

On the geodesic distance in shapes K-means clustering

Stefano Gattone, Angela de Sanctis, Stéphane Puechmorel, Florence Nicol

► **To cite this version:**

Stefano Gattone, Angela de Sanctis, Stéphane Puechmorel, Florence Nicol. On the geodesic distance in shapes K-means clustering. Entropy, MDPI, 2018, Special Issue Selected Papers from 4th International Electronic Conference on Entropy and Its Applications, 20 (9), pp 647. 10.3390/e20090647 . hal-01852144v2

HAL Id: hal-01852144

<https://hal-enac.archives-ouvertes.fr/hal-01852144v2>

Submitted on 4 Oct 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Article

On the Geodesic Distance in Shapes K-means Clustering

Stefano Antonio Gattone ^{1,*} , Angela De Sanctis ², Stéphane Puechmorel ³ and Florence Nicol ³

¹ Department of Philosophical, Pedagogical and Economic-Quantitative Sciences, University “G. d’Annunzio” of Chieti-Pescara, 66100 Chieti, Italy

² Department of Business Economics, University “G. D’Annunzio” of Chieti-Pescara, 65127 Pescara, Italy; a.desanctis@unich.it

³ Ecole Nationale de l’aviation Civile (ENAC), Université Fédérale de Toulouse, FR-31055 Toulouse CEDEX, France; stephane.puechmorel@enac.fr (S.P.); nicol@recherche.enac.fr (F.N.)

Version October 4, 2018 submitted to Entropy

Abstract: In this paper, the problem of clustering rotationally invariant shapes is studied and a solution using Information Geometry tools is provided. Landmarks of a complex shape are defined as probability densities in a statistical manifold. Then, in the setting of shapes clustering through a K -means algorithm, the discriminative power of two different shapes distances are evaluated. The first, derived from Fisher–Rao metric, is related with the minimization of information in the Fisher sense and the other is derived from the Wasserstein distance which measures the minimal transportation cost. A modification of the K -means algorithm is also proposed which allows the variances to vary not only among the landmarks but also among the clusters.

Keywords: Shape Analysis; clustering; K -means algorithm; Fisher-Rao metric; wasserstein distance

1. Introduction

Shapes clustering is of interest in various fields such as geometric morphometrics, computer vision and medical imaging. In the clustering of shapes, it is important to select an appropriate measurement of distance among observations. In particular, we are interested in classifying shapes which derive from complex systems as expression of self-organization phenomenon. We consider objects whose shapes are based on landmarks [1–3]. These objects can be obtained by medical imaging procedures, curves defined by manually or automatically assigned feature points or by a discrete sampling of the object contours.

Since the shape space is invariant under similarity transformations, that is translations, rotations and scaling, the Euclidean distance on such a space is not really meaningful. In Shape Analysis [4], to apply standard clustering algorithms to planar shapes, the Euclidean metric has to be replaced by the metric of the shape space. Examples were provided in References [5,6], where the Procrustes distance was integrated in standard clustering algorithms such as the K -means. Similarly, Lele and Richtsmeier [7] applied standard hierarchical or K -means clustering using dissimilarity measures based on the inter-landmark distances. In a model-based clustering framework, Huang and Zhu [8] and Kume and Welling [9] developed a mixture model of offset-normal shape distributions.

In Shape Analysis, it is common to assume that the landmark coordinates have an isotropic covariance structure [4]. To relax the isotropic assumption, a bivariate Gaussian model was proposed to describe the landmarks of a planar shape [10,11], where the means are the landmark geometric coordinates and

capture uncertainties that arise in the landmark placement while the variances derive from the natural variability across the population of shapes. The novelty of this shape representation is given by the fact that variances are considered as additional coordinates for the landmarks of a shape. According to Information Geometry, the space of bivariate Gaussian densities is considered as a statistical manifold [12,13] with the local coordinates given by the model parameters. In this way, distances between landmarks can be defined using the geodesic distances induced by different Riemannian metrics.

In this paper, we consider the Fisher–Rao and the Wasserstein metrics as Riemannian metrics on the statistical manifold of the Gaussian densities. The geodesic distance induced by the Fisher–Rao metric is related to the minimization of information in the Fisher sense while the Wasserstein distance is related to the minimal transportation cost. Applications of geodesics to shape clustering techniques have also been provided in a landmark-free context [14,15].

As is well known, any hierarchical clustering algorithm uses as input the pairwise distances of all possible pairs of objects under study. Using the geodesic distances induced by Wasserstein and Fisher–Rao metrics, in References [10,11], a hierarchical clustering algorithm which allows the variances to vary among the landmarks was proposed.

