MANCHESTER
1824

The University of Manchester

Information distance estimation between mixtures of multivariate Gaussians

Dodson, CTJ

2015

MIMS EPrint: **2015.108**

Manchester Institute for Mathematical Sciences
School of Mathematics

The University of Manchester

Reports available from: http://eprints.maths.manchester.ac.uk/
And by contacting: The MIMS Secretary
School of Mathematics
The University of Manchester
Manchester, M13 9PL, UK

ISSN 1749-9097

Information distance estimation between mixtures of multivariate Gaussians

C.T.J. Dodson
School of Mathematics
University of Manchester
Manchester, M13 9PL, UK.

September 20, 2015

Abstract

There are efficient software programs for extracting from image sequences certain mixtures of distributions, such as multivariate Gaussians, to represent the important features needed for accurate document retrieval from databases. This note describes a method to use information geometric methods to measure distances between distributions in mixtures of 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 is not essential and comparative values suffice for proximity testing. For two mixtures of 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 compare several 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. MSC 60D05 53B20

1 Introduction

A recent review of techniques for extracting local features for automatic object recognition in images has been given by Cao et al [4]; implicit in such techniques is computer vision and the elicitation of features that are invariant under image transformation for object classification. In a number of important areas of application the representation of local features—think of smiley, neutral or sad faces in video sequences—can be achieved through mixtures of multivariate Gaussian distributions. The Riemannian manifold of the family of k-variate Gaussians for a given k is well understood through information geometric study using the Fisher metric. For an introduction to information geometry and a range of applications see [1].

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

$$f_2 = (2, \mu_2, \Sigma_2), f_3 = (3, \mu_3, \Sigma_3)..., f_N = (N, \mu_N, \Sigma_N) \text{ and } \forall k \int_{\mathbb{R}^k} f_k = 1 \tag{1}$$

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 (1) have been obtained empirically, giving a mixture density:

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

We wish to be able to estimate the information distance $D(f^A, f^B)$ between two such distributions, $f^A = (\mu^A, \Sigma^A, w^A)$ and $f^B = (\mu^B, \Sigma^B, w^B)$. What we have analytically are natural norms, on the space of means and on the space of covariances, giving the information distance between two multivariate Gaussians of the *same* number k of variables in two particular cases:

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

Here we have the positive definite symmetric quadratic form Σ to give a norm on the difference vector of means:

$$D_\mu(f^A, f^B) = \sqrt{(\mu^A - \mu^B)^T \cdot \Sigma^{-1} \cdot (\mu^A - \mu^B)}. \tag{3}$$

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

Here we need a positive definite symmetric matrix constructed from Σ^A and Σ^B to give a norm on the space of differences between covariances; the information metric is given by Atkinson and Mitchell [2] from a result attributed to S.T. Jensen, using

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

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

In principle, (4) yields all of the geodesic distances since the information metric is invariant under affine transformations of the mean [2] Appendix 1; see also the article of P.S. Eriksen [3].

Also, we know analytically the Kullback-Leibler divergence, or relative entropy, between two multivariate Gaussians $f^A = (k, \mu^A, \Sigma^A), f^B = (k, \mu^B, \Sigma^B)$ with the same number k of variables, its square root giving a separation measurement [5]:

$$\begin{aligned} KL(f^A, f^B) &= \frac{1}{2}\log(\frac{\det \Sigma^B}{\det \Sigma^A}) + \frac{1}{2}\text{Tr}[{\Sigma^B}^{-1} \cdot \Sigma^A] \\ &+ \frac{1}{2}\left(\mu^A - \mu^B\right)^T \cdot {\Sigma^B}^{-1} \cdot \left(\mu^A - \mu^B\right) - \frac{k}{2}. \end{aligned} \tag{5}$$

