Texture Classification Using Rao's Distance on the Space of Covariance Matrices

Salem Said, Lionel Bombrun, Yannick Berthoumieu

Texture classification using Rao's distance on the space of covariance matrices

Salem Said, Lionel Bombrun, Yannick Berthoumieu

Laboratoire IMS (CNRS - UMR 5218), Université de Bordeaux
{salem.said, lionel.bombrun, yannick.berthoumieu }@ims-bordeaux.fr

Abstract. The current paper introduces new prior distributions on the zero-mean multivariate Gaussian model, with the aim of applying them to the classification of covariance matrices populations. These new prior distributions are entirely based on the Riemannian geometry of the multivariate Gaussian model. More precisely, the proposed Riemannian Gaussian distribution has two parameters, the centre of mass $\bar{Y}$ and the dispersion parameter σ. Its density with respect to Riemannian volume is proportional to $\exp(-d^2(Y;\bar{Y}))$, where $d^2(Y;\bar{Y})$ is the square of Rao's Riemannian distance. We derive its maximum likelihood estimators and propose an experiment on the VisTex database for the classification of texture images.

Keywords: Texture classification, Information geometry, Riemannian centre of mass, Mixture estimation, EM algorithm.

1 Introduction

In information geometry, a parametric family of probability densities is considered as a Riemannian manifold [1]. Precisely, the role of Riemannian metric is played by the Fisher metric, and that of Riemannian distance by Rao's distance. Rao's distance has been been widely used for several statistical applications including object detection and tracking, shape classification, and image segmentation [2–4]. Nevertheless, none of them have formulated it as a probabilistic approach to clustering on Riemannian manifolds, which is the main contribution of the paper.

More precisely, this paper introduces new Riemannian prior (denoted $G(\bar{Y},\sigma)$) as Gaussian distributions on the zero-mean multivariate Gaussian model. These distributions have a unique mode $\bar{Y}$ (the unique Riemannian centre of mass), and its dispersion away from $\bar{Y}$ is given by σ. In order to improve upon the performance obtained in [5], the present paper uses mixtures of Riemannian priors as prior distributions for classification. This allows for clustering analysis to be carried out using an expectation-maximisation, or EM, algorithm, instead of the essentially deterministic k-means approach of existing works, (*e.g.* [2–4]).

The paper is structured as follows. Section 2 recalls some definitions concerning the Riemannian geometry of covariance matrices. Section 3 introduces the

proposed Riemannian Gaussian distributions. After having presented its maximum likelihood estimators and its extension to mixture models in Section 4, an experiment on the VisTex database is proposed in Section 5 to evaluate the potential of the proposed prior for the classification of texture images. Due to the restriction length, all the mathematical proofs cannot be detailed here and will be given in a forthcoming journal paper.

2 Riemannian geometry of covariance matrices

Let $\mathcal{P}_m$ denote the space of all $m \times m$ real matrices Y which are symmetric and strictly positive definite,

$$Y^\dagger - Y = 0 \qquad x^\dagger Y x > 0 \text{ for all } x \in \mathbb{R}^m \tag{1}$$

where $\dagger$ denotes the transpose. In many applications [6], $\mathcal{P}_m$ arises as a space of tensors, such as structure tensors in image processing, or diffusion tensors in medical imaging, (in these examples, $m = 2, 3$). In general, $\mathcal{P}_m$ may also be thought of as the space of non-degenerate covariance matrices [7].

When thinking of the elements Y of $\mathcal{P}_m$ as covariance matrices, it is most suitable to do so within the framework of the normal covariance model [7][8]. This model associates to $Y \in \mathcal{P}_m$ the normal probability density function $P(x|Y)$ on $\mathbb{R}^m$, with mean $0 \in \mathbb{R}^m$ and covariance Y. Recall that $\log P(x|Y) = \ell(Y)$, where

$$\ell(Y) = -\frac{1}{2}\log\left[\det(2\pi Y)\right] - \frac{1}{2}\, x^\dagger Y^{-1} x. \tag{2}$$