In this paper, the discriminative power of these shapes distances is evaluated in the setting of shapes K -means clustering which is easier to implement and computationally faster. Furthermore, a modification of the K -means algorithm is proposed which allows the variances to vary not only among the landmarks but also among the clusters. The simulation results show that the proposed algorithm is able to cope with the effects of anisotropy in the landmark variances across different clusters.

2. Geometrical Structures for a Manifold of Probability Distributions

We call “manifold” a geometric object which is locally Euclidean then described by local coordinates. Manifolds can be used to study patterns from complex systems. Since pattern recognition essentially relies on quantitative assessment of the proximity of points, for the comparison of patterns, we need a well-suited similarity measure (distance or divergence). From Differential Geometry, we know that a Riemannian metric on a differential manifold X is induced by a metric matrix g , which defines an inner product on every tangent space of the manifold as follows: $\langle u, v \rangle = u^T g_{ij} v$ with associated norm $\|u\| = \sqrt{\langle u, u \rangle}$. Then, the distance between two points P, Q of the manifold is given by the minimum of the lengths of all the piecewise smooth paths γ joining these two points. Precisely, the length of a path is calculated using the inner product,

$$\text{Length of } \gamma = \int \|\gamma'(t)\| dt$$

thus

$$d(P, Q) = \min_{\gamma} \{\text{Length of } \gamma\}.$$

A curve that encompasses this shortest path is called a Riemannian geodesic and the previous distance is named geodesic distance. We remark that in general the concept of geodesic is related to connections defined on a manifold. If a connection is not Riemannian, then a geodesic is different from a shortest path.

Probability theory, in the presence of non-deterministic phenomena, provides a natural description of the raw data. Each measurement x is regarded as a sample from an underlying probability distribution of the measurement characterized by its probability density function $p(x|\theta)$. Measurements described by the distribution parameters, θ , may contain more information than a measurement expressed as a value and an associated error bar. Therefore, we apply pattern recognition methods directly in the space of probability distributions. Let P be a family of probability density functions $p(x|\theta)$ parameterized by $\theta \in \mathbf{R}^k$. It is well known that we can endow it with a structure of manifold, called statistical manifold, whose local

59 coordinates are the parameters of the family. As an example, we consider the family of p -variate Gaussian
60 densities:

$$f(x \mid \theta = (\mu, \Sigma)) = (2\pi)^{-\frac{p}{2}} (\det \Sigma)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right\}$$

61 where $x = (x_1, x_2, \dots, x_p)^T$, $\mu = (\mu_1, \mu_2, \dots, \mu_p)^T$ is the mean vector and Σ the covariance matrix. Note that
62 the parameter space has dimension $k = p + \frac{p(p+1)}{2}$. In particular, we are interested in the case $p = 2$.

63 Two geometrical structures have been extensively studied for a manifold of probability distributions.
64 One is based on the Fisher information metric (Fisher–Rao metric), which is invariant under reversible
65 transformations of random variables, while the other is based on the Wasserstein distance of optimal
66 transportation, which reflects the structure of the distance between random variables.

67 In the statistical manifold of bivariate Gaussian densities, we consider these two different Riemannian
68 metrics which in turn induce two types of geodesic distances.

69 *Fisher–Rao Metric for Gaussian Densities*

70 The geometry of the Gaussian manifold endowed with the Fisher–Rao metric was intensively studied
71 in References [16,17]. To avoid considering manifolds with boundaries, it is convenient to assume that all
72 densities are non-degenerate, thus the covariance matrices are invertible. In this case, one can define the
73 manifold of n -dimensional Gaussian densities as the set $\mathbb{R}^n \times \mathbb{R}^{n(n+1)/2} = \mathbb{R}^{n+n(n+1)/2}$ with local charts
74 given by the obvious identification $N_n(\mu, \Sigma) \mapsto (\mu_{i,i=1\dots n}, \sigma_{ij,i=1\dots n, j \leq n})$, where the σ_{ij} are the elements of
75 the matrix Σ . A tangent vector at a point (μ, Σ) of the manifold is just a vector from $\mathbb{R}^{n+n(n+1)/2}$. While
76 quite tractable, this choice of parameterization does not give any insight about the structure of the manifold.
77 A more enlightening approach is obtained by considering groups of transformations, as detailed below.

Let $\text{symm}^+(n)$ be a group of symmetric positive definite matrices of size $n \times n$ endowed with the product [18]:

$$(A, B) \mapsto A \circ B = A^{1/2} B A^{1/2} \quad (1)$$

78 and let us denote, using a common abuse of notation, the group of translations of \mathbb{R}^n also by \mathbb{R}^n .

