mathbf{R}^4$ with density function

$$(2\pi)^{-2}e^{-(||\mu_Y||^2/2)}\ \mathcal{I}_3[<\tau,\mu_Y>] \tag{5.64}$$

where μ_Y is the mean vector of Y. It is straightforward to check that $\beta = ||\mu_Y||^2$.

The orbits of $\mathbf{S}^3$ are the equivalence classes of pre-shapes that share a common shape, and the decomposition of $\mathbf{S}^3$ into its orbits is a decomposition of the sphere into congruent circles, all equivalent to $\mathbf{S}^1$. To calculate the shape density, we need to integrate out one additional dimension by constructing a coordinate within the orbits of $\mathbf{S}^3$ and integrating over it. Suppose $\theta \in \mathbf{S}^1$ is a coordinatization of the points in the orbits. That is, any pre-shape τ can be completely specified by the pair (σ, θ), where σ is the shape of the configuration and θ is its orientation. In more technical

language, we can say that $\mathbf{S}^3$ is a fiber bundle with fibers congruent to $\mathbf{S}^1$ and base space $\mathbf{S}^2(1/2)$. The geometric measure on $\mathbf{S}^3$ decomposes as the product

$$d\mathcal{V}_3(\tau) \;=\; d\mathcal{V}_1(\theta)\; d\mathcal{V}_2(\sigma) \tag{5.65}$$

although the fiber bundle is not a Cartesian product of $\mathbf{S}^2(1/2)$ and $\mathbf{S}^1$. Nevertheless, we do have the factorization

$$\mathcal{V}_3(\mathbf{S}^3) \;=\; \mathcal{V}_1(\mathbf{S}^1) \times \mathcal{V}_2[\mathbf{S}^2(1/2)] = 2\pi^2 \tag{5.66}$$

A formula for the shape density we seek is found by integrating formula (5.64) with respect to the geometric measure on θ. This yields the integral

$$(2\pi)^{-2} e^{-\beta/2} \int_0^{2\pi} \mathcal{I}_3[< \tau(\sigma,\theta), \mu_Y >]\; d\theta \tag{5.67}$$

Let

$$\tau_\mu \;=\; \frac{\mu_Y}{||\mu_Y||} \tag{5.68}$$

be the pre-shape of the figuration of landmark means μ_1, μ_2, μ_3. For any pre-shape τ we can draw a horizontal geodesic from τ_μ to a pre-shape in $\mathcal{O}(\tau)$. We define θ to be the angle this pre-shape $\tau_\mu(\sigma)$ makes with τ. Then we find that the integral formula in (5.67) can be rewritten as

$$(2\pi)^{-2} e^{-\beta/2} \int_0^{2\pi} \mathcal{I}_3[\sqrt{\beta} < e^{i\theta}\tau_\mu(\sigma), \tau_\mu >]\; d\theta \tag{5.69}$$

using the identification that $\mathbf{R}^4 \cong \mathbf{C}^2$.

Now, we can write

$$\begin{aligned} < e^{i\theta}\tau_\mu(\sigma), \tau_\mu > \;&=\; \Re \ll e^{i\theta}\tau_\mu(\sigma), \tau_\mu \gg \\ &=\; \cos(\theta) < \tau_\mu(\sigma), \tau_\mu > -\sin(\theta)\; \Im \ll \tau_\mu(\sigma), \tau_\mu \gg \end{aligned} \tag{5.70}$$

As $< e^{i\theta}\tau_\mu(\sigma), \tau_\mu >$ is maximized at $\theta = 0$ it follows that

$$\Im \ll \tau_\mu(\sigma),\; \tau_\mu \gg \;=\; 0 \tag{5.71}$$

Thus our expression (4.37) for the density becomes

$$(2\pi)^{-2} e^{-\beta/2} \int_0^{2\pi} \mathcal{I}_3[\sqrt{\beta}\, \cos(\theta) < \tau_\mu(\sigma), \tau_\mu >]\; d\theta \tag{5.72}$$

Using Lemma 5.5.1 we can evaluate the integral. The density in formula (5.72) becomes

$$\pi^{-1}\; e^{-\beta/2}\; [1 + \beta < \tau_\mu(\sigma), \tau_\mu >^2/2]\; e^{\beta < \tau_\mu(\sigma), \tau_\mu >^2/2} \tag{5.73}$$

The next step is to express the inner product between the pre-shapes in terms of the inner product between the shapes. First note that $\tau_\mu(\sigma)$ has been chosen to have minimum geodesic distance from τ_μ among all pre-shapes in $\mathcal{O}(\tau)$. So the geodesic distance in $\mathbf{S}^3$ from τ_μ to $\tau_\mu(\sigma)$ is the Procrustean distance from σ_μ to σ. But this is a geodesic distance on the sphere $\mathbf{S}^2(1/2)$. Identifying these two representations of the Procrustean distance we get

$$\cos^{-1}(< \tau_\mu(\sigma), \tau_\mu >) \;=\; \frac{1}{2}\cos^{-1}(4 < \sigma_\mu, \sigma >) \tag{5.74}$$

Taking the cosine of both sides, we obtain

$$\frac{1}{2} < \tau_\mu(\sigma), \tau_\mu >^2 = < \sigma_\mu, \sigma > \; + \frac{1}{4} \tag{5.75}$$

Plugging (5.75) into (5.73), and using the fact that

$$< \sigma_\mu,\; \sigma_\mu > \;=\; \frac{1}{4}$$

gives us the required result. Q.E.D.

The evaluation of the shape density for more than three non-IID normal landmarks is more complicated but can be obtained in terms of confluent hypergeometric functions. See [53], [116], and [117]. The general form of the density function on Σ_2^n is

$${}_1F_1\{2-n; 1; -\beta_{n-2}[1+\cos(2\alpha)]\}\; \exp[\beta_{n-2}\,\cos(2\alpha) - \beta_{n-2}] \tag{5.76}$$

where $\beta_{n-2} = ||\mu_Y||^2/4$ is a concentration parameter, ${}_1F_1$ is the confluent hypergeometric function, and $0 \le \alpha < \pi/2$ is the geodesic distance from σ_μ to σ on Σ_2^n. The computation of ${}_1F_1$ is straightforward here, because its representation with these values is as a finite series. In particular, we note that

$${}_1F_1(-k; 1; -x) = \sum_{j=0}^{k} \binom{k}{j} \frac{x^j}{j!} \tag{5.77}$$

To find the analogous density for Bookstein coordinates, we multiply (5.76) by (5.10).

5.6 The Poisson-Delaunay Shape Distribution

In Section 4.7.4 we obtained the pre-size-and-shape distribution of a random Delaunay simplex that is generated from a Poisson process. In this section, we shall consider a result due to Kendall [93] that extends a formula for the distribution of shape of a Delaunay triangle due to Miles [119].

Let Δ be a random Delaunay p-simplex from a Poisson process of intensity ρ in $\mathbf{R}^p$, as was obtained in Section 4.7.4. There are a variety of ways that a random Delaunay simplex can be chosen from a Poisson process. For example, we could pick a particle of the process at random from among those that fall into a bounded region. There will be a number of simplexes of the Delaunay tessellation that have this particle as one of their vertices. Among these simplexes, we could choose one at random and call it Δ. We label the vertices of Δ as $X_1, ..., X_{p+1}$.

Suppose that we represent shapes in generalized Bookstein coordinates. Let $f^{\sharp}_{\text{norm}}$ be the shape density when $X_1, X_2, ..., X_{p+1}$ are IID spherical normal landmarks in $\mathbf{R}^p$, and let $f^{\sharp}_{\text{del}}$ be the shape density when $X_1, X_2, ..., X_{p+1}$ are the vertices of a Poisson-Delaunay random simplex with distribution as in formula (4.100). We define

$$v = \sum_{j=1}^{p+1} ||X_j - \bar{X}||^2 \tag{5.78}$$

where as usual $\bar{X} = (p+1)^{-1} \sum X_j$. Define r to be the circumradius of the $(p-1)$-dimensional sphere through $X_1, X_2, ..., X_{p+1}$. We note that the quantity

$$\chi = \frac{r}{\sqrt{v}} \tag{5.79}$$

is a shape statistic. Then we have the following elegant result of Kendall [93].

Proposition 5.6.1. *Let $X_1, ..., X_{p+1}$ be the vertices of a Delaunay simplex Δ chosen at random from the Delaunay tessellation of $\mathbf{R}^p$ generated by a Poisson process. Then the shape density of $X_1, ..., X_{p+1}$ has the form*

$$f^{\sharp}_{\text{del}} = a(p)\, \chi^{-p^2}\, f^{\sharp}_{\text{norm}} \tag{5.80}$$

where $a(p)$ is a constant of integration depending only upon the dimension p.

Proof. Let us return to the setting of Section 4.7.4 and in particular formula (4.100). The first step in the proof is to calculate the Delaunay-Poisson pre-shape density by integrating over a scale variable in formula (4.100). We change coordinates, transforming from $y_1, y_2, ..., y_p$ to τ and t, where $\tau \in \mathbf{S}^{p^2-1}_{\star}$ is the pre-shape of the simplex and $t = \sqrt{v}$ is a scale variable.

The pair of variables

$$(t,\ \tau) \in \mathbf{R}^+ \times \mathbf{S}^{p^2-1}_{\star} \tag{5.81}$$

can be regarded as polar coordinates for the vector of centered vertices $(x_1 - \bar{x}, ..., x_{p+1} - \bar{x})$, which is in 1-1 correspondence with $(y_1, ..., y_p)$.

Under this change of variables, we get a typical formula for the volume element in $\mathbf{R}^{p^2}$, namely,

$$d\mathcal{V}_p(y_1) ... \, d\mathcal{V}_p(y_p) \ \propto\ t^{p^2-1}\, dt\, d\mathcal{V}_{p^2-1}(\tau) \tag{5.82}$$

Next, we rewrite (5.79) as $r = \chi t$. From this we see that

$$f(y_1, ..., y_p)\, d\mathcal{V}_p(y_1) ... \, d\mathcal{V}_p(y_p) \tag{5.83}$$

is proportional to

$$\rho^p \exp[-\rho\, \kappa_p\, \chi^p(\tau)\, t^p]\ t^{p^2}\ \frac{dt}{t}\ d\mathcal{V}_{p^2-1}(\tau) \tag{5.84}$$

To eliminate scale, we integrate, yielding

$$\int_{t=0}^{t=\infty} \left\{ -\rho^p \exp[\rho \kappa_p \chi^p(\tau) t^p]\ t^{p^2}\ d\mathcal{V}_{p^2-1}(\tau) \right\} \frac{dt}{t} \ =\ \frac{(p-1)!}{p\ \kappa_p^p\ \chi^{p^2}}\ d\mathcal{V}_{p^2-1}(\tau) \tag{5.85}$$

So, $1/\chi^{p^2}$ is proportional to the density function of the pre-shape of a Poisson-Delaunay simplex.

However, the pre-shape τ' of a simplex generated by IID spherical normal landmarks $X'_1, ..., X'_n$ has a uniform distribution on the pre-shape sphere $\mathbf{S}^{p^2-1}_{\star}$. Thus the ratio of the two densities is proportional to $1/\chi^{p^2}$ as well. This looks very much like the result we need, with the exception that it applies to pre-shapes instead of shapes. To obtain the same ratio for shapes, we note that both the Poisson-Delaunay pre-shape density and the normal pre-shape density functions are uniform over orbits of $\mathbf{S}^{p^2-1}$ that correspond to rotations in $\mathbf{R}^p$. Thus when we integrate across these orbits, the same ratio is preserved. Q.E.D.

Kendall [93] has computed the value of the coefficient $a(p)$ in formula (5.80) to be

$$a(p) = \frac{(p+1)^{p/2} \{\text{Gam}[(p+1)/2]\}^p}{2^p \pi^{(p-1)/2} \text{Gam}[(p^2+1)/2]} \tag{5.86}$$

In the case $p = 2$, we have an explicit formula for $f^{\sharp}_{\text{del}}$ from formula (5.17). Combining this with Proposition 5.6.1 and using $a(2) = 1/4$ from (5.86) we can obtain the following:

Corollary 5.6.2. *In terms of the Bookstein coordinate z for a Poisson-Delaunay triangle $X_1X_2X_3$ the shape density has the form*

$$f^{\sharp}_{\text{del}}(z) = \frac{1}{3\pi} \left\{ 1 + \left[\frac{|z|^2 - 1}{2\Im(z)} \right]^2 \right\}^{-2} \tag{5.87}$$

Proof. We can write out the formula for χ^{-4} in terms of the Bookstein coordinate z as

$$\frac{64[\Im(z)]^4(3+|z|^2)^2}{9\{4[\Im(z)]^2+(|z|^2-1)^2\}^2} \tag{5.88}$$

Plugging (5.88), (5.86) with $p=2$, and (5.17) into formula (5.80) gives us the result. Q.E.D.

Corollary 5.6.2 is essentially the same as the shape density for Delaunay triangles (in terms of interior angles of the triangles) due to Miles [119].

Although Proposition 5.6.1 does not give us an immediate formula for the shape density of Poisson-Delaunay simplexes in dimensions $p>2$, it nevertheless provides an excellent mechanism for the simulation of such simplexes without resort to generating a Poisson process. As spherical normal landmarks are easy to simulate, an acceptance method can be used that first simulates a normal simplex and accepts this simplex with probability $\chi^{-p^2}/\max(\chi^{-p^2})$. This will generate realizations of the shapes of Poisson-Delaunay simplexes. It can be demonstrated that χ is minimized when $\chi = 1/\sqrt{p+1}$. See Problem 7. For further analysis and comments on this simulation technique, see [93].