Let us now recall the definition of the Fisher information matrix [8]. Let $p = m(m+1)/2$, the dimension of $\mathcal{P}_m$, and Θ an open subset of $\mathbb{R}^p$. Assume $\theta \mapsto Y(\theta)$ is a differentiable mapping from Θ to $\mathcal{P}_m$, which is a diffeomorphism. One refers to the mapping $\theta \mapsto Y(\theta)$ as a parameterisation of $\mathcal{P}_m$, with parameters $\theta = (\theta^a\,; a = 1, \ldots, p)$. Let $\ell(\theta)$ stand for $\ell(Y(\theta))$ where $\ell(Y)$ is the function defined in (2). The Fisher information matrix $I(\theta)$ has matrix elements

$$I_{ab}(\theta) = \mathbb{E}_\theta\left[\frac{\partial \ell(\theta)}{\partial \theta^a} \times \frac{\partial \ell(\theta)}{\partial \theta^b}\right], \tag{3}$$

where $\mathbb{E}_\theta$ denotes expectation with respect to the normal probability density function $p(x|Y(\theta))$.

A Riemannian metric on $\mathcal{P}_m$ is a quadratic form $ds^2(Y)$ which measures the *squared length of a small displacement* dY, separating two elements $Y \in \mathcal{P}_m$ and $Y + dY \in \mathcal{P}_m$. Here, dY is a symmetric matrix, since Y and $Y + dY$ are symmetric, by (1). The Rao-Fisher metric is the following [9][8],

$$ds^2(Y) = \mathrm{tr}\left([Y^{-1}dY]^2\right), \tag{4}$$

where tr() denotes the trace.

The Rao-Fisher metric, like any other Riemannian metric on $\mathcal{P}_m$, defines a Riemannian distance $d : \mathcal{P}_m \times \mathcal{P}_m \to \mathbb{R}_+$. This is called Rao's distance, and is

defined as follows [9][10]. Let $Y, Z \in \mathcal{P}_m$ and $c : [0,1] \to \mathcal{P}_m$ be a differentiable curve with $c(0) = Y$ and $c(1) = Z$. The length $L(c)$ of c is defined by

$$L(c) \;=\; \int_0^1 ds(c(t)) \;=\; \int_0^1 \|\dot{c}(t)\|\, dt, \tag{5}$$

where $\dot{c}(t) = \frac{dc}{dt}$. Rao's distance $d(Y, Z)$ is the infimum of $L(c)$ taken over all differentiable curves c as above.

A major property of the Rao-Fisher metric is the following. When equipped with the Rao-Fisher metric, the space $\mathcal{P}_m$ is a Riemannian manifold of negative sectional curvature. One implication of this property, (called the Cartan-Hadamard theorem [10]), is that the infimum of $L(c)$ is realised by a unique curve γ, known as the geodesic connecting Y and Z. The equation of this curve is the following [9],

$$\gamma(t) = Y^{1/2}\, (Y^{-1/2} Z Y^{-1/2})^t\, Y^{1/2}. \tag{6}$$

Given the expression (6), it is possible to compute $L(\gamma)$ from (5). This is precisely Rao's distance $d(Y, Z)$. It turns out,

$$d^{\,2}(Y, Z) = \mathrm{tr}\,[\log(Y^{-1/2} Z Y^{-1/2})]^2. \tag{7}$$

Since the Rao-Fisher metric gives a mean of measuring length, it can also be used to measure volume. Indeed, (based on the elementary fact that the "volume of a cube is the product of the lengths of its sides"), the Riemannian volume element associated to the Rao-Fisher metric is defined to be [9]

$$dv(Y) = \det(Y)^{-\frac{m+1}{2}} \prod_{i \le j} dY_{ij}. \tag{8}$$

All matrix functions appearing in (6) and (7), (square root, power and logarithm), should be understood as symmetric positive definite matrix functions.

