Iranian Journal of Mathematical Sciences and Informatics
Vol. 15, No. 1 (2020), pp 99-109
DOI: 10.21859/IJMSI.15.1.99

Comparing Model-based Versus K-means Clustering for the Planar Shapes

Hamed Jafari, Mousa Golalizadeh*

Department of Statistics, Tarbiat Modares University, Tehran, Iran.

E-mail: hamed.jafari@modares.ac.ir
E-mail: golalizadeh@modares.ac.ir

Abstract. In some fields, there is an interest in distinguishing different geometrical objects from each other. A field of research that studies the objects from a statistical point of view, provided they are invariant under translation, rotation and scaling effects, is known as the statistical shape analysis. Having some objects that are registered using key points on the outline of the objects, the main purpose of this paper is to compare two popular clustering procedures to cluster objects. We also use some indexes to evaluate our clustering application. The proposed methods are applied to the real life data.

Keywords: Shape, Clustering, K-means, Model-based, Landmark.

2010 Mathematics Subject Classification: 62H30, 62H99.

1. Introduction

In some fields there is an interest for distinguishing different images from each other which is known as image analysis or image clustering and classification. Generally speaking, it's not necessary to know if the image has specific shapes, such as human face, specific parts of human body, organic tissues and

*Corresponding Author

Received 09 May 2017; Accepted 07 March 2018

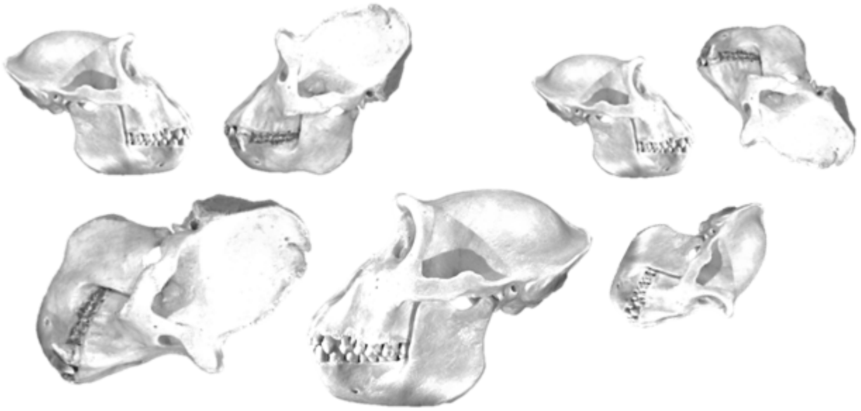


Figure 1. The pictures of some typical skulls of gorilla (quoted from https://commons.wikimedia.org). The objective is to cluster them in terms of their sizes and shapes.

son on. Instead, just the pixels are relatively compared in these cases and finally images are assigned into different groups with distinguishable properties for each group. Of course, if there is some information about particular aspects of the image, they can be helpful for further statistical analysis. As an example, consider the Figure 1, in which there are some pictures of gorilla skulls. It worths to mention that the data were first studied by [10]. One can grasp some information from these images in terms of their sizes and shapes.

One of the interesting questions in studying these skulls is: ”how can these pictures be put in two different groups?”. A field of research for answering such typical questions from statistical point of view has its root in multivariate analysis. However, those geometrical aspects of the object lie in the statistical shape analysis which was first introduced to the statistical communities by [8] in an interesting article titled “The Diffusion of Shape”.

In fact, data in the statistical shape analysis terminologies are the shape of the objects. Hence, particular geometrical aspects of the objects such as topological structure should be considered in the statistical analysis. One of main concepts of trivial statistical analysis for any real data sets is related to the way one views the data. Generally speaking, the observations can be viewed either as the realizations of the random variables generated from some particular distributions which leads to the parametric statistics, or can be seen as data from some unknown densities turning to the non-parametric statistics. These two views in clustering the objects; confined in the statistical shape analysis, are compared with each other in this article. The main is to discover how they are successful in clustering pre-labeled planar objects. In this paper, a background of the statistical shape analysis and the related definitions are first presented. Then, two clustering methods applied for the planar shapes are described. Some simulation studies along with real application of proposed methods are included in the final section. Moreover, general conclusions and some possible further researches arising from current study are given.

2. A Background on the Landmark-based Shape Analysis

It worths to recall the formal definition of shape before going forward on further analyzing the images of the objects. It has mainly been given for the statistical analysis of the shapes. Kendall [8] provided the following definition:

Definition 2.1. Shape is any geometrical information of object after removing translation, scale and rotation.