5.7 Notes

The development of this chapter roughly follows the historical order in which the distribution theory for shapes was developed. The earliest work by David Kendall, Wilfrid Kendall, and Simon Broadbent concentrated on the IID uniform and normal models with a possible eccentricity parameter. The calculation of collinearity probabilities and Broadbent factors was developed by Small [154, 155]. The earliest reference on the uniformity of shape distributions under a spherical normal model would appear to be by Kendall [87], where the model under consideration was that of the shape distribution of a set of points in the plane that diffuse independently from a common starting point as Brownian motions. Applications of the distribution theory were typically archeometric in nature. The value of general shape distributions for biometric and morphometric applications was developed by Fred Bookstein, Kanti Mardia, Ian Dryden, Colin Goodall, and others. Bookstein proposed a landmark model with normally distributed landmarks in which the landmark variability about their respective means is small compared to the distances between landmark means. This leads to the so-called "tangent approximation" (to use David Kendall's terminology) in which the shape statistics, when expressed in Bookstein coordinates, have approximately a normal distribution. This follows from a Taylor expansion of the formula for Bookstein coordinates, in which the dominant term is a linear transformation of the original normal variables. The discovery that shape variables under such circumstances can be approximately normal is reassuring, because there is a large literature on multivariate statistical analysis that can be tapped. Such models correspond to the circumstance in Proposition 5.5.2 where the concentration parameter β is large.

The idea of using statistical techniques on shape variables that are commonly associated with multivariate normal theory also arises in allometry, where the logarithms of size variables are jointly plotted and analyzed for collinearities whose presence supports the growth allometry model of formulas (1.1) and (1.2).

5.8 Problems

1. In [39], Lewis Carroll (Charles Dodgson) proposed a number of mathematical "pillow problems" that Carroll claimed to have solved in bed. On the evening of January 20, 1884, he stayed up late to solve the following, which became number 58 on his list of pillow problems. Find the probability that three points chosen at random in the plane have an obtuse angle. The solution given proceeds thus: Let $X_1X_2X_3$ be such a triangle. Without loss of generality, we can assume that X_1X_2 is the longest side. Then X_3 lies in the lune that is the intersection of the two circular disks having centers at X_1 and X_2 respectively, and common radius $|X_1-X_2|$. The triangle will have an obtuse angle if and only if X_3 lies in the circular disk with center at the midpoint of X_1X_2 and radius $|X_1-X_2|$. The ratio of the area of this circular disk to the area of the lune is

$$\frac{\pi/8}{\pi/3-\sqrt{3}/4} \tag{5.89}$$

which is taken as the required probability.

(a) Comment on this solution, discussing its assumptions and its validity.

(b) Compare the solution with Corollary 5.1.2. Which is more convincing?

(c) Find the probability of an obtuse angle for three independent points that are uniformly distributed on the boundary of a circle.

For further reading on this interesting problem, see [39] and [135].

2. Let X_1, X_2, X_3 be IID uniformly distributed in an elliptical region with stretch factor s as in 4.4.2. Find the approximate probability that the triangle having X_1, X_2, and X_3 as vertices is ϵ-blunt to within a tolerance ϵ of one degree. This can be derived following the pattern of Section 5.3, applying formula (5.56) for the elliptical uniform shape density close to alignment instead of the elliptical normal shape density.

3. Compute the approximate probability that $X_1X_2X_3$ is ϵ-blunt to

within a tolerance of one degree as in Problem 2 above using a rectangle instead of an ellipse. Assume the sides of the rectangle are in proportion $s : 1$. How does this compare with Problem 2?

4. Find the density function for the distribution of shape in Bookstein coordinates for a triangle of three independent points uniformly distributed on the circle $x_1^2 + x_2^2 = 1$.

5. For a triangle in the plane with vertices X_1, X_2, X_3 find the Bookstein coordinate Z in terms of the three internal angles of the triangle. Use this to find the joint shape density for the internal angles of a Poisson-Delaunay triangle using Corollary 5.6.2.

6. From Problem 5 above. Suppose that one of these three angles is chosen at random. Show that the distribution of this angle θ has density

$$\frac{4}{3\pi}[(\pi - \theta)\cos(\theta) + \sin(\theta)]\sin(\theta) \tag{5.90}$$

7. From Section 5.6, prove that χ of formula (5.80) is minimized when $\chi = 1/\sqrt{p+1}$.

8. Find a formula for the surface area of a sphere bounded between two parallel planes intersecting the sphere. Using the fact that this surface area is proportional to the distance between the planes, fill in the details of Corollary 5.1.2.

6

Some Examples of Shape Analysis

6.1 Introduction

In this chapter, we shall consider in greater detail some examples that were first mentioned in Chapter 1. While the Land's End data of Chapter 5 were accessible to analysis largely by shape theory alone, most spatial data sets contain scale and orientation information that should not be ignored. In many cases, a shape analysis is performed in order to find the relationship between size and shape. This is of interest in growth allometry, as was mentioned in Chapter 1. However, the relationship between size and shape is of interest more generally, as is evident in the dinosaur footprints example described in Section 6.2 below.

6.2 Mt. Tom Dinosaur Trackways

In this section, we continue the investigation through a size, orientation, and shape analysis of the dinosaur trackways of Section 1.4.2. See Figure 1.4.

The collection of footprints at this site has undergone considerable deterioration due to weathering and vandalism, making precise statistics on the distribution of footprint dimensions impossible to collect, as Ostrom noted. However, it was possible to classify the footprints into three groups, with the largest being tentatively identified as *Eubrontes giganteus*, the inter-

mediate size prints being also tentatively identified either as *Anchisauripus sillimani* or as immature *Eubrontes* prints (the former being favored), and the smallest prints as *Grallator cuneatus*. The dinosaur *Eubrontes* was an early, medium-sized, tridactylic bipedal therapod, believed to be predatory in nature. On the other hand, *Grallator* was a small dinosaur of the same period that was bipedal and tridactylic, either a predatory therapod or a herbiverous ornithischian. It was found that *Eubrontes* prints varied in length from roughly 28 to 35 cm., with the caveat that erosion makes precise determination of dimensions impossible. The *Anchisauripus* prints varied in length from 15 to 20 cm. approximately, and the *Grallator* prints varied from 9 to 12 cm. in length.

The data set was recorded on 20×20 graph paper at a scale of 10 ft. to the inch, the site being divided into five-foot squares for the purpose. Upon examination, the footprints were grouped into trackways, with some element of uncertainty in a number of cases, particularly where trackways cross. Uncertainty also arises in deciding whether two trackways along a common line were made by a single dinosaur or by two. For example, it is unclear whether footprint D should be grouped with trackway 15 or separately. Similarly, prints A, B, and C could be grouped with trackways 7, 9, and 11 respectively. For statistical purposes, it seems appropriate to analyze within and between trackway variation in size, orientation, and shape only on that subset of footprints that can be clearly classified. The loss of information by so doing is less problematic that the difficulty of outlier contamination by including all prints. Thus trackways 5, 6, and 7 are somewhat confounded with each other. For the purposes of statistical analysis, we take the first three prints of trackway 5, the first four prints of trackway 6, and the first two prints of trackway 7.

When grouped according to species, trackway 7 (counting footprint A as a continuation of trackway 7), trackway 27, and trackway 28 were classified by Ostrom as belonging to *Anchisauripus*. Trackway 14 and trackway 18 were classified as *Grallator*. All the rest were classified as trackways made by *Eubrontes*. The greatest uncertainty in this three-fold classification is in trackway 13, consisting of a single isolated print, and trackway 7, consisting of two prints.

6.2.1 *Orientation Analysis*

We have already considered in broad terms some of the orientation information in the trackways and its relationship to possible gregarious behavior of *Eubrontes*. There are two types of orientation information within a trackway: the orientation of the footprint and the direction of the trackway. Footprint orientation is typically compatible with the orientation of the trackway, and should be considered of importance in an orientation analysis in providing an ordering to the prints along a trackway. However, there is considerable damage to the footprints, making orientation of a footprint

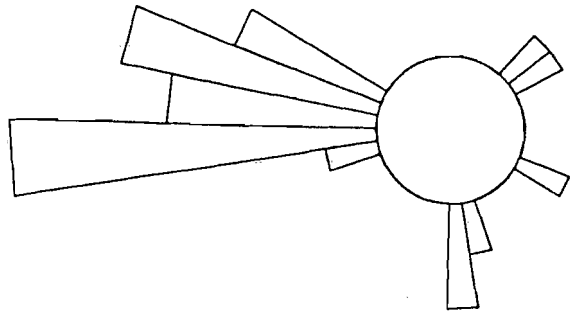

FIGURE 6.1. *Distribution of trackway orientations. The Mt. Tom dinosaur trackways can be individually oriented by taking a unit vector pointing in the direction from the first observable footprint in the trackway to the last observable footprint. Such a unit vector can be regarded as a point on the unit circle centered about the origin in the plane. Figure 6.1 shows the histogram of the scattering of trackway orientations as they are seen in Figure 1.4. It is evident that the vast majority of the tracks point in a westerly direction. These tracks are largely* Eubrontes *footprints.*

difficult to determine in isolation from other footprints in the same trackway. Thus the trackway orientation would seem to be of greater importance in the analysis. The trackways are fairly straight, with the exception of number 17, which shows a slight but systematic curvature.

We can encode the directions of the trackways by taking a vector from the initial footprint to the final footprint of the trackway and standardizing the vector to have length one. The exception to this definition is trackway 13, which contains only one footprint. The orientation of this trackway must be established roughly from the orientation of the footprint. The result is a directional data set that can be plotted on the unit circle. A histogram of the orientations can be seen in Figure 6.1.

To estimate the overall direction of dinosaurs crossing the area, the *directional median* seems appropriate as it is less sensitive to large deviations in direction away from the overall trend, in this case to the west. If $\theta_1, \theta_2, ..., \theta_n$, $0 \leq \theta_j < 2\pi$, are the angles of a set of n directional vectors, then the median of the angles is that value θ minimiz-

ing the sum of geodesic distances $\sum_j d(\theta, \theta_j)$ around the circle, where $d(\theta, \theta_j) = \min(|\theta - \theta_j|, 2\pi - |\theta - \theta_j|)$. For our data set, the directional median is achieved on an interval of angles intermediate between trackway 4, where $\theta_4 = 3.1363$, and trackway 5, where $\theta_5 = 3.146$. (The fact that these trackways are also consecutive appears to be a coincidence of Ostrom's numbering system.) Averaging the angles over this interval provides a convenient summary of the median direction. The median works out to be $\theta \simeq 3.14$, which is very close to due west, placing due north on the vertical axis of the coordinate system. (Considerable continental drift has occurred since the period when the trackways were formed. Therefore due west at present does not correspond to due west during the period when the footprints were preserved.) A total of 20 out of 28 trackways fall within a narrow $30°$ interval about the median direction.

6.2.2 *Scale Analysis*

There are two measures of size within a trackway that are relevant to the analysis of dinosaur locomotion. These are the footprint length and the stride length, the latter usually defined as the distance between successive footprints. Figure 6.2 shows a set of 28 boxplots of the sample distributions of stride lengths along the 28 trackways. The variation in stride lengths between trackways is most evident between trackways 14 and 18, classified as *Grallator*, and those trackways classified as *Eubrontes*, a much larger dinosaur. (The reader who wishes to see a comparison of *Grallator* and *Eubrontes* footprints is referred to page 128 of [133].) The differences in stride lengths can be interpreted as due to two factors, the first being the length of the dinosaur's legs from hip to foot and the second being the speed of the dinosaur. Footprint dimensions give us some indication of the size of the dinosaur, from which it is possible to estimate the speed that the dinosaur had when crossing the site. We noted the footprint dimensions above. These values are variable, even within a trackway, and so it seems safest to use species averages alone in the formulas.

Alexander [1, 2] has proposed a formula for dinosaur speed based upon footprint length and stride length using *Froude numbers*. Froude numbers and the associated theory proposed by Alexander suggests that if two bipedal animals of similar shape have a size ratio of $a : b$ in linear dimensions then their speeds will be in the ratio $\sqrt{a} : \sqrt{b}$. This suggests that the appropriate formula linking dinosaur speed to stride length and footprint length is of the form

$$\text{speed} = c \times (\text{stride length})^p \times (\text{footprint length})^{-p+0.5} \qquad (6.1)$$

for appropriate constants c and p. A regression can be performed on modern bipedal species to fit the constants c and p. From this fit we can tentatively estimate dinosaur speed. Using Alexander's empirical fit to a

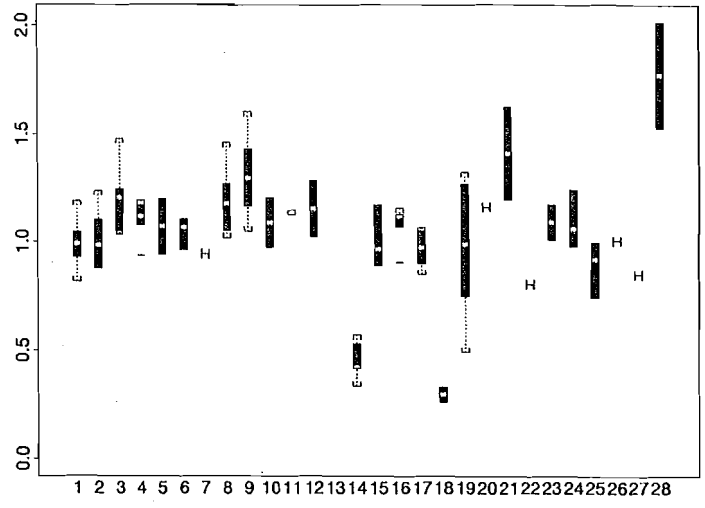

FIGURE 6.2. *Boxplots of stride lengths for dinosaur trackways. Distances are shown in meters. The trackways are ordered along the horizontal axis from 1 to 28. Each trackway has its stride length distribution displayed by a vertical boxplot that is constructed as follows: Each box appears as a thin dark rectangle with endpoints at the upper and lower quartiles of the distribution. The white horizontal strip inside each box marks the location of the median of the distribution. At each end of the box, dotted lines are drawn to the most extreme data value that lies within a distance of 1.5 times the interquartile range. Short braces mark the ends of these dotted lines.*

variety of modern species, we estimate the speed to be

$$\text{speed} = 0.49 \times (\text{stride length})^{1.67} \times (\text{footprint length})^{-1.17} \qquad (6.2)$$

where stride length and footprint length are in meters and the speed is in meters per second. The reader should note that we follow the majority here in defining stride length to be the distance between successive footprints. Alexander's definition is the distance between successive footprints of the same leg. The formula has been adjusted accordingly.