This is not symmetric, so to obtain a distance we could take the average KL-distance in both directions:

$$D_{KL}(f^A, f^B) = \sqrt{\frac{|KL(f^A, f^B)| + |KL(f^B, f^A)|}{2}} \tag{6}$$

The Kullback-Leibler distance tends to the information distance as two distributions become closer together; conversely it becomes less accurate as they move apart. Explicitly, we have for the covariance part $DKL_\Sigma(f^A, f^B)$

$$\begin{aligned} DKL_\Sigma(f^A, f^B) &= \frac{1}{2}\left(\sqrt{\left|\frac{1}{2}\log\left(\frac{\det \Sigma^B}{\det \Sigma^A}\right) + \frac{1}{2}\text{Tr}\left[\Sigma^{-B}.\Sigma^A\right] - \frac{k}{2}\right|}\right. \\ &\left. + \sqrt{\left|\frac{1}{2}\log\left(\frac{\det \Sigma^A}{\det \Sigma^B}\right) + \frac{1}{2}\text{Tr}\left[\Sigma^{-A}.\Sigma^B\right] - \frac{k}{2}\right|}\right). \end{aligned} \tag{7}$$

The true geodesic distance is plotted against $DKL_\Sigma(f^A, f^B)$ in Figure 1 for 600 bivariate Gaussian covariance matrices.

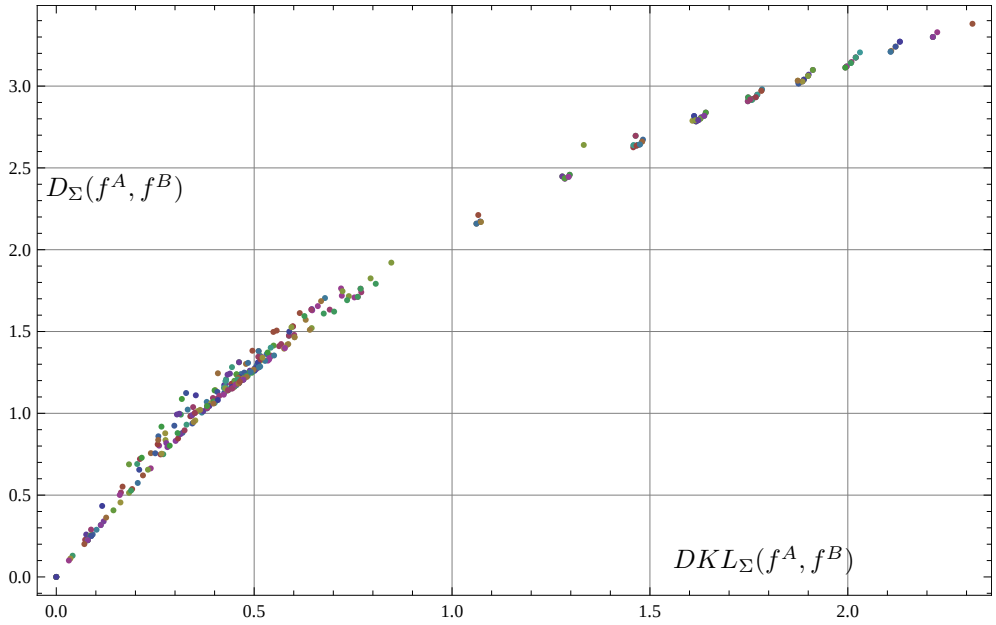

Figure 1: *Plot of $D_\Sigma(f^A, f^B)$ from (4) on $DKL_\Sigma(f^A, f^B)$ from (7) for 600 bivariate Gaussian covariance matrices.*

