Nonlinear Phenomena in Complex Systems, vol. **23**, no. 2 (2020), pp. 251 - 253

Using the Alpha Geodesic Distance in Shapes K-Means Clustering

F. D. Oikonomou*
Department of Mathematics, University of Patras, Greece

A. De Sanctis
Department of Business Economics, University of Chieti-Pescara, Pescara, ITALY

(Received 30 November, 2019)

This paper is based mainly on the relevant work [1]. In that paper the authors studied the problem of clustering of different shapes using Information Geometry tools including, among others, the Fisher Information and the resulting distance. Here we are using the same methods but for the geodesics of the alpha connection for three different values of the alpha parameter.

AMS Subject Classification: 94A17, 62H30, 94A08
Keywords: Shape Analysis, clustering, information geometry, alpha connection, alpha geodesics, Fisher–Rao metric
DOI: https://doi.org/10.33581/1561-4085-2020-23-2-251-253

1. The alpha connection and the equations of the alpha geodesics

We are considering the $2D$ normal distribution:

$$p(x, y, \mu_1, \mu_2, \sigma_1, \sigma_2) \\ = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left\{ -\frac{1}{2}(\frac{x-\mu_1}{\sigma_1})^2 - \frac{1}{2}(\frac{x-\mu_2}{\sigma_2})^2 \right\}. \tag{1}$$

Let $\xi = (\mu_1, \mu_2, \sigma_1, \sigma_2)$. Then, if $\partial_i \equiv \frac{\partial}{\partial \xi^i}$, the Fisher metric is given by the formula [2] $g_{ij} = E(\partial_i l \partial_j l)$ where $l = l(x, y, \mu_1, \mu_2, \sigma_1, \sigma_2)$ $= \log p(x, y, \mu_1, \mu_2, \sigma_1, \sigma_2)$ and the mean value has been computed for the above $2D$ normal distribution. We have found that

$$g = \text{diag}\left[\frac{1}{\sigma_1^2}, \frac{1}{\sigma_2^2}, \frac{2}{\sigma_1^2}, \frac{2}{\sigma_2^2}\right]. \tag{2}$$

Next we have calculated the coefficients of the alpha-connection from the following formula

$$\Gamma^{(\alpha)}_{ij,k} = E\left[\left(\partial_i\partial_j l + \frac{1-\alpha}{2}\partial_i l \partial_j l\right)\partial_k l\right], \tag{3}$$

and consequently

$$\Gamma^{(\alpha)d}_{ab} = g^{cd}\Gamma^{(\alpha)}_{ab,c}, \tag{4}$$

where g^{cd} are the corresponding elements of the inverse matrix g^{-1}. Then the equations of the alpha geodesics $c^i = c^i(t)$ are

$$\frac{d^2c^i}{dt^2} + \Gamma^{(\alpha)i}_{jk}\frac{dc^j}{dt}\frac{dc^k}{dt} = 0 \tag{5}$$

and the distance from one point $\xi = (\mu_1, \mu_2, \sigma_1, \sigma_2)$ to $\xi' = (\mu_1', \mu_2', \sigma_1', \sigma_2')$ along a geodesic is given by the formula

$$d(\xi, \xi') = \int_c ds = \int_t^{t'} \sqrt{g_{ij}\frac{dc^i}{dt}\frac{dc^j}{dt}}dt, \tag{6}$$

where $c(t) = \xi$ and $c(t') = \xi'$.

2. Finding the geodesics numerically

In order to find the geodesic $c = c^i(t)$ from the point $\xi = (\mu_1, \mu_2, \sigma_1, \sigma_2)$ to the point $\xi' = (\mu_1', \mu_2', \sigma_1', \sigma_2')$, we suppose

$$c^i(t) = c^i(0) + t\frac{dc^i}{dt}\Big|_{t=0} + \frac{t^2}{2}\frac{d^2c^i}{dt^2}\Big|_{t=0} \tag{7}$$

[*] E-mail: pheconom@physics.upatras.gr

for $i = 1, 2, 3, 4$. I.e. we assume the Taylor expansion of c^i for up to the second order.