Now, define the group $G(n)$ as the semi-direct product:

$$G(n) = \text{symm}^+(n) \ltimes_{\rho} \mathbb{R}^n \quad (2)$$

where the action ρ of $\text{symm}^+(n)$ on \mathbb{R}^n is given by left multiplication with the square root of the matrix, namely:

$$\rho(A)u = A^{1/2}u, A \in \text{symm}^+(n), u \in \mathbb{R}^n \quad (3)$$

79 In the sequel, we are dropping the ρ subscript in the semi-direct product and assume it implicitly.

80 An element in $G(n)$ can be represented as a couple (A, u) with $A \in \text{symm}^+(n)$, $u \in \mathbb{R}^n$. The group
81 product is obtained from the action ρ as $(A, u) \cdot (B, v) = (A^{1/2} B A^{1/2}, A^{1/2} v + u)$.

The inverse of an element (A, u) is given by $(A^{-1}, -A^{-1/2}u)$. The group $G(n)$ is a Lie group with Lie algebra $\mathfrak{g}(n) = \text{symm}^+(n) \oplus \mathbb{R}^n$ with $\text{symm}^+(n)$ the vector space of symmetric matrices. Finally, the left translation by an element (A, u) is the mapping:

$$(B, v) \mapsto L_{(A,u)}(B, v) = (A, u) \cdot (B, v) \quad (4)$$

Being an affine map, its derivative is its linear part. The Frobenius inner product on the space of square matrices of dimension n , defined as $\langle A, B \rangle = \text{tr}(A^t B) = \text{tr}(AB^t)$, jointly with the standard euclidean inner product on \mathbb{R}^n , induces a left invariant metric by:

$$\langle\langle (X, \eta), (Y, \xi) \rangle\rangle_{(A, \mu)} = K \text{tr} \left(A^{-1/2} X A^{-1} Y A^{-1/2} \right) + \eta_1^t A^{-1} \eta_1 \quad (5)$$

82 where $(X, \eta), (Y, \xi)$ are tangent vectors to $G(n)$ at (A, μ) and $K > 0$ is a fixed scaling factor that may be
83 arbitrary chosen to balance the relative contributions of the matrix part and the translation part.

It turns out that the metric obtained that way is exactly the Fisher–Rao metric on the manifold of multivariate Gaussian densities. Using the notations of Skovgaard [16], the length element of the Fisher–Rao metric g_F is:

$$ds^2 = \frac{1}{2} \text{tr} \left(\Sigma^{-1} X \Sigma^{-1} X \right) + \eta^t \Sigma^{-1} \eta \quad (6)$$

84 with (X, η) a tangent vector at (Σ, μ) .

85 The expression of ds^2 is the one of a warped product metric [19], which allows some simplifications
86 when computing the geodesics between two densities with same means.

87 A closed form for the geodesic distance between two densities with diagonal covariance matrices
88 may also be obtained as follows [17]:

$$d_F(\theta, \theta') = \sqrt{2 \sum_{i=1}^2 \left(\ln \frac{|\left(\frac{\mu_i}{\sqrt{2}}, \sigma_i\right) - \left(\frac{\mu'_i}{\sqrt{2}}, -\sigma'_i\right)| + \left|\left(\frac{\mu_i}{\sqrt{2}}, \sigma_i\right) - \left(\frac{\mu'_i}{\sqrt{2}}, \sigma'_i\right)\right|}{\left|\left(\frac{\mu_i}{\sqrt{2}}, \sigma_i\right) - \left(\frac{\mu'_i}{\sqrt{2}}, -\sigma'_i\right)\right| - \left|\left(\frac{\mu_i}{\sqrt{2}}, \sigma_i\right) - \left(\frac{\mu'_i}{\sqrt{2}}, \sigma'_i\right)\right|} \right)^2} \quad (7)$$

89 where $\theta = (\mu, \Sigma)$ with $\mu = (\mu_1, \mu_2)$ and $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2)$, $\theta' = (\mu', \Sigma')$ with $\mu' = (\mu'_1, \mu'_2)$ and $\Sigma' =$
90 $\text{diag}((\sigma'_1)^2, (\sigma'_2)^2)$.

91 For general Gaussian densities with Σ any symmetric positive definite covariance matrix, a closed form
92 for the geodesic distance is not known and one has to solve numerically a system of differential equations:

$$D_{tt}\mu - D_t \Sigma \Sigma^{-1} D_t \mu = 0 \quad (8)$$

$$D_{tt}\Sigma + D_t \mu D_t \mu^t - D_t \Sigma \Sigma^{-1} D_t \Sigma = 0 \quad (9)$$

93 where the expression D_t (respectively, D_{tt}) stands for derivative (respectively, second derivative) with
94 respect to t . A geodesic between two densities can be found by a shooting approach, which starts with
95 one density as an initial condition to the system in Equation (8) and iteratively adjusts the initial speed
96 vector of the curve so as to reduce the distance to the target density until the desired accuracy is reached. A
97 collocation algorithm can also be used, and is a common choice for solving ordinary differential equations
98 with boundary conditions. It is generally more stable than the shooting method, but may require more
99 computations. In both cases, a tricky part of the process is to ensure that the Σ matrix remains positive
100 definite. A rewrite of Equation (8) with the Cholesky decomposition $\Sigma = L^t L K$ allows this condition to
101 be satisfied by design and is the preferred choice. Another option to get an approximate value is to use
102 Equation (7) after diagonalizing the covariance matrices.