3 Riemannian Gaussian distributions

The main theoretical contribution of the present paper is to give an original exact formulation of Riemannian Gaussian distributions. These are probability distributions on $\mathcal{P}_m$, whose probability density function, with respect to the Riemannian volume element (8), is of the form,

$$p(Y \,|\, \bar{Y}, \sigma) = \frac{1}{Z(\sigma)}\, \exp\left[-\frac{d^{\,2}(Y, \bar{Y})}{2\sigma^2} \right], \tag{9}$$

where $\bar{Y} \in \mathcal{P}_m$ and $\sigma > 0$ are parameters, and where $d(Y, \bar{Y})$ is Rao's distance, given by (7). For brevity, a Riemannian Gaussian distribution, with probability density function (9), will be called a Gaussian distribution, and denoted $G(\bar{Y}, \sigma)$.

The parameter $\bar{Y}$ is called the centre of mass, and σ is called the dispersion, of the distribution $G(\bar{Y}, \sigma)$.

Distributions of the form (9) were considered by Pennec [11], defined on general Riemannian manifolds. However, in existing literature, their treatment remains incomplete, as it is based on asymptotic formulae, valid only in the limit where the parameter σ is small, (see [11] (Theorem 5., Page 140) and [12] (Theorem 3.1.1., Page 434)). In addition to being only approximations, such formulae are quite difficult, both to evaluate and to apply. These issues, (lack of an exact expression and difficulty of application), are fully overcome in the following.

Note also that a more sophisticated description by means of a concentration matrix instead of a scalar dispersion parameter σ is possible. This approach has notably been introduced in [11].

3.1 Maximum likelihood estimation

Let $Y_1, \ldots, Y_N$ be N independent samples from a Gaussian distribution $G(\bar{Y}, \sigma)$. Based on these samples, the maximum likelihood estimate of the parameter $\bar{Y}$ is the empirical Riemannian centre of mass $\hat{Y}_N$ of $Y_1, \ldots, Y_N$ defined as the unique global minimiser $\hat{Y}_N$ of $\mathcal{E}_N : \mathcal{P}_m \to \mathbb{R}$,

$$\mathcal{E}_N(Y) = \frac{1}{N} \sum_{n=1}^{N} d^2(Y, Y_n). \tag{10}$$

Moreover, the maximum likelihood estimate of the parameter σ is the solution $\hat{\sigma}_N$ of the equation, (for unknown σ),

$$\sigma^3 \times \frac{d}{d\sigma} \log Z(\sigma) = \mathcal{E}_N(\hat{Y}_N). \tag{11}$$

Both $\hat{Y}_N$ and $\hat{\sigma}_N$ exist and are unique for any realisation of the samples $Y_1, \ldots, Y_N$. In practice, $\bar{Y}$ is first estimated according to (10) then the estimation of σ is proceed by (11).

3.2 Application to $\mathcal{P}_2$

In (9), the normalising factor $Z(\sigma)$ can be expressed under an integral form as

$$Z(\sigma) = \int_{\mathcal{P}_m} f(Y | \bar{Y}, \sigma)\, dv(Y), \tag{12}$$

where $dv(Y)$ is the Riemannian volume element (8). It is interesting to note that $Z(\sigma)$ is independent from the centre of mass $\bar{Y}$. For the space of 2×2 covariance matrices (*i.e.* $m = 2$), the normalising factor admits the following close form expression:

$$Z(\sigma) = 4\pi^2\sigma^2 \ \exp(\sigma^2/4)\ \mathrm{erf}(\sigma/2), \tag{13}$$

where erf() is the error function.

4 EM algorithm for mixture estimation

While successful in application to specific data sets, the Bayesian approach of [5] summarised in the previous section fails to take into account the presence of within-class diversity. Precisely, this approach assumes that the given learning sequence is immediately subdivided into clusters, whose members display "homogeneous" properties, in the sense that they can be faithfully modelled as belonging to the same Riemannian prior. Clearly, this is a restrictive assumption. In the presence of within-class diversity, a learning sequence should be subdivided into classes, whose members display "heterogeneous" properties, in the sense that they may belong, within the same class, to different clusters, each corresponding to a different Riemannian prior.