From the equations of geodesics we have

$$\frac{d^2c^i}{dt^2}\Big|_{t=0} = -\Gamma^{(\alpha)i}_{jk}\frac{dc^j}{dt}\frac{dc^k}{dt}\Big|_{t=0}, \tag{8}$$

so the above expansion takes the form

$$c^i(t) = c^i(0) + t\frac{dc^i}{dt}\Big|_{t=0} - \frac{t^2}{2}\left(\Gamma^{(\alpha)i}_{jk}\frac{dc^j}{dt}\frac{dc^k}{dt}\right)\Big|_{t=0} \tag{9}$$

with $c(0) = \xi$.

In order to have an approximate expression for the geodesic c^i we want to know the first order derivatives at $t = 0$. These can be computed if we demand

$$c^i(1) = \xi' = (\mu_1', \mu_2', \sigma_1', \sigma_2'), \tag{10}$$

so we have a non-linear system of four equations with four unknowns which we solve numerically. Finally,

$$d(\xi, \xi') = \int_0^1 \sqrt{g_{ij}\frac{dc^i}{dt}\frac{dc^j}{dt}}dt. \tag{11}$$

3. Numerical study

The use of Information geometry tools for clustering shapes has already been considered in several papers [3, 4] However, so far only the Fisher Information and the Wasserstein distance has been evaluated. In order to evaluate the cluster recovery and to test our algorithms using the alpha-geodesic and the corresponding distance defined on the statistical manifold, we consider a modified Gaussian perturbation model where the jth configuration is obtained as follows:

$$X_j = (\mu_{g_j} + E_j)\Gamma_j + 1_8\gamma_j^T, \tag{12}$$

where $j = 1, \ldots, 40$ and

- $\mu_{g_j} = \mu_1$ for $j = 1, \ldots, 20$, $\mu_{g_j} = \mu_2$ for $j = 21, \ldots, 40$ are mean values from the data of the rat calvarial data set [5], corresponding to the skull midsagittal section of 21 rats collected at ages of $7(\mu_1)$ and $14(\mu_2)$ days;
- E_j are 8x2 random error matrices simulated from the multivariate Normal distribution with mean value zero and covariance structure Σ_E;
- Γ_j is an orthogonal rotation matrix with angle $\theta_j = j\frac{2\pi}{40}$;
- $1_8 = (1, 1, 1, 1, 1, 1, 1, 1)^T$;
- $\gamma_j^T = (\gamma_{1j}, \gamma_{2j})$ with γ_{1j}, γ_{2j} real numbers uniformly produced in the range $[-2, 2]$.

Three types of covariance matrix Σ_E are considered as described in [1]:

- Isotropic with $\Sigma_E = \sigma I_8 \bigotimes \sigma I_2$;
- Heteroscedastic with $\Sigma_E = \mathrm{diag}(\sigma_1, \ldots, \sigma_8) \bigotimes \sigma I_2$;
- Anisotropic with $\Sigma_E = \sigma I_8 \bigotimes \mathrm{diag}(\sigma_x, \sigma_y)$ with $\sigma_x \neq \sigma_y$.

The two types of algorithms (type I and type II) are described in [1] in details. In type 0 algorithm we have considered constant covariances (equal to 13) for the calculation of alpha distances. For each covariant structure we simulated only 5 samples and, for each sample, we computed the adjusted rand index.

In Figures 1, 2, 3 we can see the adjusted rand index for three different clustering algorithms, an index close to 1 for the best clustering results or to 0 otherwise. In the so-called round case we have the same constant covariance when measuring the length of the alpha geodesic. In type I and type II we allow the variances to vary during the clustering procedure. We see that the best clustering method is for type I algorithm [1], in which the covariance over the rat data is taken into account and for the values 0.7 and 0.8 of alpha. The picture is similar for all cases, that is, for the isotropic, the heteroscedastic and the anisotropic case.