In regard to the Riemannian metric g_w which induces the Wasserstein distance [20], for Gaussian densities, the explicit expression of the distance is the following:

$$d_W(\theta, \theta') = |\mu - \mu'| + \text{tr}(\Sigma) + \text{tr}(\Sigma') - 2 \text{tr}(\sqrt{\Sigma^{\frac{1}{2}} \Sigma' \Sigma^{\frac{1}{2}}}) \quad (10)$$

103 where $|\cdot|$ is the euclidean norm and $\Sigma^{\frac{1}{2}}$ is defined for a symmetric positive definite matrix Σ so that
 104 $\Sigma^{\frac{1}{2}} \cdot \Sigma^{\frac{1}{2}} = \Sigma$. We remark that, if $\Sigma = \Sigma'$, the Wasserstein distance reduces to the Euclidean distance.

105 Otto [20] proved that, with respect to the Riemannian metric which induces the Wasserstein distance,
 106 the manifold of Gaussian densities has non-negative sectional curvature. We deduce that the Wasserstein
 107 metric is different from the Fisher–Rao metric. Indeed, for example, in the univariate case, the statistical
 108 manifold of Gaussian densities with the Fisher–Rao metric can be regarded as the upper half plane with
 109 the hyperbolic metric, which has negative curvature as it is well known.

110 Once a distance is defined, it can be used for clustering on a manifold. It is proven that the distance
 111 induced from Fisher–Rao metric and Wasserstein distance are in the more general class of Bregman
 112 divergences defined by a convex function [21]. For this class, a theorem states [22] that the centroid for a
 113 set of n points $\theta_i, i = 1, 2, \dots, n$ in the statistical manifold of the Gaussian densities is the Euclidean mean
 114 $\frac{1}{n} \sum_{i=1}^n \theta_i$. We use this result in the next section where a K -mean shapes clustering algorithm is defined
 115 using geodesic distances.

116 3. Clustering of Shapes

117 We consider only planar objects, as for example a flat fish or a section of the skull. The shape of the
 118 object consists of all information invariant under similarity transformations, that is translations, rotations
 119 and scaling [4]. Data from a shape are often realized as a set of points. Many methods allow to extract
 120 a finite number of points, which are representative of the shape and are called landmarks. One way
 121 to compare shapes of different objects is to first register them on some common coordinate system for
 122 removing the similarity transformations [2,23]. Alternatively, Procrustes methods [24] may be used in
 123 which objects are scaled, rotated and translated so that their landmarks lie as close as possible to each
 124 other with respect to the Euclidean distance.

Suppose we are given a planar shape configuration, S , consisting of a fixed number K of
 labeled landmarks

$$S = \{\mu_1, \mu_2, \dots, \mu_K\}$$

with generic element $\mu_k = \{\mu_{k1}, \mu_{k2}\}$ for $k = 1, \dots, K$. Following Gattone et al. [10], the k -th landmark,
 for $k = 1, \dots, K$, may be represented by a bivariate Gaussian density as follows:

$$f(x | \theta_k = (\mu_k, \Sigma_k)) = (2\pi)^{-1} (\det \Sigma_k)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)\right\} \quad (11)$$

with x being a generic 2-dimensional vector and Σ_k given by

$$\Sigma_k = \text{diag}(\sigma_{k1}^2, \sigma_{k2}^2) \quad (12)$$

125 where $\sigma_k^2 = (\sigma_{k1}^2, \sigma_{k2}^2)$ is the vector of the variances of μ_k .

126 We remark that, in the previous representation, the means represent the geometric coordinates of
 127 the landmark and capture uncertainties that arise in the landmark placement. The variances are hidden
 128 coordinates of the landmark and reflect its natural variability across a population of shapes. Equation (11)
 129 allows assigning to the k th landmark the coordinates $\theta_k = (\mu_k, \sigma_k)$ on the four-dimensional manifold
 130 which is the product of two upper half planes.

131 Let S and S' two planar shapes registered on a common coordinate system using Procrustes method.
 132 We parameterize them as follows: $S = (\theta_1, \dots, \theta_K)$ and $S' = (\theta'_1, \dots, \theta'_K)$.

