AIMS Mathematics

http://www.aimspress.com/journal/Math

AIMS Mathematics, 3(4): 439–447
DOI:10.3934/Math.2018.4.439
Received: 25 July 2018
Accepted: 28 September 2018
Published: 19 October 2018

Research article

Information distance estimation between mixtures of multivariate Gaussians

C. T. J. Dodson*

School of Mathematics, University of Manchester, Manchester, M13 9PL, UK

* **Correspondence:** Email: ctdodson@manchester.ac.uk.

Abstract: There are efficient software programs for extracting from large data sets and image sequences certain mixtures of probability distributions, such as multivariate Gaussians, to represent the important features and their mutual correlations needed for accurate document retrieval from databases. This note describes a method to use information geometric methods for distance measures between distributions in mixtures of arbitrary multivariate Gaussians. There is no general analytic solution for the information geodesic distance between two k-variate Gaussians, but for many purposes the absolute information distance may not be essential and comparative values suffice for proximity testing and document retrieval. Also, for two *mixtures* of different multivariate Gaussians we must resort to approximations to incorporate the weightings. In practice, the relation between a reasonable approximation and a true geodesic distance is likely to be monotonic, which is adequate for many applications. Here we consider some choices for the incorporation of weightings in distance estimation and provide illustrative results from simulations of differently weighted mixtures of multivariate Gaussians.

Keywords: information geometry; multivariate spatial covariance; Gaussian mixtures; geodesic distance; approximations
Mathematics Subject Classification: 60D05, 53B20

1. Introduction

Cao et al. [4] reviewed techniques for extracting local features for automatic object recognition in images. Multivariate Gaussians can represent the important features and their mutual correlations needed for accurate document retrieval from databases. The natural choice for discrimination between pairs of such distributions is the Fisher information metric on the Riemannian manifold of smooth probability density functions coordinatized by the parameters of the distribution [1, 2]. However, it is not known analytically in some important cases of practical interest.

We have used multivariate Gaussians for face recognition using the neighbourhoods of colour pixel features at landmark points in face images [12], where we found that the spatial covariances among pixel colours was important. Craciunesco and Murari et al [5, 8] used geodesic distance on Gaussian manifolds to interpret time series in very large databases from Tokomak measurements in fusion research. Verdoolaege, Shabbir et al [10, 13] used multivariate generalized Gaussians for colour texture discrimination in the wavelet domain. In these studies the discrimination used approximations to the information distance between pairs of multivariate Gaussian probability density functions. Nielsen et al [9] suggested an entropic quantization method for approximating distances in the case of mixtures of multivariate Gaussians.

The k-variate Gaussian distributions have probability density functions:

$$f(\mu, \Sigma) = \frac{e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}}{\sqrt{(2\pi)^k |\Sigma|}}, \tag{1}$$

where $x \in \mathbb{R}^k$ is the random variable, $\mu \in \mathbb{R}^k$ a k-dimensional mean vector, and $\Sigma \in \mathbb{R}^{(k^2+k)/2}$ is the $k \times k$ positive definite symmetric covariance matrix, for features with k-dimensional representation.

The Riemannian manifold M^k of the family of k-variate Gaussians for a given k is well understood through information geometric study using the Fisher metric [1, 3, 6, 11]. For an introduction to information geometry and a range of applications see [1, 2, 7]. What we have analytically are natural metrics, on the space of means and on the space of covariances, giving the information distance between two multivariate Gaussians $f^A(\mu^A, \Sigma), f^B(\mu^B, \Sigma)$ of the *same* number k of variables in two particular cases:

1. $\mathbf{\Sigma^A = \Sigma^B = \Sigma}$: $\quad f^A(\mu^A, \Sigma), f^B(\mu^B, \Sigma)$

Here we have the positive definite symmetric quadratic form Σ to give a distance between two mean vectors:

$$D_\mu(f^A, f^B) = \sqrt{(\mu^A - \mu^B)^T \cdot \Sigma^{-1} \cdot (\mu^A - \mu^B)}. \tag{2}$$

So also we have a norm on mean vectors for each $f^A(\mu^A, \Sigma)$:

$$||\mu^A|| = \sqrt{(\mu^A)^T \cdot (\Sigma^A)^{-1} \cdot (\mu^A)} \tag{3}$$

which is evidently sensitive to the covariance.