1.1 Example: Bivariate Gaussians

$$
\begin{aligned}
f(x,y) &= \frac{1}{2\pi\sqrt{\Delta}} exp \frac{-1}{\Delta^2}(y-\mu_2)^2\sigma_{11} + (x-\mu_1)[(x-\mu_1)\sigma_{22} + 2(-y+\mu_2))\sigma_{12}] \qquad (8)\\
\mu &= (\mu_1, \mu_2), \quad \Delta = Det[\Sigma] = \sigma_{11}\sigma_{22} - \sigma_{12}^2\\
\Sigma &= \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} = \sigma_{11}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \sigma_{12}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} + \sigma_{22}\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}\\
\Sigma^{-1} &= \begin{pmatrix} \frac{\sigma_{22}}{\Delta} & -\frac{\sigma_{12}}{\Delta} \\ -\frac{\sigma_{12}}{\Delta} & \frac{\sigma_{11}}{\Delta} \end{pmatrix}
\end{aligned}
$$

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

$$\sqrt{\frac{\left(\mu_2^A - \mu_2^B\right)\left(\sigma_{11}\left(\mu_2^A - \mu_2^B\right) - \sigma_{12}\left(\mu_1^A - \mu_1^B\right)\right)}{\Delta} + \frac{\left(\mu_1^A - \mu_1^B\right)\left(\sigma_{22}\left(\mu_1^A - \mu_1^B\right) - \sigma_{12}\left(\mu_2^A - \mu_2^B\right)\right)}{\Delta}}.$$

The analytic expression for distance between two covariance matrices is cumbersome so we show a numerical example:

$$\Sigma^A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \Sigma^B = \begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix}, \quad S^{AB} = \begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix}, S^{BA} = \begin{pmatrix} 0.42857 & -0.14286 \\ -0.14286 & 0.21429 \end{pmatrix}$$

with eigenvalues : $\lambda_1^{AB} = 7, \lambda_2^{AB} = 2$ and $\lambda_1^{BA} = 0.5, \lambda_2^{BA} = 0.14286$, respectively, then

$$
\begin{aligned}
D_\Sigma(\Sigma^A, \Sigma^B) &= \sqrt{\frac{1}{2}\sum_{j=1}^{n} \log^2(\lambda_j)} = 1.46065.\\
D_{KL}(\Sigma^A, \Sigma^B) &= \frac{1}{2}\left(\sqrt{\frac{1}{2}\log 14 - \frac{19}{28}} + \sqrt{\frac{1}{2}\log\frac{1}{14} + \frac{7}{2}}\right) \approx 1.1386.
\end{aligned}
$$

2 Approximating distances between arbitrary mixtures

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 suffice for proximity testing, then the sum $D = D_\mu + D_\Sigma$ from (3) and (4) is a sufficient 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 [2, 3].

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.

One method is to combine equations (3) and (4) through the linear combination (2), obtaining an approximation as a corresponding linear combination of distances. To achieve this there are several choices of how to combine weighted sets of D_μ and D_Σ and here we mention two. The natural choice §2.1 incorporates the Gaussian component weights w_k inside the matrix operations; a simpler choice §2.2 just takes the average weighted values. Figure 2 and, Figure 3 illustrate their results on different sequences of weight vectors. However, both of those approaches suffer from the disadvantage of assuming that the k-variate components from two mixtures come from the same space but in fact there may be no connection between the contributing features they are representing.

The new implementation described in §2.3 uses the incorporated weights, 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 allows the direct calculation of a distance between two mixture distributions using moduli, without assuming any connections between the mixtures.

2.1 Incorporated weights

Given two mixture distributions $f^A = (\mu^A, \Sigma^A), f^B = (\mu^B, \Sigma^B)$ we split the distance estimate function D^* into D^*_μ and D^*_Σ as follows:

$$D^*(f^A, f^B) = D^*_\mu(f^A, f^B) + D^*_\Sigma(f^A, f^B), \quad \text{where} \tag{9}$$