Inserting mean stride lengths and mean footprint lengths for the three species, we estimate that *Eubrontes* was crossing the site at a speed of about 8.1 kilometers per hour, which is a reasonable jogging pace. On the other hand, *Grallator* is estimated to have been traveling at a speed of 7.1 kilometers per hour, which is not much slower, despite the difference in sizes of the dinosaurs. The intermediate-sized *Anchisauripus* is estimated to have been the fastest of the three. Estimates for its speed are unreliable here because the paucity of tracks is compounded with uncertainty of identification of track 7 as *Anchisauripus*. However, the large estimate for the speed of *Anchisauripus* is due in great part to the large stride length of trackway 28 passing though the site in a northeasterly direction. Based upon equation (6.2) we can estimate the speed of this individual to be about 35 kilometers per hour. With such estimates, it would be very useful to be able to provide an error analysis. However, there are far too many systematic errors to regard these values as anything more than a rough indication. Not the least of the systematic errors is the necessity of using (6.2), which is based upon modern species, to describe dinosaurs.

6.2.3 Shape Analysis

If all size variables were to scale in a similar manner, then we might expect two dinosaurs whose stride lengths were in the ratio $a : b$ to have a similar ratio in leg length, footprint length, and speed. However, this is not the case for modern species and was almost certainly not the case for dinosaurs. Differences in scaling will normally be reflected in differences in shape. So it is shape variation, and in particular the relation of size to shape, that becomes of interest.

As mentioned in the scale analysis, two factors that influence stride length along a trackway are the size of the dinosaur and its speed. If we compare dinosaurs with similar Froude numbers, we find that the speeds of the individuals will be proportional to the square root of the size of the individual, assuming that Alexander's model for dinosaur speed is correct. This can noted from the exponents of formula (6.1). Thus, if we could group individuals with common Froude numbers together, a size variable such as leg length, for example, would be a scale variable for trackway dimensions. Within such groups, trackway shape distributions would be common, with

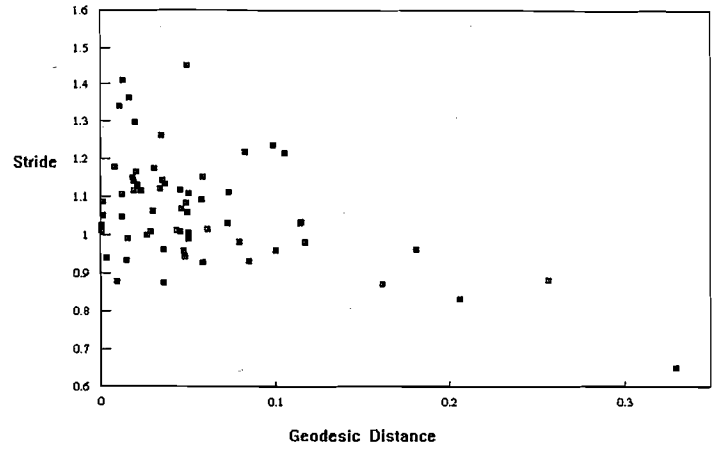

FIGURE 6.3. *Plot of geodesic distance versus stride length for* Eubrontes *triangles. Along each trackway, three successive footprints form a triangle. Some of the size and shape characteristics of these triangles are plotted above. Horizontally, a shape statistic is plotted that computes how close the triangle is to collinearity. This is measured as the geodesic distance in the sphere* $\mathbf{S}^2(1/2)$ *of triangle shapes from the great circle of collinear triangles to each shape point. Thus points plotted on the left-hand side of the figure represent nearly collinear triangles. On the vertical axis, the mean stride length of the three successive footprints is plotted in meters. The mean stride length can be defined as the average of two stride lengths: that from the first to second footprint, and that from the second to the third.*

different trackways having different scale factors. However, in actuality, a dinosaur would have had considerable variation in speed much as modern animals do. This extra variation in speed beyond scaling effects for dinosaur size would be expected to appear as a "stretching" effect along the trackway direction. Thus if we consider the shape of the triangle formed by three successive footprints, the effect of increased speed along the trackway would be to stretch the triangle closer toward the great circle of collinear triangles in shape space $\mathbf{S}^2(1/2)$.

Figure 6.3 illustrates this idea. The plotted points are statistics drawn from all triangles formed by taking three successive prints along *Eubrontes* trackways. On the horizontal axis is plotted the geodesic distance from the triangle shape to the great circle of collinearities (proportional to the absolute "latitude" taking the great circle of collinearities as the equator). On the vertical axis is the average of the two stride lengths of the footprint triangle, from first to second print and from second to third. As can be seen, the triangles at greater geodesic distances, which are closer to equilateral in shape, have smaller stride lengths. As the stride length is proportional

to estimated speed in our model, this means that triangles for which the estimated speed is greater tend to be flatter, as one would expect.

6.2.4 *Fitting the Mardia-Dryden Density*

To study the shape distribution in greater detail, we can fit a distribution such as the Mardia-Dryden density of formula (5.62) to the triangle shape data. One way to fit such a density to the data is by matching the centroid, or center of mass, of the Mardia-Dryden distribution to the centroid of the data. We recall that the shape space $\mathbf{S}^2(1/2)$ is naturally embedded in $\mathbf{R}^3$ as a sphere of radius $1/2$ centered at the origin. See formula (3.6). Thus the shapes of triangles of three successive footprints can be represented as points in $\mathbf{R}^3$. If $\sigma_1, \sigma_2, ..., \sigma_m$ are a set of triangle shapes represented as points in $\mathbf{R}^3$, then the centroid of these points will be

$$\bar{\sigma} = \frac{1}{m}\sum_{j=1}^{m}\sigma_j \tag{6.3}$$

which will lie within the interior of the sphere $\mathbf{S}^2(1/2)$. The vector

$$\frac{\bar{\sigma}}{2||\bar{\sigma}||} \tag{6.4}$$

lies once again on the sphere $\mathbf{S}^2(1/2)$ and can be interpreted as the *mean shape* of the data. This mean shape plays much the same role for the data that the shape σ_μ plays in the Mardia-Dryden density of formula (5.62). Correspondingly, the quantity $||\bar{\sigma}||$ provides a measure of how concentrated the data are about the mean shape. Its analog is not β of formula (5.62) as such, but rather

$$||\mathcal{E}(\sigma)|| = \frac{1}{2} + 4\beta^{-2}(1 - e^{-\beta/2}) - 2\beta^{-1} \tag{6.5}$$

The method of moments fit of the Mardia-Dryden density to a sample of triangle shapes is found by solving the equations

$$\sigma_\mu = \frac{\bar{\sigma}}{2||\bar{\sigma}||} \qquad ||\mathcal{E}(\sigma)|| = ||\bar{\sigma}|| \tag{6.6}$$

in the unknowns β and σ_μ.

This fitting technique is closely related to the fitting of spherical data by the Fisher distribution using maximum likelihood estimation. As Mardia and others have noted, the Mardia-Dryden density is one of a variety of densities on the sphere, including the Fisher density, the projected normal density, and the Brownian motion density, which while functionally different, form flexible families of densities that are very close to each other in shape. Like the Mardia-Dryden density, these families have a location

parameter analogous to σ_μ and have level curves for the density that are circles of points that are equidistant from the location parameter. They also include a concentration parameter that is analogous to β.

For the dinosaur trackways, there are two shape distributions associated with sets of three successive footprints, namely those with two left prints and one right, and those with two right prints and one left. For the modeling of triangles of successive footprints, we can pool the information by assuming bilateral symmetry and reflecting alternate triangle shapes along a trackway. This reflection is equivalent to multiplying w_3 by -1 in the coordinate system of formula (3.6). For many trackways, the confounding of prints with other trackways and the difficulty of distinguishing right and left prints still makes the task of pooling information from the two types of triangles problematic. However, for some trackways such as trackway 1, clear information seems to be available. In such cases, we can take triangles formed by two left prints and one right print as canonical, and reflect the shape distribution for every second triangle of three successive prints.

To encode the shapes of triangles, we take the first and third prints of any sequence of successive footprints as the base of the triangle for Bookstein coordinates. Thus in the notation of formulas (3.1) and (3.2), the point x_3 is in fact our middle footprint of the three. From the Bookstein coordinates we can encode the shape σ of each triangle using the spherical representation (w_1, w_2, w_3) of formula (3.6). For trackway 1, the tabulated shape data are as follows:

Triangle	Pattern	w_1	w_2	w_3
1	RLR	0.48647	0.02354	–0.11310
2	LRL	0.48641	0.00490	0.11569
3	RLR	0.49371	–0.00500	–0.07889
4	LRL	0.49909	0.03003	0.00316
5	RLR	0.49672	–0.03147	–0.04777
6	LRL	0.46722	0.02082	0.17684
7	RLR	0.47245	0.03951	–0.15882
8	LRL	0.48762	–0.07167	0.08422
9	RLR	0.49811	–0.03717	–0.02256

It can be seen that the sign of w_3 alternates in the table. So, we reflect the triangles of the RLR type. Fitting this to the Mardia-Dryden density gives us an estimate for σ_μ and for β, namely

$$\sigma_\mu = (0.49186,\ 0.00297,\ 0.08980) \qquad \beta = 452.5 \qquad (6.7)$$

The high value of β is indicative of the regularity of the footprint pattern, and supports the interpretation of Figure 6.2 that the *Eubrontes* of trackway 1 was moving at a fairly constant speed. The estimated mean shape σ_μ can be interpreted by noting its proximity and relationship to the shape $(0.5, 0, 0)$, which marks a triangle of three equally spaced collinear points.

6.3 Shape Analysis of Post Mold Data

6.3.1 A Few General Remarks

In this section, we apply some methods of shape analysis to the post mold data that we first considered in Section 1.4.3. In particular, we shall examine statistically the interpreted roundhouses of Aldermaston Wharf and South Lodge Camp as shown in Figures 1.5 and 1.6 respectively.

In 1973, an exhibition at the Institute of Contemporary Arts in London was entitled "Illusion in Art and Nature." It is interesting to note that one of the exhibits was a plan of an excavated Bronze Age settlement from Thorny Down in Wiltshire, England. The exhibit challenged people to interpret the configuration of post molds found at the site and to group them into recognizable patterns that would correspond to the original buildings on the site. See Gregory and Gombrich [76] for a discussion of the ambiguity of interpretation of this site in the context of the exhibit.

The reader is invited to examine Figure 6.4, which shows the layout of accepted post molds based upon J.F.S. Stone's excavation from 1937 to 1939. This figure should be studied in the light, say, of R. Wainwright's comment in *A Guide to the Prehistoric Remains in Britain* [175, p. 199] that the site contained the remains of nine circular houses called *roundhouses*. One strong indication of a roundhouse is in evidence. In other places, rough circles can be seen. However, these can equally well be interpreted as parts of structures that could conceivably have been rectangular rather than circular. The interpretation of nine buildings on the site follows directly from Stone's original report, which grouped the post molds into nine clusters. Stone found other evidence at the site such as the location of cooking holes and some pottery. Nevertheless, the post mold configuration provides most of the evidence for the number, location, and shape of the buildings.

The archeological interpretation of such sites is assisted by a certain amount of background knowledge of the cultures that were present. Thus Thorny Down has been interpreted in the light of prior knowledge that

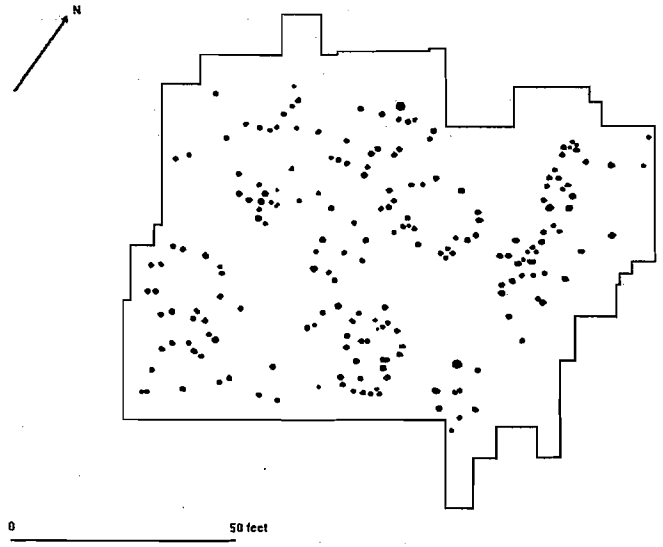

FIGURE 6.4. *Post mold arrangement at Thorny Down, Wiltshire. Thorny Down has achieved a certain amount of notoriety among archeologists for the ambiguity of its post mold pattern. In this picture, a large number of features, such as pits and cooking holes, have been eliminated so that the reader can judge the post mold evidence by itself. It has been claimed that there were nine roundhouses on the site. One is clearly evident from the picture. This figure has been redrawn from [165].*

Late Bronze Age peoples of Britain typically built roundhouses with posts that were spaced 1.6–2.2 meters apart. However, such knowledge, while of great assistance, can be misleading. For example, in the case of the post mold evidence at Thorny Down, such background knowledge can lead the researcher to interpret circular buildings in cases where the interpretation is weak. The eye is very good at interpreting patterns in chaotic pictures but is not always reliable in its interpretations.