2. $\boldsymbol{\mu^A = \mu^B = \mu}$: $\quad f^A(\mu, \Sigma^A), f^B(\mu, \Sigma^B)$

Here we use a positive definite symmetric matrix constructed from Σ^A and Σ^B to give distance between two covariance matrices; this information metric was given by Atkinson and Mitchell [3] from a result attributed to S.T. Jensen, using

$$S^{AB} = \Sigma^{A^{-1/2}} \cdot \Sigma^B \cdot \Sigma^{A^{-1/2}}, \text{ with } \{\lambda_j^{AB}\} = \mathrm{Eig}(S^{AB}) \text{ then}$$

$$D_\Sigma(f^A, f^B) \quad = \quad \sqrt{\frac{1}{2} \sum_{j=1}^{k} \log^2(\lambda_j^{AB})}. \tag{4}$$

We note that (4) is in agreement also with a special case of the geodesic distance given by Shabbir et al [10] for generalized multivariate Gaussians with the same mean.

In principle, (4) yields all of the geodesic distances since the information metric is invariant under affine transformations of the mean [3] Appendix 1; see also the article of P.S. Eriksen [6]. The equations for the geodesics were shown by Skovgaard [11] to be

$$\begin{aligned} \ddot{\mu} &= \dot{\Sigma}\Sigma^{-1}\dot{\mu} \\ \ddot{\Sigma} &= \dot{\Sigma}\Sigma^{-1}\dot{\Sigma} - \dot{\mu}\dot{\mu}^T. \end{aligned} \tag{5}$$

Eriksen [6] observed that the family M^k of k-variate Gaussians is isometric to the space $GA^+(k)/SO(k)$ where $GA^+(k)$ consists of positive affine transformations. Hence by a translation it is sufficient to restrict the geodesic to one through $\Sigma = I$ the identity, in the direction $(-B, v)$. Then, through the change of coordinates, $\Delta = \Sigma^{-1}$, $\delta = \Sigma^{-1}\mu$, Equation (5) becomes

$$\begin{aligned} \dot{\Delta} &= -B\Delta + v\delta^T \ \text{ with } \ \Delta(0) = I, \delta(0) = 0, \\ \dot{\delta} &= -B\delta + (1 + \delta^T\Delta^{-1}\delta)v. \end{aligned} \tag{6}$$

Then using

$$A = \begin{pmatrix} -B & v & 0 \\ v^T & 0 & -v^T \\ 0 & -v & B \end{pmatrix} \tag{7}$$

Eriksen proved that the geodesic solution curve is given by

$$\Lambda : \mathbb{R} :\to M^k : t \mapsto e^{At} = \begin{pmatrix} \Delta & \delta & \Phi \\ \delta^T & 1 + \delta^T\Delta^{-1}\delta & \gamma^T \\ \Phi^T & \gamma & \Gamma \end{pmatrix} \tag{8}$$

$$\text{where } \gamma = \Delta^{-1}\delta + \Phi^T\Delta^{-1}\delta, \text{ and } \delta^T\Delta^{-1}\delta = \gamma^T\Gamma^{-1}\gamma. \tag{9}$$

$$\text{So } (\Delta(-t), \delta(-t)) = (\Gamma(t), \delta(t)). \tag{10}$$

Of course, the analytic difficulty is the requirement to find the length of the geodesic between two points in M^k to obtain a distance function, that being the infimum of arc length over all curves joining the points.