$$D^*_\mu(f^A, f^B) = \sum_{k=2}^{N} \frac{1}{2} \left(\sqrt{\left(w_k^A \mu^A - w_k^B \mu^B\right)^T \cdot (w_k^A \Sigma_k^A)^{-1} \cdot \left(w_k^A \mu^A - w_k^B \mu^B\right)} + \sqrt{\left(w_k^A \mu^A - w_k^B \mu^B\right)^T \cdot (w_k^B \Sigma_k^B)^{-1} \cdot \left(w_k^A \mu^A - w_k^B \mu^B\right)} \right) \tag{10}$$

$$D^*_\Sigma(f^A, f^B) = \sum_{k=2}^{N} D_\Sigma(w_k^A \Sigma_k^A, w_k^B \Sigma_k^B), \text{ using(4)}, \quad \text{which simplifies to} \tag{11}$$

$$= \sqrt{\frac{1}{2} \sum_{k=2}^{N} (\log \lambda_k^{AB})^2}, \quad \text{with } \{\lambda_k^{AB}\} = \mathrm{Eig} W_k^{AB}, \quad \text{where} \tag{12}$$

$$W_k^{AB} = (w_k^A \Sigma_k^A)^{-1/2} \cdot w_k^B \Sigma_k^B \cdot (w_k^A \Sigma_k^A)^{-1/2} = (w_k^B / w_k^A) \left((\Sigma_k^A)^{-1/2} \cdot \Sigma_k^B \cdot (\Sigma_k^A)^{-1/2} \right).$$

Figure 2 shows the effect on D^* of differing incorporated weighting sequences using (10), (12) for the case $\Sigma_k^A = \Sigma_k^B = \Sigma_k^B$ for random k-variate Gaussians with $k = 2, 3, 4, 5$. The weight sequences are for mixtures $A : w_k^A = (0.1, 0.2, 0.3, 0.4)$, $B : w_k^B = (0.25, 0.25, 0.25, 0.25)$, $C : w_k^C = (0.4, 0.3, 0.2, 0.1)$, and we see that $D^*(f^A, f^B) < D^*(f^B, f^C)$ consistently across ten random replications using incorporated weights.