6.3.2 The Number of Patterns in a Poisson Process

Suppose the researcher observes a point process, such as a post mold pattern, within a two-dimensional region. After studying the configuration of points, the researcher comes to believe that rather than being random, the particles of the process exhibit geometric regularity that cannot be explained by chance. For example, the particles could be arranged roughly in rectangles or circles, or perhaps have an approximate lattice structure. A null hypothesis that no structure exists, so that the perceived configurations arise by chance, could be modeled by a Poisson process or any other standard model for particles in the plane. Then the number of patterned configurations observed in the data can be compared with the expected number obtained by chance under the null hypothesis. We have already encountered one such example based upon configurations of straight lines when we studied the hypothesis of ley lines in the Land's End data. We now broaden the question to include the kinds of configurations that can appear in post mold interpretations.

Suppose that $X_1, X_2, ..., X_n$ are n random planar points for $n \geq 3$. For convenience, let us write $X = (X_1, ..., X_n)$. Similarly, we will write $x = (x_1, ..., x_n)$ for any realization of X. Let $\zeta(X)$ be a function of these points taking values in the set $\{0, 1\}$. We can think of $\zeta(X)$ as an *acceptance* function, which notes that X has a certain property by assigning the value $\zeta(X) = 1$ when X has the property and $\zeta(X) = 0$ when it does not. Now let us further suppose that $\zeta(X)$ is a function of these points only through their *shape* so that

$$\zeta(X_1, \ ..., \ X_n) \ = \ \zeta(aX_1 + b, \ ..., \ aX_n + b) \tag{6.8}$$

for any complex numbers a, b with $a \neq 0$. As we are not particularly interested in the labels of the points, but rather in their geometrical characteristics as a point set, we shall also suppose that ζ is a symmetric function of its arguments.

Next, we suppose that $w(X)$ is a nonnegative real valued function that is invariant under translations and rotations, and homogeneous under scale changes of the points. That is, w has the property that

$$w(aX_1 + b, \ ..., \ aX_n + b) = |a| \ w(X_1, \ ..., \ X_n) \tag{6.9}$$

The function w is typically a measure of the *size* of the configuration. As with the function ζ, we shall suppose that w is a symmetric function of its arguments.

Now suppose we observe a Poisson process in the plane throughout some planar region A, and decide to count the number of configurations X of n particles for which $\zeta(X) = 1$ and $w_0 \leq w(X) \leq w_1$. Some configurations that satisfy these conditions will lie entirely within the region A and will be observed, while other configurations outside the window will not be observed. Configurations that overlap, having some points within and some without, will not be observed. Let N be the number of configurations X of n particles observed within A such that $\zeta(X) = 1$. We shall be interested in finding the approximate distribution of N and making a comparison of this distribution with an actual count of configurations in a post mold pattern.

Let ∂A denote the boundary of A. The exclusion of configurations that overlap ∂A is understood as a *boundary effect* which is vanishingly small as A expands to fill the entire plane. Suppose the Poisson process has intensity ρ. Then N has expectation of the form given by the following proposition:

Proposition 6.3.1. *There exists a constant $c(\zeta)$ depending upon the choice of shape function ζ such that the expected value of the count statistic N is*

$$\mathcal{E}(N) \ = \ c(\zeta) \ \rho^n \ \left(w_1^{2n-2} - w_0^{2n-2}\right) \ \mathcal{V}_2(A) \tag{6.10}$$

if boundary effects are ignored.

Proof. For convenience in this proof, we shall assume that $w(x)$ is the diameter of x. We perform a transformation of variables from

$$(x_1, ..., x_n) \ \leftrightarrow \ (x_1, w, \theta, \sigma) \tag{6.11}$$

where $w = w(x)$; the angle $\theta \in \mathbf{S}$ is the direction of the vector $x_2 - x_1$, defined except when $x_1 = x_2$; and $\sigma \in \Sigma_2^n$ is the shape of $x_1, ..., x_n$. Then we can factorize the geometric measure on x as

$$d\mathcal{V}_{2n}(x) = w^{2n-3} \ f(\sigma) \ d\mathcal{V}_2(x_1) \ d\mathcal{V}_{2n-4}(\sigma) \ d\mathcal{V}_1(\theta) \ dw \tag{6.12}$$

For the general theory of such factorizations, the reader is referred to [4, 5]. When $X_1, ..., X_n$ are IID uniform in A then X is uniformly distributed in $A^n = A \times ... \times A$. Let $W = w(X)$. Then

$$\mathcal{E}[\zeta(X) 1_{(w_0 \leq W \leq w_1)}] = \frac{1}{[\mathcal{V}_2(A)]^n} \int_{A^n} \zeta(x) 1_{(w_0 \leq w \leq w_1)} \ d\mathcal{V}_{2n}(x) \tag{6.13}$$

Applying the factorization of formula (6.12), integrating over the variables w, θ, and x_1, and ignoring the boundary effects of configurations X that

lie within a distance of w_1 from ∂A, we see that the expectation becomes

$$\mathcal{E}[\zeta(X)1_{(w_0 \le W \le w_1)}] = \frac{\pi[w_1^{2n-2} - w_0^{2n-2}]}{(n-1)[\mathcal{V}_2(A)]^{n-1}} \int_\Sigma \zeta(\sigma) f(\sigma)\, d\mathcal{V}_{2n-4}(\sigma) \quad (6.14)$$

We shall continue to ignore boundary effects in subsequent formulas. Now suppose that $m > n$ particles are uniformly and independently scattered throughout A. Then the expected number of configurations X of n particles among the m that satisfy the shape condition $\zeta(X) = 1$ and size condition $w_0 \le w(X) \le w_1$ is

$$\mathcal{E}(N) = \binom{m}{n} \frac{\pi[w_1^{2n-2} - w_0^{2n-2}]}{(n-1)[\mathcal{V}_2(A)]^{n-1}} \int_\Sigma \zeta(\sigma) f(\sigma)\, d\mathcal{V}_{2n-4}(\sigma) \quad (6.15)$$

We take a Poisson limit by letting $m \to \infty$ and letting A expand to fill the plane so that $m/\mathcal{V}_2(A) \to \rho$ and $\mathcal{V}_1(\partial A)/\mathcal{V}_2(A) \to 0$. Then the expected number of configurations of n particles is seen to be asymptotically

$$\frac{1}{n!}\,[w_1^{2n-2} - w_0^{2n-2}]\,\rho^n\,\frac{\pi \mathcal{V}_2(A)}{(n-1)} \int_\Sigma \zeta(\sigma) f(\sigma)\, d\mathcal{V}_{2n-4}(\sigma) \quad (6.16)$$

We get the required formula by setting

$$c(\zeta) = \frac{\pi}{n!(n-1)} \int_\Sigma \zeta(\sigma) f(\sigma)\, d\mathcal{V}_{2n-4}(\sigma) \quad (6.17)$$

which completes the proof. Q.E.D.

We usually wish to know more than simply the expected value of N. The Poisson approximations of Silverman and Brown [153] are useful to determine this. They show that under certain asymptotic conditions described in [153] the distribution of N is from the Poisson family. As the expected value can be approximated by Proposition 6.3.1 above, the approximate distribution of N can be specified. These asymptotics appear to be quite reasonable for post mold investigations.

Two comments on Proposition 6.3.1 should be made. As mentioned above, formula (6.12) is a special case of a factorization calculus in which geometric measures can be shown to factorize in location, scale, orientation, and shape components. See [3, 4, 5] for work by Ambartzumian and colleagues on these factorizations. Note that we can pull out a shape measure from the factorization. This corresponds to our function f in (6.12) above. The resulting shape measure might be thought of as canonical. However, upon closer examination, it is seen to depend upon the choice of size function. As we have observed on previous occasions, shape constructions cannot be fully separated from the definitions of size variables that are used in the standardization of data sets.

The second comment to be made on this result is the presence of some potentially high exponents in formula (6.10). When modeling particle scatterings as Poisson processes, we typically have to estimate the intensity ρ. Now, if we were investigating the presence of roundhouses of eight posts, say, and were to underestimate the intensity of the scattering by twenty percent, then we would underestimate the expected number of circular configurations that could be explained as chance by a factor about 0.168. One way to underestimate the value of ρ is to assume that the region A is the region of excavation as marked by the bold line in Figure 6.4. The natural estimate for ρ is then the average number of post molds per square meter across the region of excavation. An examination of Figure 6.4 shows that the post molds are not homogeneously scattered across the region of excavation. An improvement on this assumption is to suppose that the post molds are homogeneously scattered across some subregion of the region of excavation that we can call the *region of post mold activity.* This may explain why simulation studies such as those described in [44] have found a larger number of circles at Thorny Down than can be expected from a Poisson scatter over the region of excavation. Unlike the region of excavation, the region of post mold activity is unknown. Therefore we cannot directly find its area as a way of estimating ρ. Fortunately, other techniques are available to estimate ρ. One method is to fit the theoretical distribution of distance from a typical post mold to its nearest neighbor.

6.3.3 An Annular Coverage Criterion for Post Molds

An acceptance criterion that has been popular among archeologists studying and simulating post mold patterns is the *coverage criterion.* For example, Cogbill [44] and Litton and Restorick [109] have searched for patterns in post mold data by moving a set across the region of excavation. Cogbill searched for roundhouses by running an annulus, or circular ring, with fixed inner and outer radius over the post mold points. Annuli consisting of all points $y \in \mathbf{R}^2$ such that

$$w^2 \le ||y - a||^2 \le w^2(1+\epsilon)^2 \quad (6.18)$$

were used, where $\epsilon > 0$ and $w > 0$ are constants controlling the thickness and inner radius of the annulus, respectively. The constant $a \in \mathbf{R}^2$ controls the location of the annulus, and was allowed to vary as the annulus was shifted over the region of post molds. Any set of n post molds that could be covered by an annulus for some choice of a would be declared an acceptable configuration and considered as a potential roundhouse. See Figure 6.5.

In view of Proposition 6.3.1 we naturally seek an approximation to the distribution of the number of sets of n post molds that satisfy the annular criterion of Cogbill under the hypothesis that the post molds are scattered

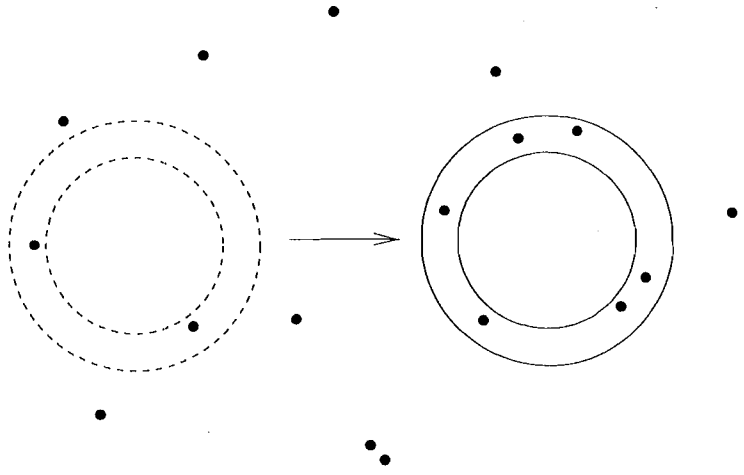

FIGURE 6.5. *The annular criterion for accepting a configuration of post molds. Cogbill [44] and others have proposed the detection of circular post mold configurations by running an annulus across the region where post molds occur. If a sufficiently large number of post molds can be covered in a given position of the annulus, these post molds are accepted as a possible roundhouse.*

as a Poisson process. We need only find the expected number of circular configurations under the annular criterion and then apply the Poisson limit theorem of Silverman and Brown.

Suppose C is some subset of the plane $\mathbf{R}^2$. We represent the translates of C as

$$C(a) = \{y + a : y \in C\} \tag{6.19}$$

for each $a \in \mathbf{R}^2$. Let $X_1, X_2, ..., X_n$ be IID uniformly distributed in the region A. What is the probability that there exists a translate $C(a)$ such that $X_j \in C(a)$ simultaneously for all $j = 1, ..., n$? That is, what is the probability that C can be translated to completely cover *all* the points $X_1, ..., X_n$? The solution to this problem will allow us to find the analogous expectation for Poisson processes.

Mack [111] has solved this problem, not only for subsets of the plane, but also for the general-dimensional problem. Using the terminology of [111] we let $\mathcal{Q}_n(n)$ be the probability that n independent uniformly distributed particles in $A \subset \mathbf{R}^p$ can be covered by a translate of a given subset $C \subset \mathbf{R}^p$. We assume that $\mathcal{V}_p(C) \ll \mathcal{V}_p(A)$ and that boundary effects of A are ignored. Then $\mathcal{Q}_n(n)$ has the general form

$$\mathcal{Q}_n(n) = n \left[1 + \sum_{j=1}^{p-1} b_j (n-1)(n-2)...(n-j) \right] \left[\frac{\mathcal{V}_p(C)}{\mathcal{V}_p(A)} \right]^{n-1} \tag{6.20}$$

for some choice of the constants $b_1, b_2, ..., b_{p-1}$. These constants can be evaluated by calculating $\mathcal{Q}_2(2), ..., \mathcal{Q}_p(p)$ directly.