133 The distances between landmarks allow defining a distance of the two shapes S and S' . Precisely,
 134 a shape metric for measuring the difference between S and S' can be obtained by taking the sum of the
 135 geodesic distances between the corresponding landmarks, according to the following definition:

$$D(S, S') = \sum_{k=1}^K d(\theta_k, \theta'_k) \quad (13)$$

Please note that this expression is not the geodesic distance on the product manifold that one would have expected from the landmark model. This last distance is given by:

$$D(S, S') = \sqrt{\sum_{k=1}^K d(\theta_k, \theta'_k)^2} \quad (14)$$

136 and is a L^2 distance instead of Equation (13) that is L^1 . It turns out that, according to simulations done, the
137 L^1 approach is more robust and gives all the time better clusterings.

138 Then, a classification of shapes, using in turn, as distance d , the distance d_F induced from Fisher–Rao
139 metric and the Wasserstein distance d_W , can be done following the standard methodology. In particular,
140 the K -means clustering procedure allows the variances to vary step by step in each cluster fitting better
141 real shape data.

142 4. K-Means Clustering Algorithm

143 The proposed shape distances are implemented in two different K -means algorithms: Type I and
144 Type II. While in the Type I algorithm the landmark coordinates variances are assumed isotropic across the
145 clusters, in Type II the variances are allowed to vary among the clusters.

146 Our task is clustering a set of n shapes, S_1, S_2, \dots, S_n into G different clusters, denoted as
147 C_1, C_2, \dots, C_G .

148 4.1. Type I Algorithm

149 1 Initial step:

150 Compute the variances of the k -th landmark coordinates $\sigma_k^2 = (\sigma_{k1}^2, \sigma_{k2}^2)$, for $k = 1, \dots, K$.

151 Randomly assign the n shapes, S_1, S_2, \dots, S_n into G clusters, C_1, C_2, \dots, C_G .

152 For $g = 1, \dots, G$, calculate the cluster center $c_g = (\theta_1^g, \dots, \theta_K^g)$ with k -th component $\theta_k^g = (\mu_{gk}, \sigma_k^2)$
153 obtained as $\theta_k^g = \frac{1}{n_g} \sum_{i \in C_g} \theta_k^i$, where n_g is the number of elements in the cluster C_g and θ_k^i is the k -th
154 coordinate of S_i given by $\theta_k^i = (\mu_{ik}, \sigma_k^2)$.

155

2 Classification:

For each shape S_i , compute the distances to the G cluster centers c_1, c_2, \dots, c_G .

The generic distance between the shape S_i and the cluster center c_g is given by:

$$D(S_i, c_g) = \sum_{k=1}^K d(\theta_k^i, \theta_k^g)$$

where the distance d could be the distance d_F induced from Fisher–Rao metric or the Wasserstein distance d_W .

Assign S_i to cluster h that minimizes the distance:

$$D(S_i, c_h) = \min_g D(S_i, c_g).$$

156 3 Renewal step:

- 157 Compute the new cluster centers of the renewed clusters c_1, \dots, c_G .
 158 The k -th component of the g -th cluster center c_g is defined as $\theta_k^g = \frac{1}{n_g} \sum_{i \in C_g} \theta_k^i$.
 159 4 Repeat Steps 2 and 3 until convergence [22].

160 4.2. Type II Algorithm

- 161 1 *Initial step:*
 162 Randomly assign the n shapes, S_1, S_2, \dots, S_n into G clusters, C_1, C_2, \dots, C_G .
 163 In each cluster, compute the variances of the k -th landmark coordinates $\sigma_{gk}^2 = (\sigma_{gk_1}^2, \sigma_{gk_2}^2)$, for
 164 $k = 1, \dots, K$ and $g = 1, \dots, G$.
 165 Calculate the cluster center $c_g = (\theta_1^g, \dots, \theta_K^g)$ with k -th component $\theta_k^g = (\mu_{gk}, \sigma_{gk}^2)$ obtained as
 166 $\theta_k^g = \frac{1}{n_g} \sum_{i \in C_g} \theta_k^i$ for $g = 1, \dots, G$, where n_g is the number of elements in the cluster C_g and
 167 $\theta_k^i = (\mu_{ik}, \sigma_{gk}^2)$ for $i \in C_g$.
 168
- 2 *Classification:*
 For each shape S_i , compute the distances to the G cluster centers c_1, c_2, \dots, c_G .
 The generic distance between the shape S_i and the cluster center c_g is given by:

$$D(S_i, c_g) = \sum_{k=1}^K d(\theta_k^i, \theta_k^g)$$

where the distance d could be the distance d_F induced from Fisher–Rao metric or the Wasserstein distance d_W .