Based upon this definition, all homologous images are considered to be in one class and so can be represented as just one object. From this point of view, one encounters with the equivalent classes and the consequence mathematical theories arise. One can consult [3] for more details on this and other relevant topics.

Taking the images of the objects, one can follow different methods to study them statistically. At the first instance, one should define particular variables as the representative of the picture (image) of the objects. One of the simplest procedures to do this in the statistical shape analysis context is to set some points in specific parts of the image. Those points are called as the **landmarks** and the statistical analysis using them is known as the landmark-based analysis. The Cartesian coordinates of the points are the observations of the variables defined in the statistical shape analysis framework. A typical illustration of such method is shown in Figure 2. As expected one can work with the matrices constituting the Cartesian coordinates instead of the whole images. Each of such matrix is then called **configuration matrix**. At the end, one has a set of matrices and should analyze them using the common statistical tools.

According to the statistical definition of the shape given above, some steps should be followed to get the shapes of the objects. In another word, the configuration matrices are not defining the equivalent classes of the objects in their initial constructions. Invoking many mathematical operations, some particular transformations should be performed on the configuration matrices in order to achieve the shapes of the objects. Those relevant transformations are **similarity transformations**, which include translation, scale and rotation effects. Theoretically, these operations transfer the configuration matrices from the Euclidean space to a non-Euclidean space. This latter space is known as **shape space** and is usually identified by different systems of coordinates. Two well-known systems are Kendall [9], and Bookstein coordinates [1]. Those

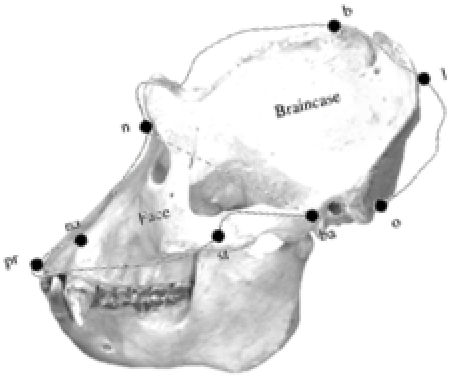


Figure 2. To set landmarks on the image of a gorilla skull (quoted from https://commons.wikimedia.org). The landmarks are used as inputs for further statistical analysis.

coordinates are now bases for statistical analysis such as clustering. To have an idea of employing these stages to get the shape coordinates, we explain those procedures applied for the Figure 2.

Suppose the landmarks set on the Figure 2 are represented in a configuration matrix, say A, as follows:

$$A = \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_8 & y_8 \end{bmatrix}.$$

By pre-multiplying the matrix A by Helmert sub-matrix, H, the translation effect is removed. Rotation and scale are omitted using rotation matrix and the common Euclidean norm, respectively. Generally, for removing all of similarity effects from the planar objects A, the following equality is useful:

$$HA \begin{pmatrix} x_2 & y_2 \\ y_2 & x_2 \end{pmatrix} \frac{1}{x_2^2 + y_2^2} = \begin{pmatrix} 1 \; u_3 \; ... \; u_k \\ 0 \; v_3 \; ... \; v_k \end{pmatrix}^T.$$

Now, the vector $\mathbf{u}$, i.e.

$$\mathbf{u} = (u_3, \, ..., \, u_k, v_3, \, ..., v_3)^T$$

is the Bookstein shape coordinates of the objects in the Figure 2 with the landmarks set in A.

As a remark, we can imagine the case in which the points in the equivalent classes can be separated into some cluster in which other categorical variables are assisting to separate the objects. In another word, the way in which the shapes of the objects are different in terms of one the nominal variables worths to study. To do this, one go through either parametric or non-parametric procedures. As a candidate for each scenarios, we confine ourselves to the model-based and k-means clustering in this paper. Furthermore, we just consider the planar shapes, i.e. 2-D objects.

3. Clustering the Planar Shapes

Clustering objects can be done using three distinct methods. For our shape analysis application, we utilize two of them below.

3.1. **Model-based clustering of shapes.** This approach for clustering shape data was just recently introduced by [7]. Below, we highlight the main theme of their research which is useful for our further statistical analysis.