Here, this situation is formulated as follows. If a class $\mathcal{C}$, whose members are points $Y_1, \ldots, Y_N \in \mathcal{P}_m$, is expected to contain K clusters, respectively corresponding to Riemannian priors $G(\bar{Y}_a, \sigma_a)$, where $a = 1, \ldots, K$, then $\mathcal{C}$ is modelled as a sample of size N, drawn from the mixture of Riemannian priors

$$p(Y|\Theta) = \sum_{a=1}^{K} \varpi_a \, p(Y|\bar{Y}_a, \sigma_a) \tag{14}$$

where $\varpi_1, \ldots, \varpi_K$ are positive weights, with $\sum_{a=1}^{K} \varpi_a = 1$, and each density $p(Y|\bar{Y}_a, \sigma_a)$ is given by (9).

Now, assume a training sequence is subdivided into classes, each containing a known numbers of clusters. In order to implement a decision rule which associates any test object, described by $Y_t \in \mathcal{P}_m$, to the most likely cluster within the training sequence, it is necessary, for each class $\mathcal{C}$, modelled by (14), to find maximum likelihood estimates of the mixture parameters $\vartheta = (\varpi_a, \bar{Y}, \sigma_a)$. Here, this task is realised using an expectation-maximisation (EM) algorithm. Following [13], the starting point for the EM algorithm is the introduction of the following quantities

$$\omega_a(Y_j) \propto \varpi_a \times p(Y_j|\bar{Y}_a, \sigma_a) \qquad n_a = \sum_{j=1}^{N} \omega_a(Y_j) \tag{15}$$

where, $\propto$ denotes proportionality, so that $\sum_a \omega_a(Y_j) = 1$. To emphasise the fact that $\omega_a(Y_j)$ and n_a are computed for a given value of $\vartheta = (\varpi_a, \bar{Y}, \sigma_a)$, they shall be denoted $\omega_a(Y_j, \vartheta)$ and $n_a(\vartheta)$. The algorithm iteratively updates $\hat{\vartheta} = (\hat{\varpi}_a, \hat{Y}_a, \hat{\sigma}_a)$, an approximation of the maximum likelihood estimate of $\vartheta = (\varpi_a, \bar{Y}_a, \sigma_a)$. Precisely, the update rules for $\hat{\varpi}_a$, $\hat{Y}_a$, and $\hat{\sigma}_a$ are repeated as long as this introduces a sensible change in the values of $\hat{\varpi}_a$, $\hat{Y}_a$, and $\hat{\sigma}_a$. As this is a non convex problem optimization, we reach a local stationary point. It is hence useful to run the algorithm several times, with different initialisations to reach the global optimum. The update rules are the following,

▶ **Update for $\hat{\varpi}_a$:** Based on the current value of $\hat{\vartheta}$, assign to $\hat{\varpi}_a$ the new value

$$\hat{\varpi}_a^{\text{new}} = \frac{n_a(\hat{\vartheta})}{\sum_{a=1}^{K} n_a(\hat{\vartheta})}. \tag{16}$$

► **Update for $\hat{Y}_a$:** Based on the current value of $\hat{\vartheta}$, compute $\hat{Y}_a$ to be the global minimiser of the following function,

$$V(Y|\hat{\vartheta}) = \frac{1}{2}\sum_{j=1}^{N} \omega_a(Y_j, \hat{\vartheta}) \times d^2(Y_j, Y). \tag{17}$$

$\hat{Y}_a$ is the empirical Riemannian centre of mass which may be estimated by a Riemannian gradient descent algorithm (See [12] for more details).

► **Update for $\hat{\sigma}_a$:** Based on the current value of $\hat{\vartheta}$, compute $\hat{\sigma}_a$ to be the solution of the following equation, for unknown σ,

$$F(\sigma) = \frac{1}{2n_a(\hat{\vartheta})}\, V(\hat{Y}_a|\hat{\vartheta}) \tag{18}$$

where $F(\sigma) = \sigma^3 \times \frac{d}{d\sigma} \log Z(\sigma)$. Practically, a Newton-Raphson procedure is employed to solve (18).

These three update rules should be performed in the above given order. Therefore, the "current value of $\hat{\vartheta} = (\hat{\varpi}_a, \hat{Y}_a, \hat{\sigma}_a)$" is different, in each one of them. For instance, in the update rule of $\hat{\sigma}_a$, the current value of $\hat{Y}_a$ is found from the minimisation of (17), just before.