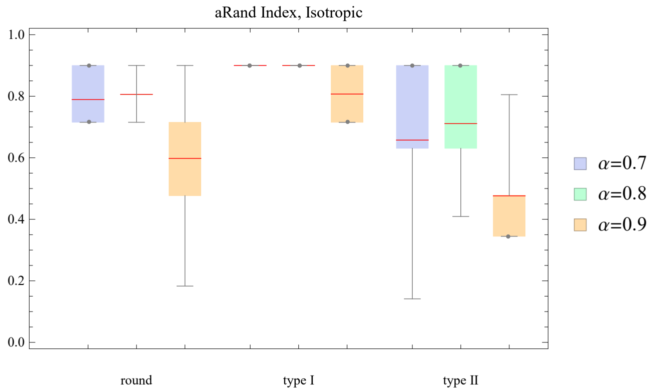

FIG. 1: (color online)

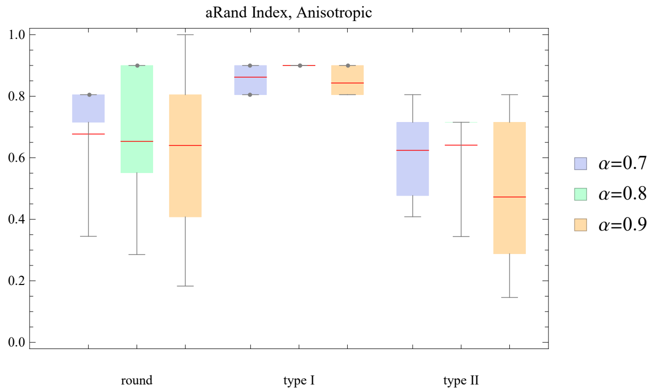

FIG. 3: (color online)

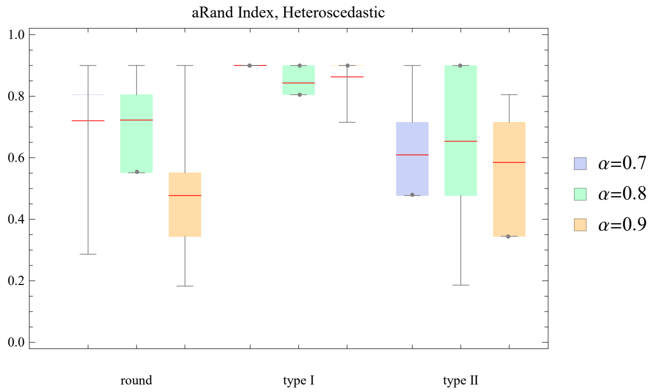

FIG. 2: (color online)

4. Further extensions of this work

This is a preliminary study. Obviously, to have better results, we need a bigger number of samples. Besides, a better Taylor expansion probably would help. Finally we have to study the behaviour of the above geodesics and the related distance under similarity transformations. All these will be an object of a future work.

п»ï

References

[1] S. A. Gattone, A. De Sanctis, S. Puechmorel, and F. Nicol, On the geodesic distance in shapes k-means clustering, Entropy **20** **(9)**, 647 (2018).

[2] S. Amari and H. Nagaoka, Methods of information geometry, Translations of mathematical monographs **191**, AMS & Oxford University Press, Providence (2000).

[3] A. De Sanctis and S. A. Gattone, Methods of information geometry to model complex shapes, European Physical Journal, Special Topics **225**, 1271 (2016).

[4] S. A. Gattone, A. De Sanctis, T. Russo, and D. Pulcini, A shape distance based on the Fisher-Rao metric and its application for shapes clustering, Phisica A **487**, 93 (2017) .

[5] F. L. Bookstein, *Morphometric Tools for Landmark Data: Geometry and Biology* (Cambridge University Press, 1991).