2. Approximating distances between arbitrary mixtures of multivariate Gaussians

Here we consider a mixture distribution consisting of a linear combination of k-variate Gaussians $\{f_k, k = 2, 3, \ldots, N\}$ with an increasing sequence of $k = 2, 3, \ldots, N$ variables and with probability density functions:

$$f_2(\mu_2, \Sigma_2), f_3(\mu_3, \Sigma_3) ..., f_N(\mu_N, \Sigma_N) \text{ and } \forall k \int_{\mathbb{R}^k} f_k = 1 \tag{11}$$

where $\mu_k \in \mathbb{R}^k$ is the k-vector of means and $\Sigma_k \in \mathbb{R}^{(k^2+k)/2}$ is the positive definite symmetric $(k \times k)$ covariance matrix with components $(\sigma_{ij}), i \leq j = 1, 2, ..., k$. The standard basis for the space of covariance matrices is $E_{ij} = 1_{ii}$ for $i = j$, $E_{ij} = 1_{ij} + 1_{ji}$ for $i \neq j$ so

$$\Sigma = \sum_{i \leq j = 1}^{k} \sigma_{ij} E_{ij}.$$

We presume that the parameters and relative weights w_k of these component probability density functions (11) have been obtained empirically, giving a mixture density:

$$f^A = \sum_{k=2}^{N} w_k^A f_k^A, \quad \text{with } w_k^A \geq 0 \text{ and } \sum_{k=2}^{N} w_k^A = 1. \tag{12}$$

Given two such distributions, $f^A = f^A(\mu^A, \Sigma^A, w^A)$ and $f^B = f^B(\mu^B, \Sigma^B, w^B)$, we wish to be able to estimate the information distance $D(f^A, f^B)$ between them.

There is no general analytic solution for the geodesic distance between two k-variate Gaussians, but for many purposes the absolute information distance is not essential and comparative values may suffice for proximity testing, then the sum $D = D_\mu + D_\Sigma$ from (2) and (4) is a natural approximation. Indeed, (4) gives the geodesic distance between f^A with $\Sigma^A = I$ and f^B with $\mu^A = \mu^B = 0$ and the information metric is invariant under affine transformations of the mean [3, 6].

So, a fortiori, also we do not have the distance between two mixtures of multivariate Gaussians: $f^A(\mu^A, \Sigma^A, w^A)$ and $f^B(\mu^B, \Sigma^B, w^B)$. For this we must resort to approximations for incorporating the weightings of component Gaussians. In practice, it may not matter greatly since the relation between a reasonable approximation and a true geodesic distance is likely to be monotonic, which may be adequate for many applications, and was what we found in our face recognition work [12].

2.1. Averaging distances over weightings

Perhaps the most natural method is to combine equations (2) and (4) through the linear combination (12), obtaining an approximation as a corresponding linear combination of distances. Given two mixture distributions $f^A(\mu^A, \Sigma^A), f^B(\mu^B, \Sigma^B)$ we could split the distance estimate function $D^\#$ into $D^\#_\mu$ and $D^\#_\Sigma$ as follows with $\delta\mu = (\mu^A - \mu^B)$:

$$D^\#_\mu(f^A, f^B) = \sum_{k=2}^{N} \frac{1}{2}\left(w_k^A \sqrt{\delta\mu^T \cdot \Sigma_k^{A^{-1}} \cdot \delta\mu} + w_k^B \sqrt{\delta\mu^T \cdot \Sigma_k^{B-1} \cdot \delta\mu} \right) \tag{13}$$

$$D^\#_\Sigma(f^A, f^B) = \sum_{k=2}^{N} \frac{1}{2}(w_k^A \sqrt{\frac{1}{2}\sum_{k=2}^{N}(\log \lambda_k^{AB})^2} + w_k^B \sqrt{\frac{1}{2}\sum_{k=2}^{N}(\log \lambda_k^{BA})^2}) \tag{14}$$

$$\{\lambda_k^{AB}\} = \mathrm{Eig} H_k^{AB}, \quad H_k^{AB} = \left((\Sigma_k^A)^{-1/2} \cdot \Sigma_k^B \cdot (\Sigma_k^A)^{-1/2} \right).$$