For our particular application, we have $p = 2$ and a family of annuli given by formula (6.18). It is easy to check that

$$\mathcal{Q}_2(2) \approx \frac{4\ \pi\ w^2}{\mathcal{V}_2(A)} \tag{6.21}$$

for small $\epsilon > 0$. Of course $\mathcal{Q}_1(1) = 1$. Solving for b_1 by evaluating $\mathcal{Q}_2(2)$ in formula (6.13) we obtain

$$\mathcal{Q}_n(n) \approx n\ [\epsilon + (n-1)]\ (2\pi)^{n-1} \left[\frac{w^2}{\mathcal{V}_2(A)} \right]^{n-1} \epsilon^{n-2} \tag{6.22}$$

again for small $\epsilon > 0$. We can check this formula by calculating $\mathcal{Q}_3(3)$ directly without appeal to formula (6.22). Using the transformations of Section I.2.3 of [147, pp. 16–17] and integrating over formula (2.18) of [147] we can show that the probability the radius of the circumcircle through X_1, X_2, X_3 is less than or equal to some value w is $6\pi^2 w^4 / [\mathcal{V}_2(A)]^2$ for large regions A, ignoring the boundary effects. For small $\epsilon > 0$, our probability $\mathcal{Q}_3(3)$ is approximately the probability that the radius of this circumcircle is between w and $w(1+\epsilon)$. This reduces to

$$\mathcal{Q}_3(3) \approx \frac{24\ \pi^2\ w^4\ \epsilon}{\mathcal{V}_2(A)^2} \tag{6.23}$$

Formula (6.23) can be seen to agree with (6.22) for $n = 3$ to first order in $\epsilon > 0$.

We are now in a position to write out the formula for circles in a Poisson process.

Proposition 6.3.2. *In a Poisson process of intensity ρ within a window A the expected number $\mathcal{E}(N)$ of circular arrangements of $n \geq 2$ particles under an annular coverage criterion with annuli of the form given in (6.18) is*

$$\mathcal{E}(N) \approx \frac{(2\pi)^{n-1}}{(n-2)!}\ \rho^n\ w^{2n-2}\ \epsilon^{n-2}\ \mathcal{V}_2(A) \tag{6.24}$$

to lowest order approximation in $\epsilon > 0$.

Proof. This is a straightforward consequence of (6.15), taking a Poisson limit as A expands to fill the plane and the number of particles m goes to infinity so that $m/\mathcal{V}_2(A) \to \rho$. Q.E.D.

The similarity between the formulas of Propositions 6.3.1 and 6.3.2 can be seen. For small $\epsilon > 0$, the annular coverage criterion factorizes into shape and size criteria that relate Proposition 6.3.2 to 6.3.1. Note also that the Poisson limiting distribution for N holds here as well.

We are now in a position to try such methods on post mold data sets. While the post molds at Thorny Down represent one of the most famous examples of ambiguous interpretations, the simplicity of the configurations at Aldermaston Wharf in Figure 1.5 and South Lodge Camp in Figure 1.6 of Section 1.4.3 make them better starting points for analysis.

6.4 Case Studies: Aldermaston Wharf and South Lodge Camp

Before it was discovered, Aldermaston Wharf was heavily plowed. Thus it can be reasonably assumed that some of the post mold evidence was destroyed by plowing. As the evidence is quite fragmentary, it is necessary to assess the strength of the interpreted roundhouses as carefully as possible. See Figure 1.4. The shaded regions are features of the site that are later than the time of the Late Bronze Age settlement. The irregular unshaded regions represent pits at the site. The archeological report on Aldermaston Wharf can be found in Bradley and Fulford [30].

The site at South Lodge Camp, shown in Figure 1.5, was re-excavated, and reported by Barrett et al. [9]. A number of buildings were identified and labeled A through D, with varying degrees of geometric regularity in the post mold evidence.

6.4.1 Scale Analysis

For a Poisson process of intensity ρ, the median distance from any particle to its nearest neighbor is $\sqrt{(\ln 2)/(\pi\rho)}$. This suggests that we estimate ρ at Aldermaston Wharf by computing the median nearest neighbor distance and solving for ρ. The median distance from any post mold to its nearest neighbor is 1.7 meters. So the intensity of the post mold scattering at Aldermaston Wharf is estimated to be $\hat{\rho} = 0.076$ post molds per square meter. A total of $m = 61$ post molds are scattered throughout the region, suggesting an area of post mold activity of $m/\hat{\rho} = 802.6$ square meters. This is considerably less than the area of excavation, which was approximately 2000 square meters. From observations at other sites, roughly contemporary with Aldermaston Wharf and South Lodge Camp, we would expect neighboring posts of buildings to be within 1.6 to 2.2 meters of each other. Counting replacement posts and the exceptional larger distance, we would expect posts belonging to a common building to be less than three meters apart. Figure 1.4 shows all post molds that satisfy this joined by a link. Neither interpreted building I nor interpreted building II is clearly defined by this linkage method. However, rough circles can be made out for both I and II, with the strength of the circular interpretation being somewhat vague. We shall examine these circles more carefully below in the shape analysis.

At South Lodge Camp, a total of $m = 71$ post molds were recorded across the area of excavation, and the median nearest neighbor distance was found to be 1.4 meters. Using the same procedure for estimating ρ as was used for Aldermaston Wharf, we obtain an estimate $\hat{\rho} = 0.112$ post molds per square meter for a Poisson process with the same median nearest neighbor distance. In turn, this allows us to estimate the area of the region of post mold activity to be $m/\hat{\rho} = 634$ square meters. Again, this is considerably less than the area of excavation, which is about 1600 square meters. Linking post molds that are within three meters of each other we see that interpreted structures B, C, and D become clearly defined, with circles evident in D and C. Structure A is less clearly defined by this linking method.

6.4.2 Shape Analysis

The report on Aldermaston Wharf by Bradley and Fulford [30] interpreted two roundhouses, which are labeled I and II in Figure 1.4. Of the two interpreted structures, the second has the stronger visual evidence of a circular configuration. A total of 6 to 8 post molds can be interpreted as possible locations for posts of a roundhouse wall. There is some evidence that on the east side of Structure II a post mold could be missing because of the presence of later features. If this is the case, an interpretation with 6 posts as in Figure 6.6 would be appropriate. We assess the fit to a circle by choosing annuli that cover the configurations having the smallest possible area. Structure II can be covered by an annulus whose inner radius is 3.66 meters and whose outer radius is 3.95 meters. According to formula (6.24) the expected number of configurations of six particles in a Poisson process of intensity $\rho = 0.076$ throughout an area of 803 square meters is 1.07. Thus such a circular configuration can be considered plausible on chance considerations alone. Structure I is cruder than Structure II with an even higher value for ϵ. The six post molds of Structure I shown in Figure 17 can be covered by an annulus with inner radius 3.15 meters and outer radius 3.62 meters. The expected number of such configurations over 803 meters for an equivalent Poisson process is 3.03.

At South Lodge Camp, Cluster D contains a circle of eight post molds as in Figure 6.6, and can be covered by an annulus of inner radius 3.95 meters and outer radius 4.21 meters. In a Poisson process of intensity $\rho = 0.112$ the number of circular configurations of eight particles that can be covered by such an annulus within a region of area 634 square meters is 0.15. Thus the circular configuration of Cluster D is more unlikely than either Structure I or Structure II at Aldermaston Wharf. Cluster C also contains a circle of seven post molds. This can be fit by an annulus with inner radius of 2.13 meters and an outer radius of 2.36 meters. The expected number of such configurations is 0.00096. The reader may find it a bit

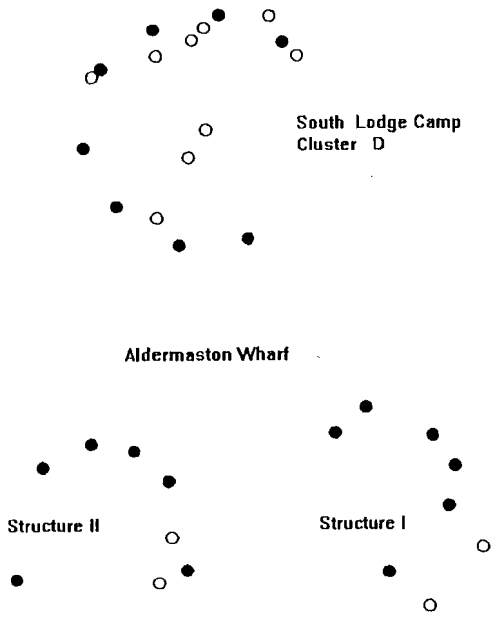

FIGURE 6.6. Roundhouse Interpretations at South Lodge Camp and Aldermaston Wharf

unusual that configurations with the fit of Cluster C are much more rare than those of Cluster D in a Poisson process, as the eye suggests otherwise. The eye also picks up the symmetry of spacing of the post molds in Cluster D as a component of the regularity. The main difference that explains the calculations is that Cluster C has a circle of smaller radius. An examination of formulas (6.10) and (6.24) shows that small configurations are much rarer than large ones for a Poisson process. A hint of a circle can be seen in Cluster A of South Lodge Camp. However, the configuration is very weak.

6.4.3 Conclusions

The case studies and formulas have not provided a clear decision procedure that would allow as to accept or reject an interpretation of a roundhouse in a post mold data set. However, they do provide us with a quantitative tool for assessing the strength of a circular configuration relative to other configurations at the same site and relative to configurations at other sites. It is perhaps the latter that is more important. The interpretation at some famous sites such as Thorny Down is problematic, whereas the interpretation at sites such as South Lodge Camp is much more straightforward. Archeologists who can supplement their post mold analysis by comparing it quantitatively with other sites can evaluate the strength of their conclusions in the context of what is known about Late Bronze Age sites. For example, we can conclude that the interpretation at Aldermaston Wharf is tentative at best, with an interpretation that is weaker than that for South Lodge Camp.

6.5 Automated Homology

6.5.1 Introduction

In this section, we shall describe an automated homology routine developed by Michael Lewis as part of his Ph.D. work at the University of Waterloo. Up until this point we have represented various shapes by assuming that they are naturally homologous or that homologous landmarks can be selected from the data, as in the case of the brooch images of Chapter 1. However, in many image data sets there are no obvious features that stand out from the rest of the image to the extent that we would wish to label them as landmarks. We would rather seek to find a mapping from each image to any other that maps each point on the image to its homologous point on the other image.

The problem of constructing a homology between images is closedly connected to the *correspondence problem* in computer vision, in which one has two images, each in two dimensions, of a three-dimensional object seen from two angles. If it is known which points in the two images are different views

of the same point in three dimensions, the images are said to be *registered.* If there is no aspect of the object that is visible in one image but not in the other, then the points in the images are homologous, with corresponding points between images being homologous if they are projections of the same point of the three-dimensional object. The images will differ in shape slightly because of the two aspects from which the object is viewed. The main difference between the correspondence problem of computer vision and the automated homology problem of shape analysis is that the shape differences of the latter are assumed to be completely general in nature, and not necessarily produced by transformations such as projectivities between images. See Besl and McKay [11] for some work on the problem of registering images based upon three-dimensional shapes.

Closer to the automated homology that we seek are the algorithms of Grenander and Miller [77]. The approach of Grenander and Miller is part of a larger program of interpretation and representation of complex images using the Pattern Theory pioneered by Ulf Grenander, and briefly surveyed in that paper. We shall examine the similarities and differences between the methods later.

For shape analysis, suppose that we have a collection of images of different objects, say images of brooches or perhaps images of faces, that vary slightly but not excessively in shape. Let us assume that in all images we are looking at essentially the same type of object, so that between any two images an approximate homology can in principle be established. For the purposes of analysis, we perform a rough standardization on the images so that all images have the same dimensions in pixel units, and so that each object within the image is centered and standardized in terms of orientation and scale. This last requirement is not required to be accomplished with careful Procrustean matching. Rather we will assume that homologous points between images are separated by small distances compared to the dimensions of the images. To simplify further, we also suppose that the images are grayscale or dithered black and white pictures such as can be produced by many image viewers.

6.5.2 *Automated Block Homology*

To describe automated homology, let us consider two images. Suppose that we wish to establish a homology from one to the other. This should be a function defined on all the pixel locations of one image and mapping to the pixel locations of the other. However, in practice we could choose a smaller set of locations by superimposing a rectangular lattice of points over each image. Equivalently, we can partition the images into blocks and suppose that these lattice points are the centers of the blocks. Our task is then to construct a correspondence between the lattice points that most nearly corresponds to the homology between the images. Consider a function $h = (h_1, h_2)$ that maps a point at the jth row and kth column

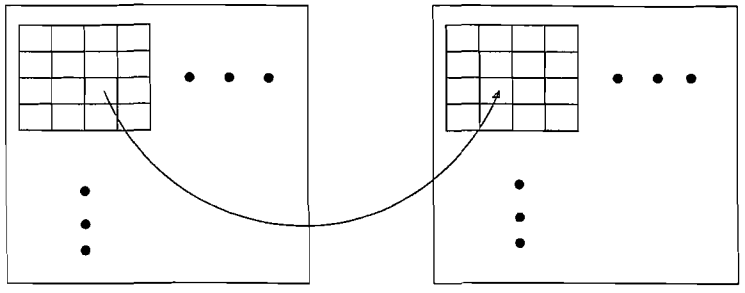

FIGURE 6.7. *Block assignment for automated homology. In order to construct an automated homology between images, the images are transformed to matrices and then divided into blocks. A first step in the construction of the homology is to find a mapping between the blocks in the first image and those in the second so that a measure of discrepancy $\mathcal{W}$ is minimized.*

of one image to the $h_1(j,k)$th row and $h_2(j,k)$th column of the other as in Figure 6.7. Let $\mathcal{W}[j,k;l,m]$ be a measure of mismatch of the homology between location (j,k) of the first image and location (l,m) of the second image. To construct the best homology from the lattice on one image to the lattice of the other we can minimize

$$\sum_j \sum_k \mathcal{W}[j,k\ ;\ h_1(j,k), h_2(j,k)] \tag{6.25}$$

over all functions h. We do not require that h be a 1-1 function, as this is much too restrictive for our purpose.