5 Application to texture image classification

The present section proposes a new decision rule, for the classification of covariance matrices, and applies it to texture classification, using the VisTex database [14]. The following numerical experiment was carried out. Half of the database was used for training, and the other half for testing. Each training image was subdivided into 169 patches of 128×128 pixels, with a 32 pixel overlap. For each training patch, 6 wavelet subbands were computed using the stationary wavelet decomposition (with 2 scale) with Daubechies' filter db4. In texture classification, multivariate models were found very effective for modelling the spatial dependency of wavelet coefficients. Hence, two spatial neighborhoods (horizontal dH and vertical dV) of one pixel were considered. Each subband s of patch n gives two bivariate normal populations $\Pi_{s,n,dH}$ and $\Pi_{s,n,dV}$, represented respectively by a point $Y_{s,n,dH}$ and $Y_{s,n,dV} \in \mathcal{P}_2$. The size of the feature space is hence $F = 12$ (6 subbands times 2 spatial supports). For the sake of simplicity, let say that the training patch n is represented by a set of F covariance matrices denoted $Y_{f,n}$.

For each training class, a set of $N = 84$ "arrays" are extracted. These arrays Y_j are a collection of F covariance matrices and are considered as multivariate

Prior	Overall Accuracy
Riemannian prior on $\mathcal{P}_2$ (K=1) (9)	$86.27 \pm 0.45\%$
Mixture prior on $\mathcal{P}_2$ (EM, K=3, (14))	$\mathbf{94.31 \pm 0.42}\%$
Mixture prior on $\mathcal{P}_2$ (K-means, K=3) [15]	$92.40 \pm 0.46\%$
Riemannian prior on $\mathcal{H}$ [5]	$83.29 \pm 0.51\%$
Mixture prior on $\mathcal{H}$ [17]	$88.50 \pm 0.88\%$
Conjugate prior on $\mathcal{H}$	$83.48 \pm 0.53\%$

Table 1. Classification performance on the VisTex database.

realisations of a mixture distribution (14), with independent components $Y_{f,n}$ since wavelet subbands are assumed independent. Each class is assumed to contain the same number K of clusters, and is modelled as a sample drawn from a mixture distribution (14). First, the EM algorithm of Section 4 is applied to each class, leading to maximum likelihood estimates $(\hat{\varpi}_a, \hat{Y}_{f,a}, \hat{\sigma}_a)$, for $a = 1, \ \ldots, K$ and $f = 1, \ \ldots, F$.

Each triple of such estimates defines a cluster within the training sequence. Denote the total number of clusters defined in this way L, and the corresponding maximum likelihood estimates $(\hat{\varpi}_c, \hat{Y}_{f,c}, \hat{\sigma}_c)$, for $c = 1, \ldots, L$ and $f = 1, \ \ldots, F$. Then, a test population represented by $Y_t \in \mathcal{P}_2$ is associated to the class of the cluster C_*, realising the minimum over c of,

$$-\log \hat{\varpi}_c + \log Z(\hat{\sigma}_c) + \frac{1}{2\hat{\sigma}_c^2} \sum_{f=1}^{F} d^2(Y_t, \hat{Y}_{f,c}). \tag{19}$$

This is the new decision rule, proposed for use with the mixture model (14). Note that the case $K = 1$ reduces to a Bayesian classifier with the proposed Riemannian Gaussian distribution.

Table 1 displays the classification performance in terms of overall accuracy on the VisTex database. The first two lines correspond to the proposed Riemannian prior (9) on $\mathcal{P}_2$ with respectively $K = 1$ and $K = 3$. The third line corresponds to a nearest centre of mass classifier classically employed in literature [15]. In such case, the centres of mass $\hat{Y}_{f,c}$ are estimated by using a K-means algorithm. Some comparisons are also carried out with univariate normal populations where the mean and standard deviation are computed on Gabor energy subbands (see [16] for more details). In such case, a Riemannian prior on the Poincaré upper half-plane $\mathcal{H}$ has been introduced in [5] and further extended to mixture models [17]. A conjugate normal-inverse gamma prior on $\mathcal{H}$ is also displayed on the last line of Table 1

As observed in Table 1, the proposed Riemannian prior on $\mathcal{P}_2$ based on a mixture model displays much better performance than other prior. A significant gain of respectively 2% and 6% is observed when compared to a nearest centre of mass classifier [15] and to a mixture prior on the Poincaré upper half-plane $\mathcal{H}$.