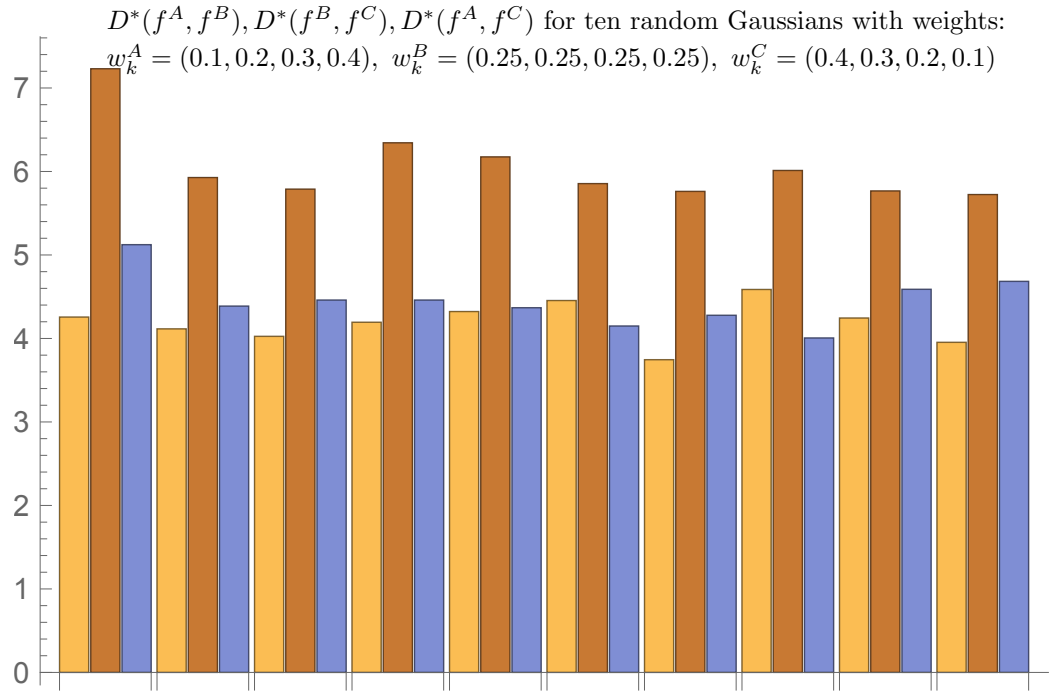

Figure 2: **Effect of incorported weights §2.1:** *Distances between pairs of mixtures of random k-variate Gaussians having $k = 2, 3, 4, 5$ variables, with increasing weights A, uniform weights B, and decreasing weights C. The three bars give $D^*(f^A, f^B), D^*(f^B, f^C), D^*(f^A, f^C)$ respectively, for ten different random sequences of k-variate Gaussians.*

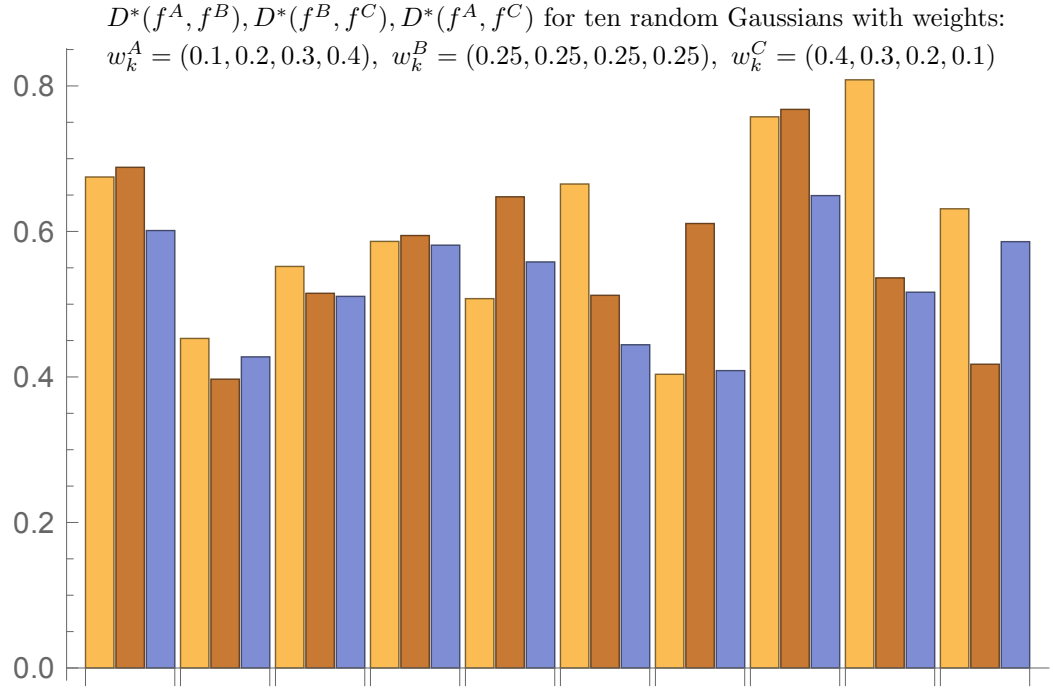

Figure 3: **Effect of averaged weights §2.2:** *Distances between pairs of mixtures of random k-variate Gaussians having $k = 2, 3, 4, 5$ variables, with increasing weights A, uniform weights B, and decreasing weights C. The three bars give $D^{\#}(f^A, f^B), D^{\#}(f^B, f^C), D^{\#}(f^A, f^C)$ respectively, for ten different random sequences of k-variate Gaussians.*

