Geometry of circular vectors and pattern recognition of shape of a boundary

(CALYAMPUDI R. RAO*)

Statistics Department, 326 Joab Thomas Building, Pennsylvania State University, University Park, PA 16802

Contributed by Calyampudi R. Rao, July 31, 1998

ABSTRACT This paper deals with pattern recognition of the shape of the boundary of closed figures on the basis of a circular sequence of measurements taken on the boundary at equal intervals of a suitably chosen argument with an arbitrary starting point. A distance measure between two boundaries is defined in such a way that it has zero value when the associated sequences of measurements coincide by shifting the starting point of one of the sequences. Such a distance measure, which is invariant to the starting point of the sequence of measurements, is used in identification or discrimination by the shape of the boundary of a closed figure. The mean shape of a given set of closed figures is defined, and tests of significance of differences in mean shape between populations are proposed.

1. Introduction

The problem of pattern recognition of the boundary of a closed figure arises in a variety of situations, such as contour coding, classification of chromosomes, interpretation of chest x-rays, scene analysis, identification of aircraft, and so on. References to applications in these areas can be found in a paper by Kashyap and Chellappa (1) and a recent book edited by Lestrel (2). If the boundary has specific landmarks, then one could use distances and angles between lines joining the landmarks as features and apply methods of multivariate statistical analysis as described in the papers by Rao and Suryawanshi (3, 4). References to applications in these areas can be found in a paper by Lestrel (2). If the boundary has specific landmarks, then one could use distances and angles between lines joining the landmarks as features and apply methods of multivariate statistical analysis as described in the papers by Rao and Suryawanshi (3, 4).

In cases where no distinct landmarks can be located on the boundary, several methods of making measurements to describe the shape of the boundary have been proposed. Some of these methods depending on the complexity of the boundary are described below.

1. Let \( A_1 \) be any point on the boundary. Mark off \((n - 1)\) points \( A_2, \ldots, A_n \) such that the angle between \( OA_i \) and \( OA_{i+1} \) is \(2\pi/n\) for all \( i \), where \( O \) is the centroid as shown in Fig. 1. Then the sequence of distances \( OA_i = r(i), i = 1, \ldots, n \), which provides a description of the boundary, is invariant to translation and rotation but not to scale and starting point. See ref. 5 for a description and an application of this method. Instead of the sequence \( r(i) \), one may consider the sequence of \( c(i) \), the curvature at \( A_i \), as in ref. 6. The sequence \( c(i) \) has the same invariance property as \( r(i) \).

2. Construct the point \( A_1, \ldots, A_n \) as in step 1 and denote the distance \( A_iA_{i+1} = b(i), i = 1, \ldots, n \). The sequence \( b(i) \) has the same invariance property mentioned in step 1.

3. Using the same points \( A_1, \ldots, A_n \), we may consider the sequence of angles \( a(i) \) between the sides \( OA_i \) and \( A_iA_{i+1}, i = 1, \ldots, n \). These angles are invariant to translation, rotation, and scale but not to the starting point.

4. Fit a polygon of \( n \) sides to the boundary such that all the sides have the same length. Call this polygon \( A_1A_2 \ldots A_nA_1 \) and let \( \phi_i \) be the angle between \( A_iA_{i+1} \) and a reference line. Consider the sequence \( \theta_i = \phi_{i+1} - \phi_i \) or \( \phi_{i+1} - \phi_i, i = 1, \ldots, n - 1 \), which is invariant for all transformations except for starting point. Zahn and Roskies (7) give a slightly different description of the method (see Fig. 2).

5. Let \( \Gamma \) be a continuous directed curve in the Euclidean plane and parametrized by its arc length \( t \). This signature \( s(t) \) of \( \Gamma \) is the function that associates each point \( t \) of \( \Gamma \) with the length that is on or to the left of a tangent line at \( t \). To remove the effect of the scale, we consider the normalized signature obtained by dividing the value of the signature at point \( t \) by the total length of the curve and by taking \( S(t) = S(Lt) \), where \( L \) is the length of the curve of \( \Gamma \) and \( t \in [0, 1] \). Instead of the length of the curve, we may consider the signature as the area of the curve to the left of the tangent line at \( t \) and normalize it suitably. For details, see ref. 8.

6. Another way of characterizing a plane curve is to express the \( x, y \) coordinates of a point with respect to a chosen coordinate frame as functions of the length of the curve from a chosen point to the given point (see Fig. 3). This way, any arbitrary curve can be described in terms of one parameter. However, the \( x, y \) coordinates are not invariant to translation, rotation, and starting point. They have to be standardized in a suitable way for further processing. A typical example of such representation is the chain-encoded contour described by Kuhl and Giardina (9).

In all of the above situations, the sequences of boundary measurements on different objects are comparable only if the...
starting point is a well specified landmark on the boundary. If no such landmark exists or if the landmarks on the boundary are not considered relevant for characterizing the shape of a boundary, the sequences of measurements may not be comparable because of the arbitrariness of the starting point in each sequence. To overcome this problem, the sequence of measurements or its Fourier descriptors are transformed into functions invariant to the starting point, and such functions are used in comparing shapes. The articles in ref. 2 describe such methods in different applications.