We note $D^\#_\mu(f^A, f^B)$ in (13) does give a potentially useful distance measure between the two Gaussian mixtures, since it incorporates both means and covariances. Figure 1 shows the effect on $D^\#_\mu$ values between differing averaged weighting sequences for random k-variate Gaussians having $k = 2, 3, 4, 5$ variables, with increasing weights f^A, uniform weights f^B, and decreasing weights f^C. The g^A, g^B, g^C are for the same mixtures except that Σ_2^C has been replaced by $\Sigma_2^C/5$ and h^A, h^B, h^C are for the same mixtures except that Σ_5^C has been replaced by $\Sigma_5^C/5$ to show the effect of a change in one covariance component. For these simulations *Mathematica* was used to generate random mean vectors $\mu_i \in [5, 10]$ and random covariance matrix elements $\sigma_{ij} \in [5, 10], i \neq j, \sigma_{ii} \in [-6, 6]$, choices that ensured positive definite symmetric matrices for covariances. This approach was chosen to try to illustrate the effects of weighting sequences and isolated covariance changes on the measurements of distances between mixtures. Table 1 shows mean values $\overline{D^\#_\mu}$ and $\overline{D^\#_\Sigma}$ for the pairs of mixtures $(f^A, f^B), (f^B, f^C), (f^A, f^C)$

for the 20 random Gaussians with weights $w_k^A = (0.1, 0.2, 0.3, 0.4)$, $w_k^B = (0.25, 0.25, 0.25, 0.25)$, $w_k^C = (0.4, 0.3, 0.2, 0.1)$.

Table 1. Mean values $\overline{D_\mu^\#}$ and $\overline{D_\Sigma^\#}$ for the pairs of mixtures $(f^A, f^B), (f^B, f^C), (f^A, f^C)$ for the 20 random Gaussians with weights $w_k^A = (0.1, 0.2, 0.3, 0.4)$, $w_k^B = (0.25, 0.25, 0.25, 0.25)$, $w_k^C = (0.4, 0.3, 0.2, 0.1)$.

	$\overline{D_\mu^\#}(A, B)$	$\overline{D_\mu^\#}(B, C)$	$\overline{D_\mu^\#}(A, C)$
Mixture f	0.6085	0.5635	0.5868
Mixture g	0.7087	0.7694	0.7522
Mixture h	0.9002	0.7126	0.7774

	$\overline{D_\Sigma^\#}(A, B)$	$\overline{D_\Sigma^\#}(B, C)$	$\overline{D_\Sigma^\#}(A, C)$
Mixture f	0.8607	0.8110	0.8537
Mixture g	0.8607	0.8110	0.8537
Mixture h	0.8607	0.8110	0.8537

As expected from the form of $D_\Sigma^\#$ in equation (14), its values in the table of means is unaffected by the scale changes in covariances through mixtures f, g, h, but is sensitive to weighting sequences.

However, the k-variate components from two mixtures might not come from the same feature space in some applications so there may be no connection between the contributing features they are representing. On the other hand, a commonly used feature space is that of pixel colours in different locations, as for example in texture and face recognition and those feature spaces are the same.

2.2. *Mixtures projected onto the complex plane*

The new implementation described here uses the information geometric norm on the mean vectors and the Frobenius norm on the covariance matrices to project the mixture distributions onto the complex plane. This 2-dimensional representation reveals influences of the means and covariances in the mixtures, which itself may be a valuable It allows also the direct calculation of a distance between two mixture distributions using moduli, without assuming any connections between the mixtures, though this has the effect of smoothing the component influences of the means and covariances.