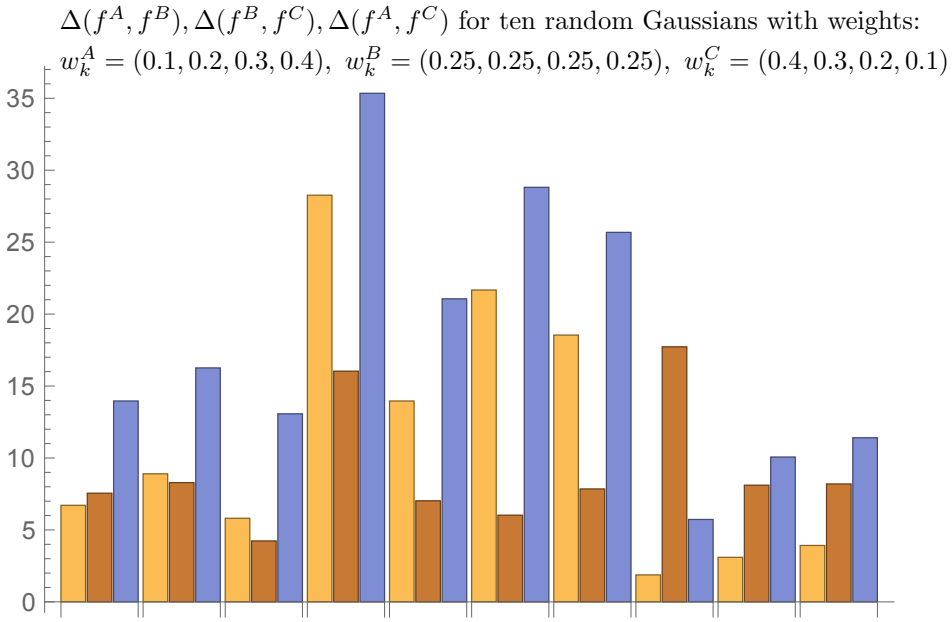

Figure 4: **Effect of weightings using mixture projection onto $\mathbb{C}$ §2.3:** *Distances between pairs of mixtures of random k-variate Gaussians having $k = 2, 3, 4, 5$ variables, with increasing weights A, uniform weights B, and decreasing weights C. The three bars give $\Delta(f^A, f^B), \Delta(f^B, f^C), \Delta(f^A, f^C)$ respectively, for ten different random sequences of k-variate Gaussians.*

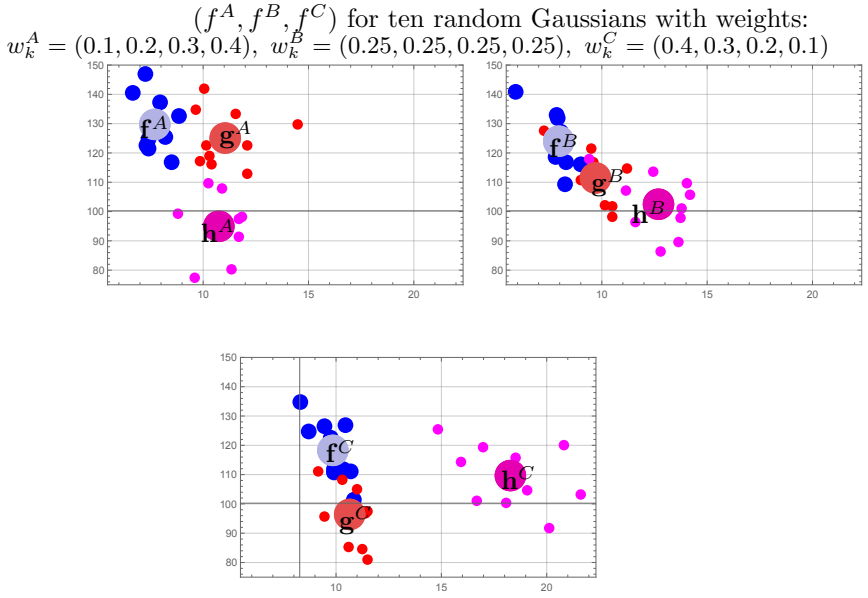

Figure 5: **Mixture projection onto $\mathbb{C}$ §2.3:** *Mixtures are shown plotted in $(||\mu||, ||\Sigma||)$-space for the 10 random k-variate Gaussians having $k = 2, 3, 4, 5$ variables, with increasing weights f^A, uniform weights f^B, and decreasing weights f^C. The g^A, g^B, g^C are for the same 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. The mean for each over the ten replications is shown as a large point.*