In this paper we offer a different approach. First, we observe that a given sequence of measurements is cyclical in nature and as such the measurements can be represented as a clockwise cyclical sequence with the last measurement placed to the left of the first (as a circular vector) with no specified starting point, and develop methods of comparing such circular vectors. Kashyap and Chellappa (1) develop a stochastic model for circular vectors, but our method is based on a geometry of such vectors as developed in the next section.

2. Geometry of Circular Vectors

Let $X$ be an $n$-vector of a sequence of measurements $x_1, \ldots, x_n$ starting with $x_1$. Then we can generate the sequence $x_i, x_{i+1}, \ldots, x_{i+n-1}$ from $X$ by the matrix operation $C^{n-1}X$, where $C$ is the matrix

$$C = \begin{pmatrix}
0 & 1 & 0 & \cdots & 0 \\
0 & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 0 & 0 & \cdots & 0
\end{pmatrix}. \tag{2.1}
$$

We consider a space of circular vectors (SCV) whose points are sets

$$\tilde{X} = (X, CX, \ldots, C^{n-1}X). \tag{2.2}
$$

Note that the set $(2, 2)$ can be generated by choosing any member of the set and multiplying by powers of the matrix $C$. We define the distance between two points $X$ and $Y$ as

$$d(X, Y) = \min \|X - CY\|. \tag{2.3}
$$

where $X$ and $Y$ are any members of the set $\tilde{X}$ and $\tilde{Y}$, respectively, and $\| \cdot \|$ is any vector norm function.

**Definition:** We say that $X \in \tilde{X}$ and $Y \in \tilde{Y}$ are in an optimal position to each other if

$$d(X, Y) = \|X - Y\|. \tag{2.4}
$$

We establish the following properties of the distance function $2.3$.

**Theorem 1.**

(a) The distance function $2.3$ is independent of any choice of vectors $X \in \tilde{X}$ and $Y \in \tilde{Y}$.

(b) $d(X, Y) = d(Y, X)$.

(c) $d(X, \tilde{Y}) = 0$ if and only if $X = \tilde{Y}$.

(d) $d(X, \tilde{Y}) + d(\tilde{X}, \tilde{Z}) \equiv d(\tilde{X}, \tilde{Z})$.

**Proof:** The results a, b, and c follow from the definition $2.3$ of distance. To prove d, let $X \in \tilde{X}, Y \in \tilde{Y}, Z \in Z$ and $(X, CY), (X, CZ), (Y, CZ)$ be pairs of vectors, where the members in each pair are in optimal position to each other as defined in $2.4$. Then, by using the properties of a norm,

$$\|X - CY\| + \|X - CZ\| \geq \|CY - CZ\|$$

$$= \|Y - CZ\|$$

$$\geq \|Y - CY\|,$$

which implies that

$$d(X, \tilde{Y}) + d(\tilde{X}, \tilde{Z}) \equiv d(\tilde{Y}, \tilde{Z}), \tag{2.5}
$$

which proves d.

The space of circular vectors (SCV) with the definition of distance $2.3$ is a metric space but not a linear vector space. However, we can define the sum and mean of a set of elements $X_1, \ldots, X_k$.

**Definition:** Let $X_i \in \tilde{X}_i, i = 1, \ldots, k$, be any selection of $n$-vectors. Then an $n$-vector $m$ that generates $\tilde{m}$, the mean of $X_1, \ldots, X_k$, is defined to be $m_0$, obtained by minimizing

$$\|m - CX_1\|^2 + \cdots + \|m - CX_k\|^2 \tag{2.6}
$$

with respect to $r_1, \ldots, r_k$ and $m$. The sum of $\tilde{X}_1, \ldots, \tilde{X}_k$ is $k\tilde{m}_0$.

Let us write $2.6$ at the optimum values of the arguments as

$$\|m_0 - X_1\|^2 + \cdots + \|m_0 - X_k\|^2. \tag{2.7}
$$

Then $m_0$ is the vector that generates the mean $\tilde{m}_0$, and $X_1, \ldots, X_k$, the representatives of $\tilde{X}_1, \ldots, \tilde{X}_k$, are said to be in optimal position to each other. This does not imply that any two $X_1^0$ and $X_2^0$ are in optimal position to each other.
Let us now choose the norm function as the Euclidian vector norm

$$\|X\|^2 = X'X,$$  \[2.8\]

where $X'$ is the transpose of $X$. We have the following theorems under the Euclidian norm.

**Theorem 2.** Let

$$m_0, x_1^0, \ldots, x_k^0$$  \[2.9\]

be an optimum solution to 2.6. Then

(a) $m_0 = k^{-1}(x_1^0 + \cdots + x_k^0) = X_0$

(b) For every $i = 1, \ldots, k$, the vector $x_i^0$ is in optimal position to $X_0$.

**Proof:** Under Euclidian norm, we have the identity

$$\sum_{i=1}^k \|x_i - x_i^0\|^2 = k\|m_0 - x_i^0\|^2 + \sum_{i=1}^k \|X_0 - x_i^0\|^2.$$

Then the results a and b follow from the definition of $m_0$ and $x_i^0$, $i = 1, \ldots, n$.