The idea here is simple: for each mixture distribution f^A given by a weighted sum (12) we obtain two numbers $||\mu^A||$ and $||\Sigma^A||$ being the weighted sums of norms of means and covariances. The norm on mean vectors is given by (3) and for the covariance matrices we need a matrix norm, which here we choose as the Frobenius norm given for an $n \times n$ matrix $M_{\alpha\beta}$ by the square root of the sum of squares of its elements $m_{\alpha\beta}$,

$$||M_{\alpha\beta}||^2 = \sum_{\alpha=1}^{n} \sum_{\beta=1}^{n} (m_{\alpha\beta})^2$$

Note that if $M_{\alpha\beta}$ has eigenvalues $\{\lambda_\alpha\}$ and is represented on a basis of eigenvectors then

$$\|M_{\alpha\beta}\|^2 = \sum_{\alpha=1}^{n} (\lambda_\alpha)^2.$$

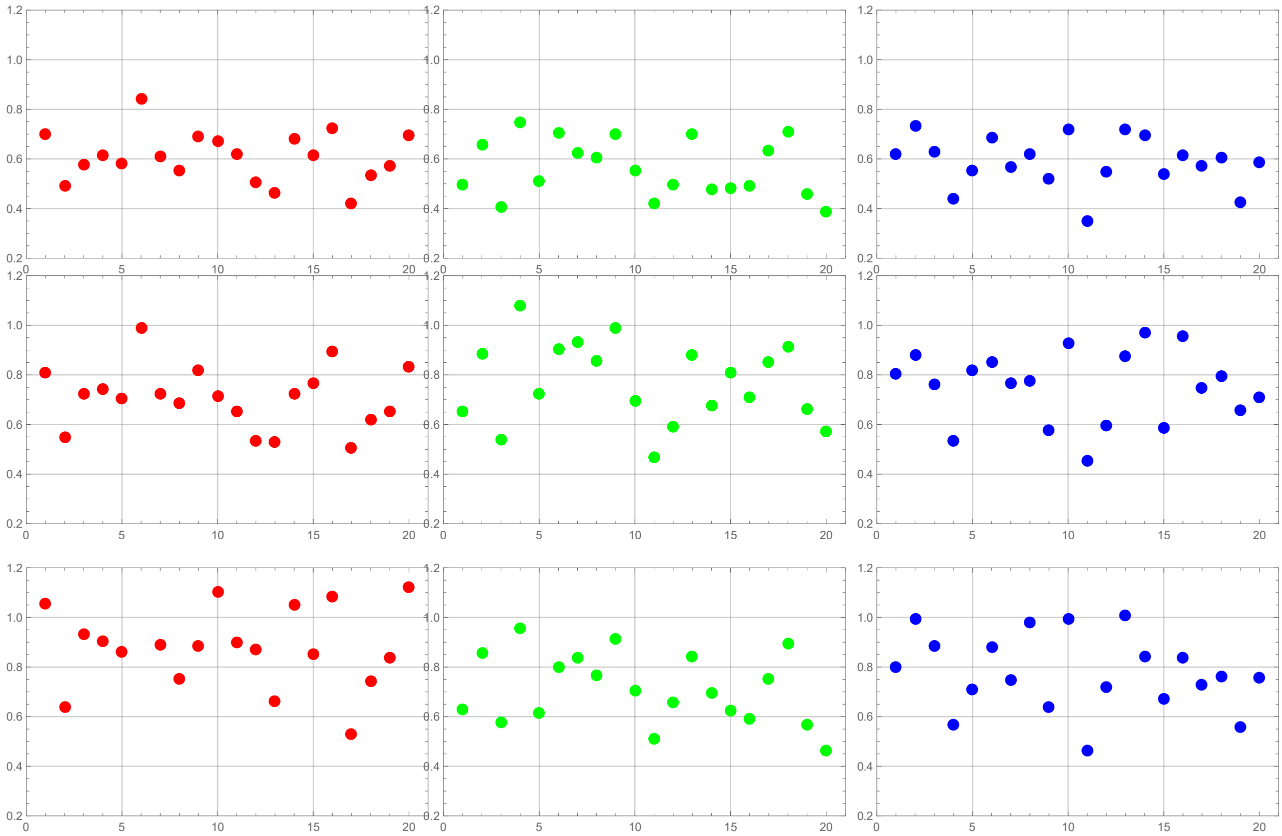

Figure 1. *Illustration of the effects of weighting sequences and covariances in mixtures. Upper row:* $D^{\#}_{\mu}(f^A, f^B), D^{\#}_{\mu}(f^B, f^C), D^{\#}_{\mu}(f^A, f^C)$, *Middle row:* $D^{\#}_{\mu}(g^A, g^B), D^{\#}_{\mu}(g^B, g^C), D^{\#}_{\mu}(g^A, g^C)$, *Lowest row:* $D^{\#}_{\mu}(h^A, h^B), D^{\#}_{\mu}(h^B, h^C), D^{\#}_{\mu}(h^A, h^C)$ *for 20 random k-variate Gaussians for* $k = 2, 3, 4, 5$ *with weights* $w^A_k = (0.1, 0.2, 0.3, 0.4)$, $w^B_k = (0.25, 0.25, 0.25, 0.25)$, $w^C_k = (0.4, 0.3, 0.2, 0.1)$.

Given a mixture distribution f^A consisting of M different multivariate Gaussians: $G^A = \{G^A_i(\mu^A_i, \Sigma^A_i)\}_{i=1,M}$ with weights $w^A = \{w^A_i\}_{i=1,M}$ we have

$$f^A = \sum_{m=1}^{M} w^A_m G^A_m$$

$$\|\mu^A\| = \sqrt{\sum_{m=1}^{M} w^A_m (\mu^A_m)^T.(\Sigma^A_m)^{-1}.(\mu^A_m)} \tag{15}$$