Assign S_i to cluster h that minimizes the distance:

$$D(S_i, c_h) = \min_g D(S_i, c_g).$$

- 169 3 *Renewal step:*
 170 Update the variances of the k -th landmark coordinates in each cluster by computing $\sigma_{gk}^2 = (\sigma_{gk_1}^2, \sigma_{gk_2}^2)$,
 171 for $k = 1, \dots, K$ and for $g = 1, \dots, G$.
 172 Calculate the new cluster centers of the renewed clusters c_1, \dots, c_G .
 173 The k -th component of the g -th cluster center c_g is defined as $\theta_k^g = \frac{1}{n_g} \sum_{i \in C_g} \theta_k^i$.
 174 4 Repeat Steps 2 and 3 until convergence [22].

175 5. Numerical Study

The purpose of the simulation study was to evaluate the cluster recovery of the proposed shape K -means algorithm and to test its sensitiveness with respect to different shape distances defined on the manifold of the probability distributions. The shapes were simulated according to a Gaussian perturbation model where the i th configuration is obtained as follows:

$$X_{ig} = (\mu_g + E_i)\Gamma_i + \mathbf{1}_K \gamma_i^T \quad (15)$$

176 where

- 177 • E_i are zero mean $K \times 2$ random error matrices simulated from the multivariate Normal distribution
- 178 with covariance structure Σ_E ;
- 179 • μ_g is the mean shape for cluster g ;
- 180 • Γ_i is an orthogonal rotation matrix with an angle θ uniformly produced in the range $[0, 2\pi]$; and

- 181 • γ_i^T is a 1×2 uniform translation vector in the range $[-2, 2]$.

182 Three types of covariance structures are considered:

- 183 • Isotropic with $\Sigma_E = \sigma \mathbf{I}_K \otimes \sigma \mathbf{I}_2$
 184 • Heteroscedastic with $\Sigma_E = \text{diag}[\sigma_1, \sigma_2, \dots, \sigma_K] \otimes \sigma \mathbf{I}_2$
 185 • Anisotropic with $\Sigma_E = \sigma \mathbf{I}_K \otimes \text{diag}[\sigma_x, \sigma_y]$ with $\sigma_x \neq \sigma_y$

186 The data were generated from the model in Equation (15) with sample size $n = 100$ and the number
 187 of clusters equal to $G = 2$. The mean shapes in each cluster were taken from the rat calvarial dataset [1]
 188 corresponding to the skull midsagittal section of 21 rats collected at ages of 7 and 14 days. In the isotropic
 189 case, σ was equal to 13. In the heteroscedastic case, the values of $\sigma_1, \sigma_2, \dots, \sigma_K$ were equal to 13 for 3
 190 randomly chosen landmarks and equal to 1.3 for the remaining 5 landmarks of each shape. Finally, the
 191 anisotropic case was simulated by setting $\sigma_x = 13$ and $\sigma_y = 1.3$ in one cluster and $\sigma_x = 1.3$ and $\sigma_y = 13$ in
 192 the other cluster.

193 Examples of the simulated data under the different covariance structures are provided in Figures 1–3.
 194 In the isotropic case (Figure 1), all landmark coordinates exhibit the same independent spherical variation
 195 around the mean. In the heteroscedastic case (Figure 2), the spherical variation is allowed to vary between
 196 the landmarks and the clusters. Finally, Figure 3 shows the anisotropic case where the variability of the
 197 landmark coordinates is different in the horizontal and vertical directions.

198 The shape K -means algorithm is implemented by using:

- 199 • The Fisher–Rao distance under the round Gaussian model representation (dr) where the variance is
 200 considered as a free parameter, which is isotropic across all the landmarks
 201 • The Fisher–Rao distance under the diagonal Gaussian model representation (dd1 (Type I K -means
 202 algorithm) and dd2 (Type II K -means algorithm))
 203 • Wasserstein distance (dp)

`./figures/isotropic.pdf`

Figure 1. Independent spherical variation around each mean landmark.

`./figures/heteroscedastic.pdf`

Figure 2. Heteroscedastic variation around each mean landmark.

`./figures/anisotropic.pdf`

Figure 3. Anisotropy in the x and y directions around each mean landmark.

204 For each covariance structure, we simulated 150 samples and, for each sample, we computed the
 205 adjusted Rand index [25] of each clustering method. The adjusted Rand index is a measure of agreement
 206 between two partitions. It ranges from about 0 when the compared partitions are completely random to 1
 207 when they are equal. The index is generally used as a measure of cluster recovery, the closer the index to 1
 208 the better the clustering results.

