

CHAPTER 5

The Statistics of Shape

D. G. Kendall, *University of Cambridge, U.K.*

This is a brief report of work in progress, or in preparation for publication elsewhere. In this investigation 'shape' always means the shape of a (possibly random) configuration of points in euclidean space. The basic idea is to think of k points in m dimensions labelled as P_1, P_2, \dots, P_k and to identify the shape of that k -ad with the equivalence class of k -ads similar to it relative to changes of location, scale, and orientation, but not normally of handedness. When $m = 1$ (the simplest case) there turns out to be a representation for the quotient space (in which such shapes are points) as the sphere S^{k-2} , while when $m = 2$ the corresponding quotient space is the complex projective space CP^{k-2} , which can be thought of as the sphere S^2 when $k = 3$, so that 'shapes of triangles live on a sphere' (see Figure 5.1).

For larger values of m the most interesting case is that in which $k = m + 1$, and it turns out that what I call the *shape-manifold* is actually endowed with a smooth structure, a group, an invariant riemannian metric, and an invariant measure. Part of the group accounts for the transformations associated with re-labelling points, and it is often convenient to use this fact in the construction of a sub-region of the manifold on which unlabelled shapes can be plotted; Figure 5.1 gives an example of this. The metric can be derived from procrustean considerations concerning the best way of matching two labelled figures when translations, scale changes and rotations are permitted; asymptotic analysis of this then leads to the riemannian metric. This natural metric of itself induces a probability distribution on the shape-manifold which turns out to have connections with a statistical example in which the $m + 1$ points are independent, identically distributed, and have a gaussian distribution. It should be mentioned that the homogeneity of the shape-manifold when $k = 3$ and $m = 2$ is exceptional; for larger values of m the intrinsic geometry of the manifold depends on where one is on it, and the analysis of this is very interesting.

Returning to the general situation, suppose that the k points are generated by an arbitrary random mechanism (not necessarily independently). This will automatically induce a probability distribution on the shape-manifold itself, which I call the *shape-measure*. Moreover, in most cases of interest it turns out

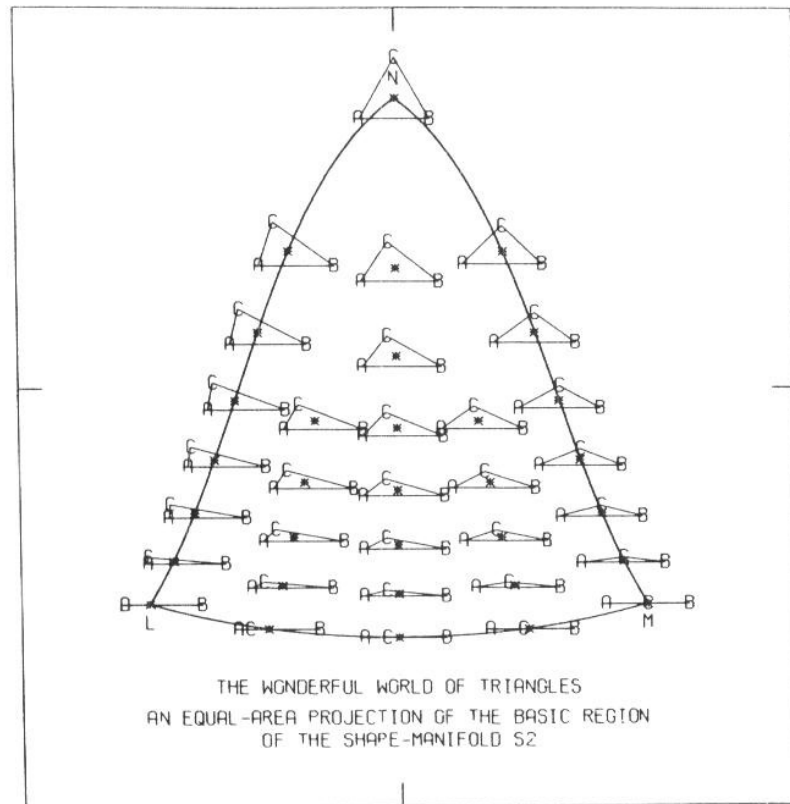


Figure 5.1 The spherical blackboard, showing 32 triangles located according to their shapes

that this shape-measure possesses a density relative to the invariant measure mentioned above, and this density is therefore more conveniently taken as a representation of the way in which the shapes are distributed on the shape-manifold. It has now been computed in one or two cases. For example, I have found the shape-density in the general multivariate gaussian case, and when $m = 2$ this distribution (which is closely related to work by W. S. Kendall, 1981) has proved to be important in the analysis of alignments (whether 3-point or k -point) in empirical sets of points. When $m = 3$ or more, these results generalize in the obvious way and it becomes particularly interesting to see how non-isotropy in the generating gaussian distribution is reflected in the form of the induced density on the shape-manifold. Alternatively, instead of considering gaussian distributions, one can consider the case in which the k -points are independently and identically distributed, and lie uniformly inside some given compact convex region. This problem has been investigated by C. G. Small of the University of Cambridge for the case $m = 2$, and he has

also done very interesting work on the associated questions of unicity and stability which naturally arise, both in this case and in more general cases.

Now let us forget about statistics for a moment, and think of the k points as given. Then each shape-function (e.g. the maximum angle of the triangle when $k = 3$ and $m = 2$) can be discussed in terms of the real function which it determines over the shape-manifold. In the case just mentioned I have written computer programs to draw contour lines for that particular shape-function, and these have proved useful in the interpretation of observed alignments. Going back to the statistical problem, it will be seen that the theoretical distribution of values of that shape-function (e.g. the distribution of the maximum angle of a random gaussian triangle, originally determined by other methods by W. S. Kendall, 1981) can be obtained by fitting together two components. The first of these is the shape-density determined by the stochastic mechanism generating the points, while the second is the shape-function viewed as a real-valued function over the shape-manifold. The distribution to be found is obtained from these components by an appropriate integration, and a computer program has been constructed to perform it. Examples of the output of that program will be found in Kendall and Kendall (1980).