$$\|\Sigma^A\| = \sqrt{\sum_{m=1}^{M} w^A_m \|\Sigma^A_m\|^2}. \tag{16}$$

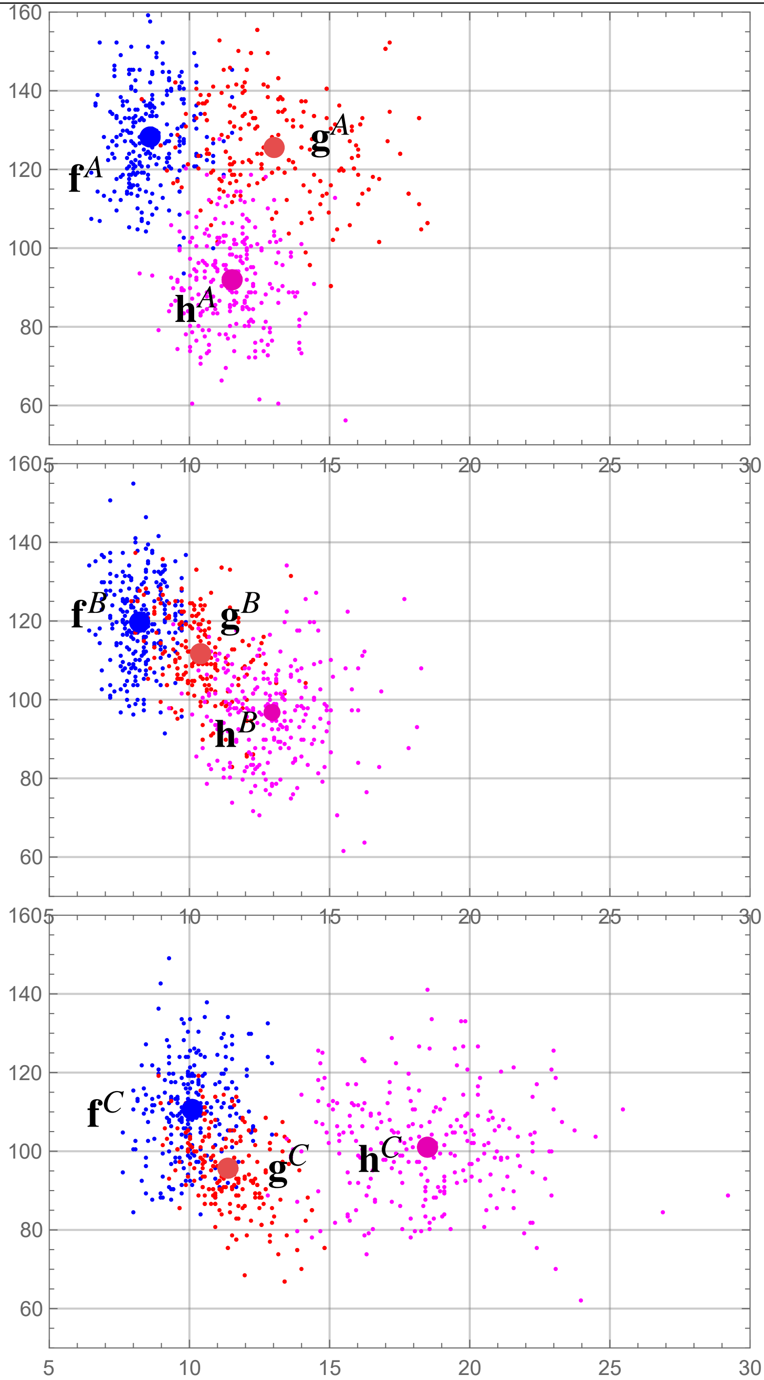

Figure 2. Mixture projection onto $\mathbb{C}$. *Mixtures (f^A, f^B, f^C) for 250 random Gaussians with weights $w_k^A = (0.1, 0.2, 0.3, 0.4)$, $w_k^B = (0.25, 0.25, 0.25, 0.25)$, $w_k^C = (0.4, 0.3, 0.2, 0.1)$ are shown plotted in ($\|\mu\|, \|\Sigma\|$)-space for k-variate Gaussians having $k = 2, 3, 4, 5$ variables, with increasing weights f^A, uniform weights f^B, and decreasing weights f^C. The g^A, g^B, g^C are for the same mixtures except that Σ_2^C has been replaced by $\Sigma_2^C/5$ and h^A, h^B, h^C are for the same mixtures except that Σ_5^C has been replaced by $\Sigma_5^C/5$ to show the effect of a change in one covariance component. The mean for each over the 250 replications is shown as a larger point.*

Now we can represent f^A by the complex number $\phi^A = ||\mu^A|| + i||\Sigma^A||$ and its difference from another such complex number ϕ^B for f^B gives us a distance measure in our reduced space of mixtures:

$$\Delta(f^A, f^B) = |\phi^B - \phi^A|. \tag{17}$$

Figure 2 shows a plot of the points $(||\mu||, ||\Sigma||) \in \mathbb{C}$ for the 250 mixtures of random k-variate Gaussians having $k = 2, 3, 4, 5$ variables, with increasing weights f^A, uniform weights f^B, and decreasing weights f^C. The g^A, g^B, g^C are for the same mixtures except that Σ_2^C has been replaced by $\Sigma_2^C/5$ and