As the first stage, one needs to model the random behavior of the landmarks for each objects by some distributions. Suppose the vector resulted from stacking columns of the matrix A has the multivariate normal distribution. It is shown in [4] that the distribution of **u** is then the offset normal with the density function

$$f_u(\mathbf{u};\mu,\Sigma)=\frac{|\Gamma|^{\frac{1}{2}}\exp(-g/2)}{(2\pi)^{k-2}|\Sigma|^{\frac{1}{2}}}\sum_{i=0}^{k-2}\begin{pmatrix} k-2 \\ i \end{pmatrix}E(\iota_x^{2i}|\xi_x,\sigma_x^2)E(\iota_y^{2k-4-2i}|\xi_y,\sigma_y^2),$$

where the components can be achieved by following sets of the formulas:

$$\begin{aligned}
&vec(AH)=Wh,\\
&\Gamma=(W^T\Sigma^{-1}W)^{-1},\\
&\upsilon=\Gamma W^T\Sigma^{-1}vec(\mu),\\
&g=vec(\mu)^T\Sigma^{-1}vec(\mu)-\upsilon^T\Gamma^{-1}\upsilon,\\
&\Gamma=\Psi D\Psi^T\\
&diag(\sigma_x^2,\sigma_y^2)=D,\\
&(\xi_x,\xi_x)^T=\Psi^T\upsilon.
\end{aligned}$$

Here, $E(\iota^p|b,c)$ is the p-th momentum of the normal distribution with the mean b and variance c, i.e. $N(b,c)$, which according to [11] can be derived using the following recursive equality:

$$E(\iota^{p+1}|b,c)=bE(\iota^p|b,c)+pcE(\iota^{p-1}|b,c).$$

The parameters in the density $f_u(\mathbf{u};\mu,\Sigma)$ can be estimated through employing the EM algorithm [2].

To invoke the model-based clustering of the planar shapes, it is assumed the observations, i.e. the shape coordinates, come from a mixture of distributions. In particular,

$$g_{\mathbf{u}}(\mathbf{u};\theta)=\sum_{m=1}^{M}\pi_{mi}f_u(\mathbf{u},\mu_m,\Sigma_m),$$

where each distribution shows a separate group and each observation is assumed to come only from one of M densities. By estimating the parameters π_{mi} and the other parameters in the model, one can assign the observations to each group in a trivial manner. More details on this can be found in [7].

3.2. **K-means clustering of shapes.** One of the popular non-parametric approach for clustering the objects is k-means clustering. Using this method, the number of clusters (groups), i.e. k, is fixed initially. So, one encounters with the observations set in k distinct clusters. Then, the centroid, usually means of shapes , for each group are computed. Obviously, the optimal situation is occurred on allocating each observation to the nearest group in sense which its distance with the centroid of the cluster is minimal. To reach this objective, one observation is repeatedly removed from each group and added to another cluster until there is no noticeable change in Within Sum of Squares of the Groups (WGSS). The main difference between various versions of the k-means clustering arises from the different distance measures used in computing the WGSS. There are some algorithms to handle this substitution and insertion in a proper way.

Since the shape data are members of a non-Euclidean space, calculating distance between them cannot be done using the common Euclidean metrics. Instead, some distances, which are appropriate to measure the difference between shape data and applicable in the shape space, should be utilized. A comprehensive account of those distances are provided in [4]. Below, we review some of them which are useful for our analysis. In all definitions given bellow, the X_1 and X_2 represent two objects, represented by their corresponding configuration matrices.

Definition 3.1. The full procrustes distance between X_1 and X_2 is

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

where $Z_i = \frac{HX_i}{\|HX_i\|}$.

Definition 3.2. The partial Procrustes distance between X_1 and X_2 is defined as

$$d_P(X_1, X_2) = \inf_{\Gamma \in SO(m)} \|Z_2 - Z_1 \Gamma\|.$$

Definition 3.3. Riemannian (Procrustes) distance on a manifold is defined as a choice of positive-definite inner product on each tangent space $T(m)$ at point m.

Definition 3.4. Tangent space distance between X_1 and X_2 is the distance between corresponding two points mapped into the planar tangent space which is the linearized version of the shape space in the vicinity of a particular point (usually mean shape) of shape space.

Definition 3.5. Size-and-shape distance is the distance between the size-and-shape forms of the configurations and is found by minimizing the Euclidean distance over rotations:

$$d_S(X_1, X_2) = \inf_{\Gamma \in SO(m)} \|HX_1 - HX_2\Gamma\|.$$

From practical point of view, we follow algorithm proposed by [5]. In fact, this algorithm works only for Euclidean distance and so it should be adapted for the shape data with the aforementioned distances.