2.2 Averaged weights

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 $\delta\mu = (\mu^A - \mu^B)$:

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

$$\begin{aligned} D^{\#}_{\Sigma}(f^A, f^B) &= \sum_{k=2}^{N} \frac{1}{2}(w^A_k + w^B_k) D_{\Sigma}(\Sigma^A_k, \Sigma^B_k) \text{ using(4), which simplifies to} \\ &= \sum_{k=2}^{N} \frac{1}{2}(w^A_k + w^B_k) \sqrt{\frac{1}{2}\sum_{k=2}^{N} (\log \lambda^{AB}_k)^2} \quad \text{where } \{\lambda^{AB}_k\} = \mathrm{Eig} H^{AB}_k \text{ with} \end{aligned} \tag{14}$$

$$H^{AB}_k = \left((\Sigma^A_k)^{-1/2} \cdot \Sigma^B_k \cdot (\Sigma^A_k)^{-1/2} \right).$$

In this case, if $f^A = (\mu^A, \Sigma^A)$, and $f^B = (\mu^B, \Sigma^B)$ arise as differently weighted sums of the same sequence of covariances, then $\Sigma^A_k = \Sigma^B_k$ so H^{AB}_k is the identity matrix and $D^{\#}_{\Sigma}(f^A, f^B) = 0$. Figure 3 shows the effect on $D^{\#}$ of differing averaged weighting sequences using (13), (14).

2.3 Mixtures projected onto the complex plane

The idea here is simple: for each mixture distribution f^A given by a weighted sum (2) 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

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

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

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

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

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

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:

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

The result of using (17) to project mixtures onto the complex plane is shown in Figure 4. The three bars give $\Delta(f^A, f^B), \Delta(f^B, f^C), \Delta(f^A, f^C)$ respectively, for ten different random sequences of k-variate Gaussians. The three barcharts, in Figure 2, Figure 3 and Figure 4, use the same mixtures of multivariate Gaussians. It appears that the projection of mixtures onto the complex plane, Figure 4, as described in the present section gives a wider range of differences and shows the intuitively expected largest differences mostly between increasing and decreasing weight sequences, $\Delta(f^A, f^C)$ in the third columns of each replication.

Figure 5 shows a plot of the points $(||\mu||, ||\Sigma||) \in \mathbb{C}$ for the ten 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 ten replications is shown as a large point.

2.4 Acknowledgement

The methods described in §2.1 and §2.2 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 will be reported elsewhere in a joint paper.

References

[1] K. Arwini and C.T.J. Dodson. **Information Geometry Near Randomness and Near Independence.** Lecture Notes in Mathematics. Springer-Verlag, New York, Berlin, 2008.

[2] C. Atkinson and A.F.S. Mitchell. Rao's distance measure. *Sankhya: Indian Journal of Statistics* 48, A, 3 (1981) 345-365.

[3] P.S. Eriksen. Geodesics connected with the Fisher metric on the multivariate normal manifold. In C.T.J. Dodson, Editor, **Proceedings of the GST Workshop**, Lancaster (1987), 225-229. http://trove.nla.gov.au/version/21925860

[4] Jian Cao, Dian-hui Mao, Qiang Cai, Hai-sheng Li and Jun-ping Du. A review of object representation based on local features. *Journal of Zhejiang University-SCIENCE C (Computers & Electronics)* 14, 7 (2013) 495-504. doi:10.1631/jzus.CIDE1303

[5] 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. pp 2012-2016.