The next step that needs to be considered is the construction of the mismatch function $\mathcal{W}[j,k;l,m]$. To construct such a function, the images must be placed into an environment in which they can be quantitatively compared. A variety of packages are available at the time of writing to assist in the analysis. The description that follows represents an approach found useful in the analysis of the Iron Age brooches of Figures 1.1 and 3.7.

The images are first transformed to matrices of real numbers by conversion to ASCII format. For a grayscale image, the entries in the matrix will denote the degree of darkness at a particular pixel location. For a dithered image, the matrix will consist of entries of zeros and ones corresponding to black and white pixel values. A standard tool for conversion of an image to a matrix is the XV viewer available on X-windows terminals and the UNIX operating system. The matrices can then read into MATLAB, which provides special tools for the manipulation of matrices. Each matrix representing an image can then be subdivided into blocks of size $p \times p$, say. We can think of the lattice points (j,k) as being centered in the middle of these blocks so that the mismatch $\mathcal{W}[j,k;l,m]$ between lattice point (j,k) in the first image and (l,m) in the second is interpreted as

a measure of mismatch between their corresponding blocks. Thus we shall seek a function that measures the mismatch between block (j,k) in the first matrix and block (l,m) in the second. Suppose the matrix of the first image is divided up into blocks (A_{jk}) where each A_{jk} is *itself* a $p \times p$ matrix. Similarly, let us suppose that the matrix from the second image is divided up in blocks (B_{lm}) that are also of the same dimension. Now, a function h mapping block A_{jk} to block B_{lm} is a candidate for a homology between the images. To measure the discrepancy between the images, we can find some measure of distance between the matrices A_{jk} and B_{lm}. If the entries of these matrices are zeros and ones a suitable measure of mismatch could be obtained by counting the number $||A_{jk} - B_{lm}||$ of discordant entries between them. More generally, since A_{jk} and B_{lm} are each $p \times p$ matrices, we can regard them as vectors in $\mathbf{R}^{p \times p}$. The distance $||A_{jk} - B_{lm}||$ can then be taken as Euclidean distance between these vectors.

An optimization algorithm can then be run using this choice of $\mathcal{W}$. However, it should be noted that such an automated homology algorithm has no respect for the natural topology of the image. Points that are the centers of neighboring blocks in the first image could be mapped by the optimal h to opposite sides of the second image. To counter this trend we can introduce another term to the formula for $\mathcal{W}[j,k;l,m]$ that measures the distance between the coordinates. Thus our formula for $\mathcal{W}[j,k;l,m]$ becomes

$$\lambda_1 \mathcal{T}_1 \left\{ (||A_{jk} - B_{lm}||^2 \right\} + \lambda_2 \mathcal{T}_2 \left\{ [(j-l)^2 + (k-m)^2] \right\} \qquad (6.26)$$

for appropriate weights $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$, and suitable nondecreasing functions $\mathcal{T}_j : \mathbf{R}^+ \to \mathbf{R}^+$. The functions $\mathcal{T}_1$ and $\mathcal{T}_2$ can be chosen to be the identity functions. However, in many cases it seems reasonable to rule out transformations h that will effect too radical a distortion of the image. This can be accomplished by making $\mathcal{T}_1$ and $\mathcal{T}_2$ increase faster than linearly. To restrict the search to transformations that perturb points to a maximum distance of $\epsilon > 0$, say, a reasonable choice for $\mathcal{T}_2$ would be one for which $\mathcal{T}_2(x) = \infty$ when $x > \epsilon$. If λ_2 is large compared to λ_1 then the optimal homology found will be inelastic because transformations close to the identity transformation will be favored. On the other hand, when λ_1 is large compared to λ_2 the algorithm will allow more drastic transformations to match features in the images.

Having established a preliminary correspondence between blocks in two images, we still have to construct a full homology between the images by smoothly extending from the correspondence between the centers of the blocks to the rest of the image. We begin by assigning a vector v_{jk} to the center x_{jk} of each block A_{jk} pointing from x_{jk} to the center of the corresponding block $B_{h(j,k)}$ in the second image. Thus v_{jk} can be regarded as a vector field on the lattice of centers x_{jk} of blocks. The vector field is smoothed across the entire image by assigning a vector to each point x in the first image whose value is a weighted average of neighboring lattice vectors v_{jk}. This smoothing can be performed using a Gaussian kernel. Many other choices are also reasonable. Associated with the vector field $v(x)$ on the first image is a transformation h' from the first image to the second that has this vector field as its field of difference vectors. So, we can write $h'(x) = x + v(x)$.

We could stop at this point and let this transformation h' be the required homology. However, this construction, while transforming the shape of the first image closer towards the second, still tends to be too rough an approximation to the desired homology to be suitable for shape analysis. This would appear to be due to the coarseness of the block size $p \times p$, which has to be sufficiently large to allow the discrimination of salient features in the images. To compensate for this, we can allow the features of the first image to undergo a small displacement *in the direction of the initial homology* h'. This is achieved by shrinking the displacement vector so that a point x is mapped to the new point $x + \epsilon v(x)$ for $\epsilon > 0$. The choice of $\epsilon = 1$ represents the full transformation by h'. However, this transformation is too drastic in some cases. For these cases, a more modest perturbation is preferred with some $\epsilon < 1$. The perturbed image then replaces the original first image and is partitioned into blocks. A homology h'' is then constructed in a similar fashion between the perturbed first image and the second image. This procedure continues with h''', h'''', ... until a satisfactory transformation (the composition of the small perturbations) from the original first image to the second image has been achieved.

6.5.3 An Application to Three Brooches

Michael Lewis has developed and implemented these techniques for an automated block homology routine. In Figure 6.8 we see an automated homology between the three brooches of Figure 1.1. There are a number of weighting factors and choices to be made by the researcher, such as the choice of λ_1 and λ_2 above. A certain amount of standardization of the images must be done in advance. This need not be very precise, and can be incorporated into the automatic search procedure. However, it is advisable to have the researcher interacting with the procedure at this level as well, as the fitting is accomplished with any computer interpretation of the images as a whole. Thus the algorithm is a compromise between the expert selection and spline interpolation methods of Bookstein and a fully automated procedure that would be expected in computer vision. The latter is typically context sensitive. Perhaps a more precise description of the method would be to call it computer assisted homology.

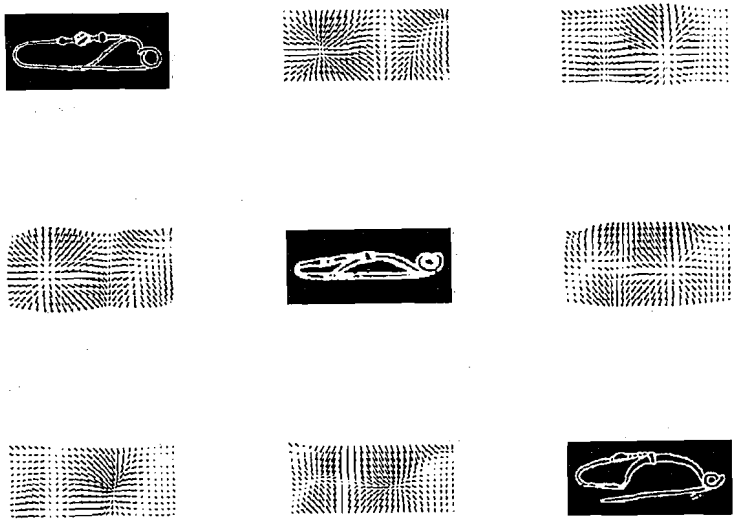

FIGURE 6.8. *Automated homology for three Iron Age brooches. The images of the three brooches can be seen along the diagonal of the* 3×3 *matrix of diagrams. Off the main diagonal, in the* (j,k)*th position, is a* quiver diagram *for the vector field of displacements for the homology that attempts to transform the* j*th brooch into the* k*th brooch. The vectors displayed are not the displacement vectors described above, but rather rescaled versions of these, shrunk for convenience of graphical presentation. The* 3×3 *matrix of images has a natural antisymmetry property: the sum of the vector fields in the* (j,k)*th and* (k,j)*th positions is zero. Blocks of size* $p \times p = 16 \times 16$ *were used in the partition of the images.*

6.6 Notes

This chapter has considered a few examples of the application of shape analysis. While the theme of shape is common to all examples, the methodologies used are quite diverse. This is even more the case for the entire range of applications in the literature. For this reason, a single chapter of applications cannot do justice to the variety of techniques and examples. The reader who is interested in particular applications will find, grouped by topic below, some references that can serve as a starting point for the exploration of the literature. I have taken the liberty of including examples in which shape plays an important role without necessarily being the primary topic of concern. Such applications leave room for future work involving the theory of shape.

6.6.1 Anthropology, Archeology, and Paleontology

[1], [2], [7], [9], [10], [25], [26], [30], [31], [32], [33], [44], [45], [58], [59], [95], [106], [107], [109], [110], [129], [130], [131], [133], [155], [165], [175], [177].

6.6.2 Biology and Medical Sciences

For a more complete list of biomedical applications, the reader is referred to the references given in [24].

[10], [13], [14], [15], [16], [17], [18], [19], [20], [21], [24], [25], [26], [27], [45], [48], [49], [50], [62], [63], [68], [69], [85], [106], [107], [108], [126], [131], [142], [143], [145], [152], [169], [170], [172], [176], [181], [182].

6.6.3 Earth and Space Sciences

[28], [29], [37], [42], [110], [118], [130], [173], [185].

6.6.4 Geometric Probability and Stochastic Geometry

[3], [4], [5], [6], [29], [30], [33], [46], [47], [89], [91], [93], [95], [96], [97], [100], [101], [102], [103], [111], [112], [118], [119], [128], [135], [141], [147], [148], [153], [154], [155], [157], [159], [160], [161], [166], [167], [171], [178], [179], [183].

6.6.5 Industrial Statistics

[11], [40], [54], [84].

6.6.6 *Mathematical Statistics and Multivariate Analysis*

[8], [35], [38], [52], [53], [55], [56], [57], [60], [66], [67], [70], [71], [72], [74], [75], [80], [81], [86], [88], [89], [90], [91], [92], [93], [94], [95], [98], [99], [101], [102], [103], [104], [105], [111], [112], [113], [114], [116], [117], [123], [124], [134], [139], [140], [144], [150], [151], [153], [154], [159], [166], [168].

6.6.7 *Pattern Recognition, Computer Vision, and Image Processing*

[11], [36], [41], [77], [83], [148].

6.6.8 *Stereology and Microscopy*

[47], [48], [49], [50], [62], [147], [160], [161], [180], [181], [182].

6.6.9 *Topics on Groups and Invariance*

[12], [36], [120], [127], [132], [136], [137], [138], [156], [162], [184].

Bibliography

[1] Alexander, R. M. "Estimates of speeds of dinosaurs." *Nature* 261 (1976), 129–130.

[2] Alexander, R. M. "How dinosaurs ran." *Scientific American* 264 (1991), 130–136.

[3] Alexandrian, A. R. "Random quadrangular shapes by factorization." In *Stochastic Geometry, Geometric Statistics, Stereology*, Teubner-Texte zur Mathematik 65, ed. R. V. Ambartzumian and W. Weil. Teubner, Leipzig (1984), 7–13.

[4] Ambartzumian, R. V. "Random shapes by factorization." In *Statistics in Theory and Practice*, ed. B. Ranneby. Umea: Section of Forest Biometry, Swedish University of Agricultural Sciences (1982), 35–41.

[5] Ambartzumian, R. V. *Factorization Calculus and Geometric Probability.* Cambridge University Press, Cambridge (1990).

[6] Baddeley, A. "Stochastic geometry: an introduction and reading list." *Int. Statist. Rev.* 50 (1982), 179–193.

[7] Barker, P. A. "An exact method for describing metal weapon points." In *Computer Applications in Archaeology*, Proc. Annual Conf. Computer

Centre, University of Birmingham (1975), 3–8.

[8] Barndorff-Nielsen, O. E., Blaesild, P., and Eriksen, P. S. *Decomposition and invariance of measures, and statistical transformation models.* Springer Lecture Notes in Statistics 58, Springer-Verlag, New York (1989).

[9] Barrett, J., Bradley, R., Bowden, M., and Mead, B. "South Lodge after Pitt Rivers." *Antiquity* 57 (1983), 193–204.

[10] Benson, R. H., Chapman, R. E., and Siegel, A. F. "On the measurement of morphology and its change." *Paleobiology* 8 (1982), 328–339.

[11] Besl, P. J. and McKay, N. D. "A method for registration of 3-d shapes." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 14 (1992), 239–256.

[12] Bondesson, L. "Characterizations of probability laws through constant regression." *Z. Wahrscheinlichkeits. verw. Gebiete* 30 (1974), 93–115.

[13] Bookstein, F. L. *The measurement of biological shape and shape change.* Lecture Notes in Biomathematics 24. Springer-Verlag, New York (1978).

[14] Bookstein, F. L. "Linear machinery for morphological distortion." *Comput. Biomed. Res.* 11 (1978), 435–458.

[15] Bookstein, F. L. "Foundations of morphometrics." *Ann. Rev. Ecol. Syst.* 13 (1982), 451–470.

[16] Bookstein, F. L. "A statistical method for biological shape comparisons." *J. Theor. Biol.* 107 (1984), 475–520.

[17] Bookstein, F. L. "Tensor biometrics for changes in cranial shape." *Ann. Human Biol.* 11 (1984), 413–437.

[18] Bookstein, F. L. "Transformations of quadrilaterals, tensor fields, and morphogenesis." In *Mathematical Essays on Growth and the Emergence of Form*, University of Alberta Press, Edmonton (1985), 221–265.