4. Simulation Studies and Real Data Analysis

In order to compare two proposed methods on clustering the shapes data, we need to recall some relevant criteria. Those validation measures can be found in [6]. Among many of them, we consider one of the simple criteria in this paper. Coinciding the consequence of a clustering procedure with the distinctions of the data made by the initial categorical labels is the core root of the clustering validation indexes. This objective is properly adopted into the CVI, defined as follows

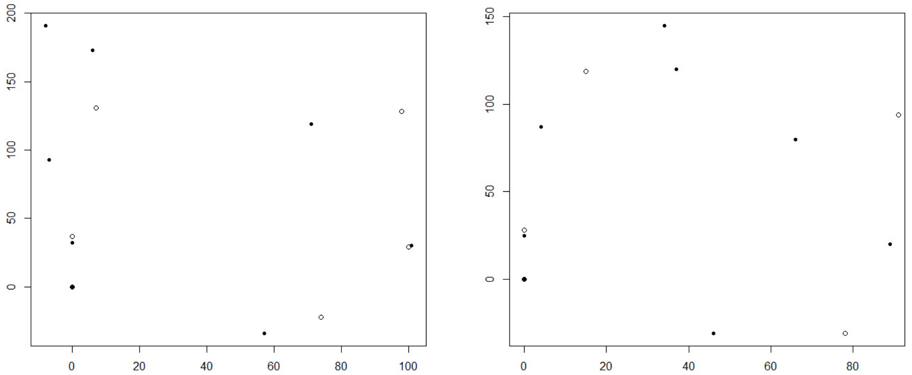
$$CVI = \frac{TCO}{TO} \times 100,$$

where TCO stands for the True Clustered Observations and TO for the Total Observations.

Next, we provide our simulation studies to compare two aforementioned shape clustering procedures.

4.1. **Simulation Studies.** In order to conduct the simulation studies, we assume that there are two types of data differentiated by the male and female labels. In fact, the objects were simulated in such a way that they can mimic the real data described in the next section. A schematic representation of these data can be seen in Figure 3. One can efficiently compute the sample mean and variance of the real data. Mathematical treatment of this is provided in [3] and codes to do this task are given in [4]. Then, one can simulate the data using those summary statistics. Hence, mimicking pattern within and between each group of shape data will occur. Note that simulated data can also be generated using the multivariate normal distribution with the mean and variance parameters to be the same as those for the real data. Then, the entire observations need to be transformed to the offset-normal density to assure having shape data. One can consulate [3] for more details.

The structure of our simulation studies are as follows. The 50 samples were simulated for the pair of gorilla and orangutan skulls identified with eight landmarks. Hence, there are 100 samples for each type of the skulls. First, the model-based clustering of the shape data was invoked on these objects and the CVI was derived. This scenario was repeated for 100 iterations. Then, the k-means clustering procedure was performed using the same data. The results



(a) (b)

FIGURE 3. The shape of simulated data. They are generated such that their shapes are mimicking the gorilla (a) and orangutan (b) skulls.

gained in this study are shown in the Table 1. As seen, the results for the Euclidean distance is also included for a proper comparison purposes.

TABLE 1. The values of CVI for the simulated data using different clustering methods and various shape distances. Data are mimicking the shape of two objects.

Method	**Distance**	**Gorilla-like**	**Orangutan-like**
Model-based	—	75.5	100.00
K-means	Full procrustes	87.08	73.64
	Riemannian	87.04	73.43
	Tangent space	87.84	72.03
	Partial procrustes	87.04	70.54
	Size-and-shape	94.51	96.00
	Euclidean	99.86	95.98

As the Table 1 shows we see different behaviour for two objects. For the gorilla-like simulated data, the model-based was weaker than the k-means even using various distances. However, this is not the case for the orangutan-like simulated data. Confining to the k-means clustering, the Euclidean distance outperforms the other distances for the gorilla-like data. For the orangutan-like data, the candidate is the size-and-shape distance. Generally, it might be argued that the size-and-shape distance is a reasonable distance to use whenever the size is playing the key role on identifying the geometrical structure of the objects. On the other hand, if this is not the case, one can ignore the structure of the shape space and switch to the common multivariate statistics via projecting the shape data into the Euclidean space in the vicinity of the mean shape.

4.2. **Real data analysis.** In this section, we consider the real data set to check the performance of two clustering methods. The data contain 59 samples (29 male and 30 female) for gorilla skulls and 54 samples (30 male and 24 female) for the orangutan. The data are available in the package `shapes`; freely available in the software `R`.