In the context of the present volume these ideas are relevant because they demand a solution to the problem of visually presenting such data, and this problem is studied in detail in a forthcoming paper (D. G. Kendall, 1981a) in the important case where $m = 2$ and $k = 3$. Here, as mentioned above, the shape-manifold is the sphere, which can be thought of as the union of six regions equivalent under the re-labelling group. If we are also indifferent to reflection (as sometimes in practice we are) then each such region is halved and we have twelve in all. Choosing any one of these as representative of all the others, this basic region is a spherical triangle on which the invariant measure is just Lebesgue, and we can preserve that by a radial projection from any axis through the centre of the sphere onto an enveloping right circular cylinder having that axis. Cutting and opening out the cylinder then gives a measure-preserving map from the shape-manifold to the plane and on this we can display statistical distributions, whether empirical or theoretical, in a manner which is very easy to assimilate once one has learnt by heart the information contained in Figure 5.1.

What might be called the 'bell-shaped region' in that figure is in fact one of the twelve regions derived from relabelling plus reflection, and the situation is that the equilateral shapes are represented by a point at the top of the bell, while collinear shapes are represented by points lying along the curve at the foot of the bell. The isosceles shapes appear on the two sides of the bell and it is interesting that they have to be segregated into two groups; those where the vertical angle is greater than either of the two equal angles—and here the shape-points occur on the right—and those where the vertical angle is less than each of the two equal angles—and here the shape-point occurs on the

s located

iriant measure
ntly taken as a
on the shape-
xample, I have
ase, and when
W. S. Kendall,
ents (whether
or more, these
y interesting to
reflected in the
ely, instead of
e in which the
lie uniformly
en investigated
= 2, and he has

left. Once this has been explained, the way in which the shape of the triangle changes as one moves about inside the bell can be seen from the diagram. Thus the reader will notice that collinear triplets in which two members of the triplet coincide or nearly coincide will determine shapes represented by points near the corner of the bell marked by the letter L .

These procedures have been exploited in various ways. The study of nearly collinear (blunt) triangles initiated by Broadbent (1980) with its archaeological and other applications provides one with suitable data, and I have constructed a computer program which accepts such data, calculates the shapes of all the triangles contained in it, and plots the shape-points on the bell. Actually with data of the size of Broadbent's there are too many points to plot (22 100); in order to avoid an output which would be merely an inky mess, the program is therefore directed to draw contour lines on the bell indicating the form of the empirical density function. This can be compared with the output of the same program in another mode when a theoretical stochastic generator for the points is proposed and supplied via a sub-routine, and either the shape-density is theoretically calculated and contour lines drawn on the bell, or alternatively samples derived from that stochastic mechanism are generated by simulation and analysed as if they were data.

It may be interesting to mention that Broadbent's data, when analysed in this way, show an unexpected concentration in the neighbourhood of the point L . This is easily explained because in fact there are one or two cases in the Broadbent data when two sites are very close together. Each such pair of sites can be associated with any one of the fifty remaining sites (there were 52 sites altogether), and therefore one obtains fifty isosceles triangles with a very small vertical angle; this automatically produces a lot of points, and therefore a high empirical density, near the corner marked L . In this way, and in other ways too complicated for summary here, one learns by looking at the bell-plot of the Broadbent data to recognize many intrinsic features of the data-set which were not immediately obvious from earlier visual presentation. Because of its very flexible use, I like to call the bell diagram 'the spherical blackboard'; the main function of the various computer programs I am writing and testing at the moment is to enable one to put onto the spherical blackboard as many different objects of interest as one can.

To take another example, recent work by Mardia, Edwards and Puri (1977) has suggested that one might use the Delaunay triangles associated with an empirical plane point process to investigate whether in fact there is any evidence for the influence of what is called by geographers 'central place theory'. The reason for using the Delaunay triangles in this rôle is that if central place theory is indeed dominant in the true model generating the data then one would expect more nearly equilateral triangles than would be predicted by the Delaunay distribution calculated by R. E. Miles (see the paper by Mardia *et al.* for an appropriate reference) for data coming from a Poisson

REFERENCES

- Broadbent, S. R. (1980) Simulating the ley hunter. *J. Roy. Statist. Soc. A*, **143**, 109–40.
- Kendall, D. G. (1977) The diffusion of shape. *Adv. Appl. Prob.*, **9**, 428–30.
- Kendall, D. G. (1981a) Shape-manifolds, Procrustean metrics and complex projective spaces (in preparation).
- Kendall, D. G. (1981b) Foundations of a theory of random shape. *Bull. London Math. Soc.* (to appear).
- Kendall, D. G. and Kendall, W. S. (1980) Alignments in two-dimensional random sets of points. *Adv. Appl. Prob.*, **12**, 380–424.
- Kendall, W. S. (1981) Random gaussian triangles and k -point collinearities (in preparation).
- Mardia, K. V., Edwards, R., and Puri, M. L. (1977) Analysis of central place theory. *Bull. Int. Statist. Inst.*, **7**, (2), 93–110.
- Small, C. G., Random uniform triangles and the alignment problem. *Ph.D. thesis* (in preparation), University of Cambridge.
- Small, C. G. Characterization of distributions through shapes of samples. *Ph. D. thesis*, (in preparation), University of Cambridge.

PART II

Re
Ar
an