[19] Bookstein, F. L. "Size and shape spaces for landmark data in two dimensions (with discussion)." *Statist. Sci.* 1 (1986), 181–242.

[20] Bookstein, F. L. "From medical images to the biometrics of form." In *Information Processing and Medical Imaging*, ed. S. L. Bacharach. Nijhoff, Dordrecht (1986), 1–18.

[21] Bookstein, F. L. "Soft modeling and the measurement of biological shape." In *Theoretical Empiricism: A General Rationale for Scientific Model-Building*, ed. H. Wold. Paragon Press, New York (1987).

[22] Bookstein, F. L. "Discussion of 'A survey of the statistical theory of shape' by D. G. Kendall." *Statist. Sci.* 4 (1989), 99–105.

[23] Bookstein, F. L. "Principal warps: thin-plate splines and the decomposition of deformations." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11 (1989), 567–585.

[24] Bookstein, F. L. *Morphometric Tools For Landmark Data: Geometry and Biology.* Cambridge University Press, Cambridge (1991).

[25] Bookstein, F. L., Chernoff, B., Elder, R., Humphries, J., Smith, G., and Strauss, R. *Morphometrics in Evolutionary Biology: The Geometry of Size and Shape Change with Examples from Fishes.* Academy of Natural Sciences of Philadelphia, Philadelphia (1985).

[26] Bookstein, F. L. and Reyment, R. A. "Microevolution in Miocene *Brizalina* (Foraminifera) studied by canonical variate analysis and analysis of landmarks." *Bull. Math. Biol.* 51 (1989), 657–679.

[27] Bookstein, F. L. and Sampson, P. D. "Statistical models for geometric components of shape change." *Comm. Statist. Theory Meth.* 19 (1990), 1939–1972.

[28] Bookstein, F. L., Sampson, P. D., Lewis, S., Guttorp, P., and Hurley, C. B. "Computation and interpretation of deformations for landmark data in morphometrics and environmetrics." *Computing Sc. and Statist. Proc. of the 23th Symposium on the Interface* (1991), 534–541.

[29] Boots, B. N. "Delaunay triangles: an alternative approach to point pattern analysis." *Proc. Assoc. Amer. Geographers* 6 (1974), 26–29.

[30] Bradley, R. and Fulford, M. "Sherd size in the analysis of occupation debris." *Bull. Univ. London Inst. Archaeology* 17, 85–94.

[31] Bradley, R. and Small, C. "Looking for circular structures in post hole

distributions: quantitative analysis of two settlements from Bronze Age England." *J. Archaeol. Sci.* 12 (1985), 285–297.

[32] Bradley, R. and Small, C. "Statistical analysis of structures at two settlements from Bronze Age England." *MASCA J.* 4 (1986), Museum Applied Science Center for Archaeology, University of Pennsylvania, Philadelphia, 86–95.

[33] Broadbent, S. "Simulating the ley hunter (with discussion)." *J. Roy. Statist. Soc. Ser. A* 143 (1980), 109–140.

[34] Bronshtein, I. N. and Semendyayev, K. A. *Handbook of Mathematics.* Van Nostrand Reinhold, New York (1985).

[35] Brown, G. W. "Reduction of a certain class of composite statistical hypotheses." *Ann. Math. Statist.* 11 (1940), 254–270.

[36] Bumcrot, R. J. *Modern Projective Geometry.* Holt, Rinehart and Winston, New York (1969).

[37] Cahn, J. W. "The generation and characterization of shape." *Adv. Appl. Probab.* Special Supplement (1972), 221–242.

[38] Carne, T. K. "The geometry of shape spaces." *Proc. London Math. Soc.* 61 (1990), 407–432.

[39] Carroll, L. *Curiosa Mathematica, Part II: Pillow Problems.* Macmillan, London (1893).

[40] Chen, G. and Chen, J. "Geometric quality inspection." Submitted to *Statistica Sinica* (1996).

[41] Chow, Y., Grenander, U., and Keenan, D. M. "HANDS, a pattern theoretic study of biological shapes." *Technical Report, Division of Applied Math.* Brown University, Providence (1988).

[42] Christaller, W. *Die zentralen Orte in Süddeutschland* Jena (1933). Translated by C. C. Baskin, *Central Places in Southern Germany*, Prentice-Hall, New Jersey (1966).

[43] Chung, K. L. *A Course In Probability Theory.* Academic Press, New York (1974).

[44] Cogbill, S. "Computer post hole analysis with reference to the Bronze Age." *Computer Appl. Archaeol. 1980*, University of Birmingham (1980), 35–38.

[45] Corner, B. D. and Richtsmeier, J. T. "Morphometric analysis of craniofacial growth in *Cebus apella.*" *Amer. J. Phys. Anthrop.* 84 (1991), 323–342.

[46] Cover, T. and Efron, B. "Geometrical probability and random points in a hypersphere." *Ann. Math. Statist.* 38 (1967), 213–220.

[47] Cowan, R. "A collection of problems in random geometry." In *Stochastic Geometry, Geometric Statistics, Stereology*, Teubner-Texte zur Mathematik 65, Edited by R. V. Ambarzumian and W. Weil. Leipzig (1984), 64–68.

[48] Cruz-Orive, L.-M. "Particle size-shape distributions: the general spheroid problem, part 1." *J. Microsc.* 107 (1976), 235–253.

[49] Cruz-Orive, L.-M. "Particle size-shape distributions: the general spheroid problem, part 2." *J. Microsc.* 112 (1978), 153–167.

[50] Cruz-Orive, L.-M. and Weibel, E. R. "Recent stereological methods for cell biology: a brief survey." *Am. J. Physiol.* 258 (1990) L148–L156.

[51] Dieudonné J. *Treatise on Analysis, Vol. 3.* Academic Press (1969).

[52] Dryden, I. L. "The statistical analysis of shape data." PhD Thesis, Dept. of Statistics, University of Leeds, Leeds (1989).

[53] Dryden, I. L. and Mardia, K. V. "General shape distributions in the plane." *Adv. Appl. Probab.* 23 (1991), 259–276.

[54] DuPuis, P. and Oliensis, J. "An optimal control formulation and related numerical methods for a problem in shape reconstruction." *Ann. Appl. Probab.* 4 (1994), 287–346.

[55] Fisher, N. I. *Statistical Analysis of Circular Data.* Cambridge University, Cambridge (1993).

[56] Fisher, N. I., Lewis, T., and Embleton, B. J. J. *Statistical Analysis of Spherical Data.* Cambridge University, Cambridge (1987).

[57] Fisher, R. A. "Two new properties of mathematical likelihood." *Proc. Roy. Soc. London Ser. A* 144 (1934), 285–307.

[58] Fletcher, M. and Lock, G. "Computerised pattern perception with post hole distributions." *Sci. Archaeol.* 23 (1981), 15–20.

[59] Fletcher, M. and Lock, G. "Post built structures at Danebury hillfort: an analytical search method with discussion." *Oxford J. Archaeol.* 6 (1984), 175–196.

[60] Flury, B. A. "Principal points." *Biometrika* 77 (1990), 33–41.

[61] Fukugawa, H. and Pedoe, D. *Japanese Temple Geometry Problems.* Charles Babbage Research Centre, Winnipeg (1989).

[62] Girling, A. J. "Shape analysis for the anisotropic corpuscle problem." *J. Roy. Statist. Soc. Ser. B* 55 (1993), 675–686.

[63] Goodall, C. R. "The statistical analysis of growth in two dimensions." PhD Dissertation, Department of Statistics, Harvard University (1983).

[64] Goodall, C. R. "The growth of a two-dimensional figure: strain crosses and confidence regions." *Proc. Statist. Comput. Sect. Amer. Statist. Assoc.* (1984), 165–169.

[65] Goodall, C. R. "Discussion of 'Size and shape spaces for landmark data in two dimensions' by F. L. Bookstein." *Statist. Sci.* 1 (1986), 181–242.

[66] Goodall, C. R. "Procrustes methods in the statistical analysis of shape (with discussion)." *J. Roy. Statist. Soc. Ser. B* 53 (1991), 285–339.

[67] Goodall, C. R. and Bose, A. "Procrustes techniques for the analysis of shape and shape change." In *Computer Science and Statistics*, ed. R. M. Heiberger. American Statistical Association, Alexandria (1987), 86–92.

[68] Goodall, C. R. and Green, P. B. "Quantitative analysis of surface growth." *Bot. Gaz.* 147 (1986), 1–15.

[69] Goodall, C. R., Lange, N. and Moss, M. L. "Growth-curve models for repeated triangular shapes." Manuscript.

[70] Goodall, C. R. and Mardia, K. V. "A geometric derivation of the shape density." *Adv. Appl. Probab.* 23 (1991), 496–514.

[71] Goodall, C. R. and Mardia, K. V. "The noncentral Bartlett decompositions and shape densities." *J. Multivariate Anal.* 40 (1992), 94–108.

[72] Goodall, C. R. and Mardia, K. V. "Multivariate aspects of shape theory." *Ann. Statist.* 21 (1993), 848–866.

[73] Gough, J. and Mardia, K. V. "Shape modelling using deformable polygons." *Technical Report.* Department of Statistics, University of Leeds (1990).

[74] Gower, J. C. "Some distance properties of latent root and vector methods used in multivariate analysis." *Biometrika* 53 (1966), 325–338.

[75] Gower, J. C. "Generalized Procrustes analysis." *Psychometrika* 40 (1975), 33–50.

[76] Gregory, R. L. and Gombrich, E. H. *Illusion in Art and Nature.* Duckworth, London (1973).

[77] Grenander, U. and Miller, M. I. "Representation of knowledge in complex systems." *J. Roy. Statist. Soc. Ser. B* 56 (1994), 549–603.

[78] Guillemin, V. and Pollack, A. *Differential Topology.* Prentice Hall, New Jersey (1974).

[79] Hewitt, E. and Ross, K. A. *Abstract Harmonic Analysis, Vols. I, II.* Springer-Verlag, Berlin (1963).

[80] Heyer, H. *Probability Measures on Locally Compact Groups.* Springer-Verlag, Berlin (1977).

[81] Hills, M. "Allometry." In *Encyclopedia of Statistical Sciences Vol. 1.* Wiley, New York (1982), 48–54.

[82] Hogg, R. V. and Craig, A. T. *Introduction to Mathematical Statistics.* Fourth Edition. Macmillan, New York (1978).

[83] Horn, B. K. P. and Brooks, M. J. *Shape From Shading.* MIT Press, Cambridge (1989).

[84] Hulting, F. L. "Methods for the analysis of coordinate measurement data." *Computing Science and Statistics* 24 (1992), 160–169.

[85] Huxley, J. S. *Problems of Relative Growth*, 2nd ed. Dover, New York (1972).

[86] Jones, M. C. and Sibson, R. "What is projection pursuit? (with discussion)." *J. Roy. Statist. Soc. Ser. A* 150 (1987), 1–36.

[87] Kendall, D. G. "The diffusion of shape." *Adv. Appl. Probab.* 9 (1977), 428–430.

[88] Kendall, D. G. "The statistics of shape." In *Interpreting Multivariate Data*, ed. V. Barnett. Wiley, New York (1981), 75–80.

[89] Kendall, D. G. "The shape of Poisson-Delaunay triangles." In *Studies in Probability and Related Topics*, ed. M. C. Demetrescu and M. Iosefescu. Nagard, Sophia (1983).

[90] Kendall, D. G. "Shape manifolds, Procrustean metrics, and complex projective spaces." *Bull. London Math. Soc.* 16 (1984), 81–121.

[91] Kendall, D. G. "Exact distributions for shapes of random triangles in convex sets." *Adv. Appl. Probab.* 17 (1985), 308–329.

[92] Kendall, D. G. "A survey of the statistical theory of shape." *Statist. Sci.* 4 (1989), 87–120.

[93] Kendall, D. G. "Random Delaunay simplexes in $\mathbf{R}^m$." *J. Statist. Plann. Inf.* 25 (1990), 225–234.

[94] Kendall, D. G. "Spherical triangles revisited." *The Art of Statistical Science. A Tribute to G. S. Watson*, ed. K. V. Mardia (1992). Wiley, New York, 105–113.

[95] Kendall, D. G. and Kendall, W. S. "Alignments in two-dimensional random sets of points." *Adv. Appl. Probab.* 12 (1980), 380–424.

[96] Kendall, W. S. "Symbolic computation and the diffusion of shapes of triads." *Adv. Appl. Probab.* 20 (1988), 775–797.

[97] Kendall, W. S. "The diffusion of Euclidean shape." In *Disorder in Physical Systems*, ed. D. Welsh and G. Grimmet. Oxford University Press, Oxford (1990), 203–217.

[98] Kent, J. T. "The complex Bingham distribution and shape analysis." *J. Roy. Statist. Soc. Ser. B* 56 (1994), 285–299.

[99] Kovalenko, I. N. "On recovering the additive type of a distribution over a sequence of runs of independent observations." *Trudy Vsesoyuzn. Soveshcheniya po Teorii Veroyatnostei i Matematicheskoi Statistike, Erevan* (1960), 148–159. In Russian.

[100] Langford, E. "Probability that a random triangle is obtuse." *Biometrika* 56 (1969), 689–690.

[101] Le, H.-L. "Explicit formulae for polygonally generated shape-densities in the basic tile." *Math. Proc. Camb. Philos. Soc.* 101 (1987), 313–321.

[102] Le, H.-L. "Singularities of convex-polygonally generated shape-densities." *Math. Proc. Camb. Philos. Soc.* 102 (1987), 587–596.

[103] Le, H.-L. "A stochastic calculus approach to the shape distribution induced by a complex normal model." *Math. Proc. Cambridge Philos. Soc.* 109 (1990), 221–228.