h^A, h^B, h^C are for the same mixtures except that Σ_5^C has been replaced by $\Sigma_5^C/5$ to show the effect of a change in one covariance component. In each case the mean for each over the 250 replications is shown as a large point. This was done using *Mathematica* with random mean vectors $\mu_i \in [5, 10]$ and random covariance matrix elements $\sigma_{ij} \in [5, 10], i \neq j, \sigma_{ii} \in [-6, 6]$.

3. Discussion

There are efficient software programs for extracting from large data sets and image sequences certain mixtures of probability distributions, such as multivariate Gaussians, to represent the important features and their mutual correlations needed for accurate document retrieval from databases. The lack of an analytic solution to the geodesic distance equations between points in the Riemannian space of multivariate Gaussian mixtures, Equation (12), with an information metric, means that approximate solutions need to be found for practical applications. We have illustrated a new approximation for the case of 250 mixtures of k-variate Gaussians for $k = 2, 3, 4, 5$ with four weightings $w_k^A = (0.1, 0.2, 0.3, 0.4)$, $w_k^B = (0.25, 0.25, 0.25, 0.25)$, $w_k^C = (0.4, 0.3, 0.2, 0.1)$ of the component Gaussians that are increasing, uniform and decreasing. These simulations show the effects of covariance changes and the effects of weighting sequences on each given collection of k-variate Gaussians for $k = 2, 3, 4, 5$.