**Corollary.** Let $x_1, \ldots, x_k$ be any set of representatives from $x_1, \ldots, x_k$, and let $X = k^{-1}(x_1 + \cdots + x_k)$. Then

$$\sum_{i=1}^k \|x_i - X\|^2 \geq \sum_{i=1}^k \|x_i^0 - X_0\|^2.$$

**Theorem 3.** Let $X$ be a given n-vector and $X_i^1$ in $X_i$ be in optimal position to $X$ for $i = 1, \ldots, k$. Then

$$\sum_{i=1}^k \|X - X_i^1\|^2 \geq \sum_{i=1}^k \|X_i - X_i^1\|^2.$$

where $X_i = k^{-1}(x_1^1, \ldots, x_k^1)$.

**Proof:** The result is easy to establish, and it also provides an algorithm for computing the mean of $x_1, \ldots, x_k$.

**Algorithm:** Start with any vector, say $X_0 = x_1 \in X_1$, and find $x_1^1 \in X_1$ that is in optimal position to $x_1^0$. For $i = 1, \ldots, k$, $X_i = k^{-1}(x_1 + x_2 + \cdots + x_k)$ is therefore safe to repeat the algorithm with different starting vectors and choose the one with the smallest value of the sum of squares of the residuals.

The geometry developed for real circular vectors can be extended to complex circular vectors in an obvious way by using the norm of a complex vector space.

3. An Example

For illustrative purposes, we give a simple example and explain the computations involved. Consider the following vectors of size 3:

$$X_1 = (3, 2, 5), \quad X_2 = (1, 2, 3), \quad X_3 = (2, 5, 5).$$

To find the mean value of $X_1, X_2, X_3$, let us start with

$$X_0 = X_1 = (3, 2, 5)$$

and find vectors in $X_1, X_2, X_3$ in optimal position to $X_0$. They are

$$X_1^1 = (3, 2, 5), \quad X_2^1 = (1, 2, 3), \quad X_3^1 = (5, 2, 5),$$

giving the first approximation to the mean

$$X_1 = \frac{1}{3}(3, 2, 5) = (3, 2, 13/3),$$

which is an improvement over $X_0$ because

$$\sum_{i=1}^3 (X_0 - X_i^1)^2 \geq \frac{10}{3} = \sum_{i=1}^3 (X_1 - X_i^1)^2.$$

For the next iteration, we have to find $X_1^2 \in \tilde{X}_1, X_2^2 \in \tilde{X}_2, X_3^2 \in \tilde{X}_3$, which are in optimal position to $X_1$. They are

$$X_1^2 = (3, 2, 5), \quad X_2^2 = (1, 2, 3), \quad X_3^2 = (5, 2, 5),$$

giving

$$X_2 = \frac{1}{3}(3, 2, 5) = (3, 2, 13/3),$$

which is the same as $X_1$, so that no further improvement is possible. We accept $(3, 2, 13/3)$ as the mean value. In the above example, the iteration converged in two steps. In real situations when $k$ is large, the iteration may not converge in a finite number of steps, in which case we may have to stop when the improvement in the residual sum of squares is negligible.

Yasui (5) suggested an algorithm for computing the mean in a finite number of steps. A particular order of $X_1, \ldots, X_k$ of the measurements on $k$ objects is chosen. First, we find $X_2^0 \in \tilde{X}_2$, which is in optimal position to $X_1$, and compute the mean $m_1 = (X_1 + X_2^0)/2$. Next, we find $X_3^0 \in \tilde{X}_3$ that is in optimal position to $m_1$ and compute the mean $m_2 = (2m_1 + X_3^0)/3$. At the $r$th stage, we have the mean

$$m_r = (rm_{r-1} + X_1^0)/(r + 1)$$

for $r = 1, \ldots, n - 1$. The suggested mean is $m_{n-1}$. The method depends on the order in which the measurements on different objects is chosen and not all orders lead to the same mean value.

4. Applications

The measurements on the boundary of a closed figure are not scale free in all of the cases described in Section 1. If only the differences in shape are of interest, the measurements have to be standardized to have variance unity.

In the absence of a suitable stochastic model for circular vector data, one could use nonparametric distance-based methods in tests of hypotheses of equality of shape distributions in two or more populations as in ref. 11 and in discriminant and cluster analyses as in refs. 12–14.

**Remark:** If there are well specified landmarks $L_1, \ldots, L_m$ on the boundary, we may have a more informative description of the shape of the boundary by considering the following features. Let $O$ be the centroid of the boundary or one of the landmarks and obtain as the first set of features characterizing the configuration of landmarks the angles $\phi_i = \angle L_i OL_{i+1}$, $i = 1, \ldots, m$ with $L_{m+1} = L_1$. Now consider each piece of the boundary connecting the successive landmarks and take measurements of the type described in Section 1 at equal values of a suitably chosen argument. An example of such a description is given figure 7 of ref. 4. Then, standard multivariate techniques could be used.

The research for this paper is supported by the Army Research Office under Grant DAAH04-96-1-0082.