209 Figures 4–6 display the boxplots of the adjusted Rand index over 150 simulated samples for each
 210 clustering method. Outliers are plotted individually with the + symbol. When the covariance structure
 211 is isotropic (Figure 4), all distances show a similar behavior. In particular, the Fisher–Rao distance with
 212 round Gaussian distribution (dr) and the Wasserstein distance (dw) yield the best clustering results with
 213 median values of the adjusted Rand index both equal to 0.96 versus 0.85 and 0.88 obtained by the diagonal
 214 Gaussian distribution with Type I (dd1) and Type II (dd2) algorithms, respectively. In the heteroscedastic
 215 setting (Figure 5), both the Fisher–Rao with the round Gaussian distribution (median adjusted Rand
 216 index equal to 0.54) and the Wasserstein distance (median adjusted Rand index equal to 0.43) perform

217 poorly in comparison to the Fisher–Rao distance based on the diagonal distribution. As expected, the
 218 models which take into account different landmark variances (dd1-Type I algorithm) and also differences
 219 in the variances between the clusters (dd2-Type II algorithm) show a very good behavior with median
 220 values of the adjusted Rand index equal to 0.84 and 1, respectively. A very similar pattern is observed
 221 when anisotropy is also added in the covariance structure (Figure 6). As expected, the Type II algorithm
 222 significantly increases the computational time needed to meet the convergence condition since in each
 223 iteration both the means and the variances of the cluster centers have to be updated. On average, the Type
 224 II algorithm’s running time is approximately 20 times longer than Type I.

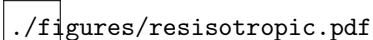
./figures/resisotropic.pdf

Figure 4. Isotropic case: Boxplots of the adjusted Rand index over 150 simulated samples for each clustering method; aRand index median values are 0.96 (dr), 0.85 (dd1), 0.88 (dd2), 0.96 (dw).

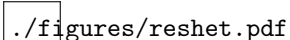
./figures/reshet.pdf

Figure 5. Heteroscedastic case: Boxplots of the adjusted Rand index over 150 simulated samples for each clustering method; aRand index median values are 0.54 (dr), 0.84 (dd1), 1.00 (dd2), and 0.43 (dw).

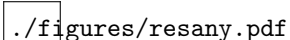
./figures/resany.pdf

Figure 6. Anisotropic case: Boxplots of the adjusted Rand index over 150 simulated samples for each clustering method; aRand index mean values are 0.51 (dr), 0.79 (dd1), 1 (dd2), and 0.44 (dw).

225 6. Conclusions

226 In this study, Information Geometry was used as a useful tool in the area of shape clustering. We first
 227 described a shape representing each landmark by a Gaussian model using the mean and the variance as
 228 coordinates, reflecting the geometrical shape of the configuration and the variability across a family of
 229 patterns, respectively. Within this framework, we considered the Fisher–Rao and the Wasserstein metric
 230 for quantifying the difference between two shapes.

231 Two version of the Fisher–Rao metric were proposed, depending on how the variances in the data are
 232 employed. In one case (round Gaussian distribution model), the variance was considered a free parameter
 233 that is isotropic across all the landmarks. In the second case, the isotropic assumption was relaxed allowing
 234 the variances to vary among the landmarks (diagonal Gaussian distribution model).

235 The results of the numerical study have shown that the violation of the isotropic assumption on the
 236 landmarks variability may cause a severe loss in the clustering recovery. Indeed, this assumption is rarely
 237 satisfied in practice where it is regularly seen that landmarks have different variances. In such a case, the
 238 relative importance among landmarks must be taken into account in the similarity measure adopted in the
 239 clustering algorithm. The proposed geodesic distance under the diagonal Gaussian model representation
 240 is able to face this problem. A further assumption that may be violated is that in all clusters the landmarks
 241 coordinates have a common covariance matrix. To cope with this issue, a new K -means shape algorithm
 242 was implemented that allows for differences among the clusters in the landmark coordinates variability.

243 Other extensions of the current work deserve further investigation, for example, the use of geodesics
 244 in the case of the general multivariate Gaussian model and considering more general shape measures,
 245 such as α -divergences.

246 **Author Contributions:** Although all authors contributed equally to the research, in particular A.D.S. proposed the
247 idea and the methodology, S.P. and F.N. improved the theoretical derivation and S.A.G. implemented the algorithms
248 and performed the numerical calculations.

249 **Funding:** This research received no external funding.

250 **Conflicts of Interest:** The authors declare no conflict of interest.