4. Conclusions

Whereas there are not analytic expressions for the information geometric distance between pairs of mixtures of multivariate Gaussians, we have shown that there are several choices for good information geometric approximate distances which are easy to compute. The new method yielded evident discrimination between pairs of these mixtures, shown in easily interpretable graphical form, Figure 2, distinguishing effects of covariance changes and effects of weighting sequences.

Acknowledgement

The methods described in §2.1 were developed with J. Scharcanski and J. Soldera during a visit to UFRGS, Brazil with a grant from The London Mathematical Society in 2013 and the author is grateful to CAPES (Coordeanao de Aperfeioamento de Pessoal de Nivel Superior, Brazil) for partially funding this project. An application of the results to face recognition was reported elsewhere in a joint paper [12].

Conflict of interest

The author declares that there are no conflicts of interest in this paper.

References

1. S-I. Amari and H. Nagaoka, *Methods of Information Geometry,* Oxford, American Mathematical Society, Oxford University Press, 2000.
2. K. Arwini and C. T. J. Dodson, *Information Geometry Near Randomness and Near Independence,* Lecture Notes in Mathematics, Springer-Verlag, New York, Berlin, 2008.
3. C. Atkinson and A. F. S. Mitchell, *Rao's distance measure*, Sankhya: Indian Journal of Statistics, **43** (1981), 345–365.
4. J. Cao, D. Mao, Q. Cai, et al. *A review of object representation based on local features*, Journal of Zhejiang University-SCIENCE C (Computers & Electronics), **14** (2013), 495–504.
5. T. Craciunescu and A. Murari, *Geodesic distance on Gaussian manifolds for the robust identification of chaotic systems*, Nonlinear Dynam, **86** (2016), 677–693.
6. P. S. Eriksen, Geodesics connected with the Fisher metric on the multivariate normal manifold. In C.T.J. Dodson, Editor, *Proceedings of the Geometrization of Statistical Theory Workshop*, Lancaster (1987), 225–229.
7. F. Nielsen and F. Barbaresco , *Geometric Science of Information*, LNCS 8085, Springer, Heidelberg, 2013.
8. A. Murari, T. Craciunescu, E. Peluso, et al. *Detection of Causal Relations in Time Series Affected by Noise in Tokamaks Using Geodesic Distance on Gaussian Manifolds*, Entropy **19** (2017), 569.
9. F. Nielsen, V. Garcia and R. Nock, Simplifying Gaussian mixture models via entropic quantization. In Proc. 17^{th} *European Signal Processing Conference*, Glasgow, Scotland 24-28 August 2009, 2012–2016.
10. A. Shabbir, G. Verdoolaege and G. Van Oost. Multivariate texture discrimination based on geodesics to class centroids on a generalized Gaussian manifold. In F. Nielsen and F. Barbaresco (Eds) *Geometric Science of Information*, LNCS 8085, Springer, Berlin-Heidelberg, 2013, 853–860.
11. L. T. Skovgaard, *A Riemannian geometry of the multivariate normal model*, Scand. J. Stat., **11** (1984), 211–223.
12. J. Soldera, C. T. J. Dodson and J. Scharcanski, *Face recognition based on texture information and geodesic distance approximations between multivariate normal distributions*, Measurement Science and Technology, 2018.
13. G. Verdoolaege and A. Shabbir. *Color Texture Discrimination Using the Principal Geodesic Distance on a Multivariate Generalized Gaussian Manifold*, International Conference on Networked Geometric Science of Information, Springer, Cham Berlin-Heidelberg, 2015, 379–386.