6 Conclusion

This paper has addressed the problem of classification using Rao's distance on the space of covariance matrices. To this aim, a Riemannian Gaussian distributions has been introduced. Analogous to the classical multivariate Gaussian distribution, the proposed Riemannian Gaussian distribution has two parameters, the centre of mass $\bar{Y}$ and the dispersion parameter σ. The main difference relies on the use of the Riemannian distance in the exponential of the pdf instead of the Mahalanobis distance. After having presented its maximum likelihood estimators and its extension to mixture models, an experiment on the VisTex database have shown the potential of the proposed model for the classification of texture images.

7 Acknowledgment

This study has been carried out with financial support from the French State, managed by the French National Research Agency (ANR) in the frame of the "Investments for the future" Programme IdEx Bordeaux-CPU (ANR-10-IDEX-03-02).

References

1. Amari, S., Nagaoka, H.: Methods of information geometry. American Mathematical Society (2000)
2. Porikli, F., Tuzel, O., Meer, P.: Covariance tracking using model update based means on Riemannian manifolds. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2006) 728–735
3. Kurtek, S., Klassen, E., Ding, Z., Avison, M., Srivastava, A.: Parameterization invariant shape statistics and probabilistic classification of anatomical surfaces. In Székely, G., Hahn, H., eds.: IPMI. Volume 6801 of Lecture Notes in Computer Science., Springer (2011) 147–158
4. Gu, X., Deng, J., Purvis, M.: Improving superpixel-based image segmentation by incorporating color covariance matrix manifolds. In: International Conference on Image Processing (ICIP). (2014) 4403–4406
5. Said, S., Bombrun, L., Berthoumieu, Y.: New Riemannian priors on the univariate normal model. Entropy **16**(7) (2014) 4015–4031
6. Weickert, J., Hagen, H., eds.: Visualization and processing of tensor fields. Mathematics visualization. Springer, Berlin, Heidelberg (2006)
7. Muirhead, R.J.: Aspects of multivariate statistical theory. John Wiley & Sons, New York (1982)
8. Atkinson, C., Mitchell, A.: Rao's distance measure. Sankhya Ser. A **43** (1981) 345–365
9. Terras, A.: Harmonic analysis on symmetric spaces and applications, Vol. II. Springer-Verlag, New York (1988)
10. Helgason, S.: Differential geometry, Lie groups, and symmetric spaces. Americal Mathematical Society (2001)

11. Pennec, X.: Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements. J. Math. Imaging Vis. **25**(1) (2006) 127–154
12. Lenglet, C., Rousson, M., Deriche, R., Faugeras, O.: Statistics on the manifold of multivariate normal distributions. J. Math. Imaging Vis. **25**(3) (2006) 423–444
13. Mengersen, K., Robert, C., Titterington, M.: Mixtures : estimation and applications. Wiley (2011)
14. : MIT Vision and Modeling Group. Vision Texture. VisTex : Vision Texture Database, `http://vismod.media.mit.edu/pub/vistex`
15. Barachant, A., Bonnet, S., Congedo, M., Jutten, C.: Multiclass brain–computer interface classification by Riemannian geometry. IEEE Trans. Biomed. Eng. **59**(4) (2012) 920–928
16. Grigorescu, S., Petkov, N., Kruizinga, P.: Comparison of texture features based on Gabor filters. IEEE Trans. Im. Proc. **11**(10) (2002) 1160–1167
17. Said, S., Bombrun, L., Berthoumieu, Y.: Texture classification using Rao's distance: An EM algorithm on the Poincaré half plane. In: International Conference on Image Processing (ICIP). (2015)