251 References

- 252 1. Bookstein, F.L. *Morphometric Tools for Landmark Data: Geometry and Biology*; Cambridge University Press:
253 Cambridge, UK, 1991.
- 254 2. Kendall, D.G. Shape manifolds, Procrustean metrics and complex projective spaces. *Bull. Lond. Math. Soc.* **1984**,
255 *16*, 81–121. [[CrossRef](#)]
- 256 3. Cootes, T.; Taylor, C.; Cooper, D.H.; Graham, J. Active shape models-their training and application. *Comput. Vis.*
257 *Image Underst.* **1995**, *61*, 38–59. [[CrossRef](#)]
- 258 4. Dryden, I.L.; Mardia, K.V. *Statistical Shape Analysis*; John Wiley & Sons: London, UK, 1998.
- 259 5. Stoyan, D.; Stoyan, H. A further application of D.G. Kendall’s Procrustes Analysis. *Biom. J.* **1990**, *32*, 293–301.
260 [[CrossRef](#)]
- 261 6. Amaral, G.; Dore, L.; Lessa, R.; Stosic, B. K-means Algorithm in Statistical Shape Analysis. *Commun. Stat. Simul.*
262 *Comput.* **2010**, *39*, 1016–1026. [[CrossRef](#)]
- 263 7. Lele, S.; Richtsmeier, J. *An Invariant Approach to Statistical Analysis of Shapes*; Chapman & Hall/CRC: New York,
264 NY, USA, 2001.
- 265 8. Huang, C.; Martin, S.; Zhu, H. Clustering High-Dimensional Landmark-based Twodimensional Shape Data. *J.*
266 *Am. Stat. Assoc.* **2015**, *110*, 946–961. [[CrossRef](#)] [[PubMed](#)]
- 267 9. Kume, A.; Welling, M. Maximum likelihood estimation for the offset-normal shape distributions using EM. *J.*
268 *Comput. Gr. Stat.* **2010**, *19*, 702–723. [[CrossRef](#)]
- 269 10. Gattone, S.A.; De Sanctis, A.; Russo, T.; Pulcini, D. A shape distance based on the Fisher–Rao metric and its
270 application for shapes clustering. *Phys. A Stat. Mech. Appl.* **2017**, *487*, 93–102. [[CrossRef](#)]
- 271 11. De Sanctis, A.; Gattone, S.A. A Comparison between Wasserstein Distance and a Distance Induced by Fisher-Rao
272 Metric in Complex Shapes Clustering. *Proceedings* **2018**, *2*, 163. [[CrossRef](#)]
- 273 12. Amari, S.; Nagaoka, H. Translations of Mathematical Monographs. In *Methods of Information Geometry*; AMS &
274 Oxford University Press: Providence, RI, USA, 2000.
- 275 13. Murray, M.K.; Rice, J.W. *Differential Geometry and Statistics*; Chapman & Hall: London, UK, 1984.
- 276 14. Srivastava, A.; Joshi, S.H.; Mio, W.; Liu, X. Statistical Shape analysis: Clustering, learning, and testing. *IEEE Trans.*
277 *PAMI* **2005**, *27*, 590–602. [[CrossRef](#)] [[PubMed](#)]
- 278 15. Mio, W.; Srivastava, A.; Joshi, S.H. On Shape of Plane Elastic Curves. *Int. J. Comput. Vis.* **2007**, *73*, 307–324.
279 [[CrossRef](#)]
- 280 16. Skovgaard, L.T. A Riemannian geometry of the multivariate normal model. *Scand. J. Stat.* **1984**, *11*, 211–223.
- 281 17. Costa, S.; Santos, S.; Strapasson, J. Fisher information distance: A geometrical reading. *Discret. Appl. Math.* **2015**,
282 *197*, 59–69. [[CrossRef](#)]
- 283 18. Pennec, X.; Fillard, P.; Ayache, N. A Riemannian Framework for Tensor Computing. *Int. J. Comput. Vis.* **2006**,
284 *66*, 41–66. [[CrossRef](#)]
- 285 19. Bishop, R.L.; O’Neill, B. Manifolds of negative curvature. *Trans. Am. Math. Soc.* **1969**, *145*, 1–49. [[CrossRef](#)]
- 286 20. Takatsu, A. Wasserstein geometry of Gaussian measures. *Osaka J. Math.* **2011**, *48*, 1005–1026.
- 287 21. Amari, S. Applied Mathematical Sciences. In *Information Geometry and Its Applications*; Springer: Berlin, Germany,
288 2016.
- 289 22. Banerjee, A.; Merugu, S.; Dhillon, I.; Ghosh, J. Clustering with Bregman Divergence. *J. Mach. Learn. Res.* **2005**,
290 *6*, 1705–1749.
- 291 23. Bookstein, F.L. Size and shape spaces for landmark data in two dimensions. *Stat. Sci.* **1986**, *1*, 181–242. [[CrossRef](#)]
- 292 24. Goodall, C.R. Procrustes methods in the statistical analysis of shape. *J. R. Stat. Soc.* **1991**, *53*, 285–339.

- 293 25. Hubert, L.; Arabie, P. Comparing Partitions. *J. Classif.* **1985**, *2*, 193–218. [[CrossRef](#)]

© 2018 by the author. Submitted to *Entropy* for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).