[104] Le, H.-L. "On geodesics in Euclidean shape spaces." *J. London Math. Soc.* 44 (1991), 360–372.

[105] Le, H.-L. and Kendall, D. G. "The Riemannian structure of Euclidean spaces: a novel environment for statistics." *Ann. Statist.* 21 (1993), 1225–1271.

[106] Lele, S. R. "Some comments on coordinate free and scale invariant methods in morphometrics." *Amer. J. Phys. Anthrop.* 85 (1991), 407–418.

[107] Lele, S. R. and Richtsmeier, J. T. "Euclidean distance matrix analysis: a coordinate-free approach for comparing biological shapes using landmark data." *Amer. J. Phys. Anthrop.* 86 (1991), 415–427.

[108] Lele, S. R. and Richtsmeier, J. T. "On comparing biological shapes: detection of influential landmarks." *Amer. J. Phys. Anthrop.* 87 (1992), 49–65.

[109] Litton, C. D. and Restorick, J. "Computer analysis of post hole distributions." *Computer Applic. Archaeology 1983*, University of Birmingham (1983), 85–92.

[110] Lohmann, G. P. "Eigenshape analysis of microfossils: a general morphometric procedure for describing changes in shape." *Math. Geol.* 15 (1983), 659–672.

[111] Mack, C. "Expected number of aggregates in a random distribution of n points." *Proc. Cambridge Philos. Soc.* 46 (1950), 285–292.

[112] Mannion, D. "A Markov chain of triangle shapes." *Adv. Appl. Probab.* (1988), 348–370.

[113] Mardia, K. V. *Statistics of Directional Data.* Academic Press, London (1972).

[114] Mardia, K. V. "Shape analysis of triangles through directional techniques." *J. Roy. Statist. Soc. Ser. B* 51 (1989), 449–458.

[115] Mardia, K. V. "Discussion of 'A survey of the statistical theory of shape' by D. G. Kendall." *Statist. Sci.* 4 (1989), 108–111.

[116] Mardia, K. V. and Dryden, I. L. "Shape distributions for landmark data." *Adv. Appl. Probab.* 21 (1989), 742–755.

[117] Mardia, K. V. and Dryden, I. L. "The statistical analysis of shape data." *Biometrika* 76 (1989), 271–281.

[118] Mardia, K. V., Edwards, R., and Puri, M. L. "Analysis of central place theory." *Bull. Int. Statist. Inst.* 47 (1977), 93–110.

[119] Miles, R. "On the homogeneous Poisson process." *Math. Biosciences* 6 (1970), 85–127.

[120] Montgomery, D. and Zippin, L. *Topological Transformation Groups.* Interscience Publishers, New York (1955).

[121] Morgan, F. *Geometric Measure Theory: A Beginner's Guide.* Academic Press, Boston (1988).

[122] Morgan, F. *Riemannian Geometry: A Beginner's Guide.* Jones and Bartlett, Boston (1992).

[123] Mosier, C. I. "Determining a simple structure when loadings for certain tests are known." *Psychometrika* 4 (1939), 149–162.

[124] Mosimann, J. E. "Size allometry: size and shape variables with characterizations of the lognormal and generalized gamma distributions." *J. Amer. Statist. Assoc.* 65 (1970), 930–948.

[125] Mosimann, J. E. "Size and shape analysis." In *Encyclopedia of Statistical Sciences Vol. 8.* Wiley, New York (1988), 497–507.

[126] Mosimann, J. E. and James, F. C. "New statistical methods for allometry with application to Florida red-winged blackbirds." *Evolution* 33 (1979), 444–459.

[127] Neyman, J. and Scott, E. "Consistent estimates based on partially consistent observations." *Econometrics* 16 (1948), 1–32.

[128] Okabe, A., Boots, B., and Sugihara, K. *Spatial Tessellations Concepts and Applications of Voronoi Diagrams.* Wiley, New York (1992).

[129] Orton, C. *Mathematics in Archaeology.* Cambridge University, Cambridge (1980).

[130] Ostrom, J. H. "Were some dinosaurs gregarious?" *Paleogeog. Paleoclimat. Paleoecol.* 11 (1972), 287–301.

[131] Oxnard, C. E. *The Order of Man: A Biomathematical Anatomy of the Primates.* Yale University, New Haven (1984).

[132] Parthasarathy, K. R. *Probability Measures on Metric Spaces.* Academic Press, New York (1967).

[133] Paul, G. S. *Predatory Dinosaurs of the World: A Complete Illustrated Guide.* Simon and Schuster, New York (1989).

[134] Pitman, E. J. G. "The estimation of the location and scale parameters of a continuous population of any given form." *Biometrika* 30 (1939), 391–421.

[135] Portnoy, S. "A Lewis Carroll pillow problem: probability of an obtuse

triangle." *Statist. Sc.* 9 (1994), 279–284.

[136] Prakasa Rao, B. L. S. "On a characterization of probability distributions on locally compact abelian groups." *Z. Wahrscheinlichkeits. verw. Gebiete* 9 (1968), 98–100.

[137] Prohorov, Yu. V. "On a characterization of a class of probability distributions by distribution of some statistics." *Theor. Probab. Appl.* 10 (1965), 438–445.

[138] Rao, C. R. *Linear Statistical Inference and Its Applications.* Wiley, New York (1965).

[139] Reyment, R. A. Blacklith, R. E., and Campbell, N. A. *Multivariate Morphometrics*, 2nd ed. Academic Press, New York (1984).

[140] Ripley, B. D. *Spatial Statistics.* Wiley, New York (1981).

[141] Ripley, B. D. and Rasson, J.-P. "Finding the edge of a Poisson forest." *J. Appl. Probab.* 14 (1977), 483–491.

[142] Rohlf, F. J. and Bookstein, F. L. *Proceedings of the Michigan Morphometrics Workshop.* The University of Michigan Museum of Zoology, Ann Arbor (1990). Special Publication Number 2.

[143] Rohlf, F. J. and Slice, D. "Methods for comparison of sets of landmarks." *Syst. Zool.* 39 (1990), 40–59.

[144] Rukhin, A. "Charakterisierung der Transformationsparameterfamilie." *Z. Wahrscheinlichkeits. verw. Gebiete* 38 (1977), 287–291.

[145] Sampson, P. D. "Dental arch shape: a statistical analysis using conic sections." *Amer. J. Orthodont.* 79 (1981), 535–548.

[146] Sampson, P. D. and Siegel, A. F. "The measure of 'size' independent of 'shape' for multivariate lognormal populations." *J. Amer. Statist. Assoc.* 80 (1985), 910–914.

[147] Santalo, L. A. *Integral Geometry and Geometric Probability.* Encyclopedia of Mathematics and Its Applications. Addison-Wesley, Reading (1976).

[148] Serra, J. P. *Image Analysis and Mathematical Morphology.* Academic Press, New York (1982).

[149] Schwerdtfeger, H. *Geometry of Complex Numbers.* Dover, New York (1962).

[150] Sibson, R. "Studies in the robustness of multidimensional scaling: Procrustes statistics." *J. Roy. Statist. Soc. Ser. B* 40 (1978), 234–238.

[151] Sibson, R. "Studies in the robustness of multidimensional scaling: perturbational analysis of classical scaling." *J. Roy. Statist. Soc. Ser. B* 41 (1979), 217–229.

[152] Siegel, A. F. and Benson, R. H. "A robust comparison of biological shapes." *Biometrics* 38 (1982), 341–350.

[153] Silverman, B. W. and Brown, T. C. "Short distances, flat triangles and Poisson limits." *J. Appl. Probab.* 15 (1978), 815–825.

[154] Small, C. G. "Distributions of shape and maximal invariant statistics." PhD Thesis. University of Cambridge, Cambridge (1981).

[155] Small, C. G. "Random uniform triangles and the alignment problem." *Math. Proc. Camb. Philos. Soc.* 91 (1982), 315–322.

[156] Small, C. G. "Characterization of distributions from maximal invariant statistics." *Z. Wahrsch. verw. Gebiete* 63 (1983), 517–527.

[157] Small, C. G. "A classification theorem for planar distributions based on the shape statistics of independent tetrads." *Math. Proc. Camb. Philos. Soc.* 96 (1984), 543–547.

[158] Small, C. G. "Discussion of 'A survey of the statistical theory of shape' by D. G. Kendall." *Statist. Sci.* 4 (1989), 105–108.

[159] Small, C. G. "Techniques of shape analysis on sets of points." *Int. Statist. Rev.* 56 (1988), 243–257.

[160] Small, C. G. "Reconstructing convex bodies from random projected images." *Can. J. Statist.* 19 (1991), 341–347.

[161] Small, C. G. "A counterexample to a conjecture on random shadows."

Can. J. Statist. 20 (1992), 463–468.

[162] Small, C. G. and Murdoch, D. J. "Nonparametric Neyman-Scott problems: telescoping product methods." *Biometrika* 80 (1993), 763–779.

[163] Spivak, M. *A Comprehensive Introduction to Differential Geometry, Volume I.* Publish or Perish, Berkeley (1970).

[164] Sprent, P. "The mathematics of size and shape." *Biometrics* 28 (1972), 23–37.

[165] Stone, J. F. S. "The Deverel-Rimbury settlement on Thorny Down, Winterbourne gunner, S. Wilts." *Proc. Prehistoric Soc.* 7 (1941), 114–133.

[166] Stoyan, D. "Estimation of distances and variances in Bookstein's landmark model." *Biometr. J.* 32 (1990), 843–849.

[167] Stoyan, D., Kendall, W. S., and Mecke, J. *Stochastic Geometry and Its Applications.* Akademie, Berlin (1987).

[168] Stoyan, D. and Stoyan, H. "A further application of D. G. Kendall's procrustes analysis." *Biometr. J.* 32 (1990), 293–301.

[169] Strauss, R. E. and Bookstein, F. L. "The truss: body form reconstruction in morphometrics." *Systematic Zool.* 31 (1982), 113–135.

[170] Stuetzle, W., Gasser, T. Molinari, L., Largo, R. H., Prader, A., and Huber, P. J. "Shape invariant modelling of human growth." *Ann. Human Biol.* 7 (1980), 507–528.

[171] Sukiasian, H. S. "Two results on triangle shapes." In *Stochastic Geometry, Geometric Statistics, Stereology*, Teubner-Texte zur Mathematik 65, ed. R. V. Ambartzumian and W. Weil. Teubner, Leipzig (1984), 210–221.

[172] Thompson, D'A. W. *On Growth and Form.* Cambridge University Cambridge (1961). Reprinted from 1917.

[173] Tobler, W. R. "The comparison of plane forms." *Geograph. Anal.* 10 (1978), 154–162.

[174] Wagner, H. M. *Principles of Operations Research.* Prentice-Hall, Englewood Cliffs (1975).

[175] Wainwright, R. *A Guide to the Prehistoric Remains in Britain, Volume I: South and East.* Constable, London (1978).

[176] Wainwright, S. A. *Axis and Circumference: The Cylindrical Shape of Plants and Animals.* Harvard University Press, Harvard (1988).

[177] Watkins, A. *The Old Straight Track.* Methuen (1925). Republished Garnstone Press (1970).

[178] Watson, G. S. "Random triangles." In *Proc. Conf. Stochastic Geometry*, ed. O. Barndorff-Nielsen. Aarhus University, Aarhus (1983).

[179] Watson, G. S. "The shape of a random sequence of triangles." *Adv. Appl. Probab.* 18 (1986), 156–169.

[180] Weibel, E. R. *Stereological Methods.* Academic Press, London (1980).

[181] Wicksell, S. D. "The corpuscle problem, part 1." *Biometrika* 17 (1925), 84–99.

[182] Wicksell, S. D. "The corpuscle problem, part 2." *Biometrika* 18 (1926), 152–172.

[183] Ziezold, H. "On expected figures in the plane." *Math. Res. Ser.* 51 (1989), 105–110.

[184] Zinger, A. A. and Linnik, Yu. V. "On characterizations of the normal distributions." *Theor. Probab. Appl.* 9 (1964), 624–626.

[185] Zuiderwijk, E. J. "Alignment of randomly distributed objects." *Nature* 295 (1982), 577–578.

Index

Springer Series in Statistics

(continued from p. ii)

Pollard: Convergence of Stochastic Processes.
Pratt/Gibbons: Concepts of Nonparametric Theory.
Read/Cressie: Goodness-of-Fit Statistics for Discrete Multivariate Data.
Reinsel: Elements of Multivariate Time Series Analysis.
Reiss: A Course on Point Processes.
Reiss: Approximate Distributions of Order Statistics: With Applications to Non-parametric Statistics.
Rieder: Robust Asymptotic Statistics.
Rosenbaum: Observational Studies.
Ross: Nonlinear Estimation.
Sachs: Applied Statistics: A Handbook of Techniques, 2nd edition.
Särndal/Swensson/Wretman: Model Assisted Survey Sampling.
Schervish: Theory of Statistics.
Seneta: Non-Negative Matrices and Markov Chains, 2nd edition.
Shao/Tu: The Jackknife and Bootstrap.
Siegmund: Sequential Analysis: Tests and Confidence Intervals.
Simonoff: Smoothing Methods in Statistics.
Small: The Statistical Theory of Shape.
Tanner: Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions, 3rd edition.
Tong: The Multivariate Normal Distribution.
van der Vaart/Wellner: Weak Convergence and Empirical Processes: With Applications to Statistics.
Vapnik: Estimation of Dependences Based on Empirical Data.
Weerahandi: Exact Statistical Methods for Data Analysis.
West/Harrison: Bayesian Forecasting and Dynamic Models.
Wolter: Introduction to Variance Estimation.
Yaglom: Correlation Theory of Stationary and Related Random Functions I: Basic Results.
Yaglom: Correlation Theory of Stationary and Related Random Functions II: Supplementary Notes and References.