Unfortunately, the model-based method did not work and the algorithm failed repeatedly while calculating the results. However, the k-means method performed reasonably well. The results are appeared in the Table 2. The reported data in the table show that the size-and-shape outperforms the other distances, although there is not too much difference between the CVI of this and Euclidean distance for the gorilla data. However, this difference is remarkable for the gorilla shapes. The interesting point goes back to the same values of the CVI for the full procrustes and riemannian regardless of the type of the data. However, there is still room to investigate the difference of shape clustering via relating it to the relevant covariates. Also, adopting the available software packages that are implementing the clustering data for the shape objects which are members of the non-Euclidean space is vital. This will help the researchers to get more insights into the clustering of the objects and facilities to compare clustering methods.

Table 2. The value of CVI on clustering the real data sets using different shape distances in K-means algorithm.

Distance	**Gorilla**	**Orangutan**
Full procrustes	92.00	87.00
Riemannian	92.00	87.40
Tangent space	88.00	78.00
Partial procrustes	92.00	96.50
Size-and-shape	100.00	98.00
Eculidean	98.00	72.00

5. Conclusion and Future Works

Shape as the data play a great role in various scientific fields. The main objects of this paper was to study and compare two well-known clustering methods applied for the shape objects. These typical data are members of a non-Euclidean space and so invoking clustering methods require a great care. We investigated the methods in both simulation and real application studies.

The results gained from applying the model-based method show that the performance of clustering for simulated samples from normal distributions is promising. This means that, if we have observations with normal distribution (which can be evaluated using, for example, the normality tests), we can use model-based clustering. The interesting point is that if one is away from the normality assumption, the performance of the model-based clustering gets weaker.

The results gained from applying k-means method show that size-and-shape distance, among different distances, has better results. This might be because of considering size of the shapes which is important in differentiating between male and female sexes. It means that we are not allowed to remove all of the similarity effects to compare shapes whenever there is clear difference between groups. As seen, the investigation in this paper consists of two different sections, i.e. simulation study and real data analysis. Further, both real data sets are discriminated in terms of male and female genders. It means that we have two apparent clusters in our real-life data analysis. Moreover, we have followed the same structure inherited among the real data in our simulation study. So, there is no concern on selecting the numbers of clusters in our investigation. However, this is the case in where there is no prior knowledge on explicit numbers of the clusters. Discussion on this critical issue is out of the scope of our paper. There are some directions to extend the current research. A well-known distribution in the shape analysis context is the size-and shape offset normal. One can consider it for clustering the shape and conduct new research in this context rather than using the offset normal distribution in our study.

Acknowledgments

The authors would like to express their thanks to the referees for their useful comments and suggestions.

References

1. F. L. Bookstein, Size and Shape Spaces for Landmark Data in Two Dimensions, *Statistical Science*, **10**, (1986), 181-222.
2. A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society, Series B*, **39**, (1977), 1-38.
3. I. L. Dryden, K. V. Mardia, *Statistical Shape Analysis*, Wiley, Chichester, 1998.
4. I. L. Dryden, K. V. Mardia, *Statistical Shape Analysis: With Application in R*, Wiley, Chichester, 2016.
5. J. A. Hartigan, M. A. Wong, Algorithm AS 136: A k-means Clustering Algorithm, *Applied Statistics*, **28**, (1979), 100-108.
6. M. Halkidi, Y. Batistakis, M. Vazirgiannis, On Clustering Validation Techniques, *Journal of Intelligent Information Systems*, **17**, (2001), 107-145.

7. C. Huang, , C. Styner, H. Zhu, Clustering High-Dimensional Landmark-based Two-Dimensional Shape Data, *Journal of the American Statistical Association*, **110**, (2015), 946-961.
8. D. G. Kendall, The Diffusion of Shape, *Advances in Applied Probability*, **9**, (1977), 428-430.
9. D. G. Kendall, Shape Manifolds, Procrustean Metrics, and Complex Projective Spaces, *Bulletin of the London Mathematical Society*, **16**, (1984), 81-121.
10. P. O'Higgins, *A Morphometric Study of Cranial Shape in the Hominoidea*, Phd Dissertation, University of Leeds, Leeds, 1989.
11. R. Willink, Normal Moments and Hermite Polynomials. *Statistics and Probability Letters*, **73**, (2005), 271-275.