Riemannian Laplace distribution on the space of symmetric positive definite matrices

Hatem Hajri, Ioana Ilea, Salem Said, Lionel Bombrun, Yannick Berthoumieu

Riemannian Laplace distribution on the space of symmetric positive definite matrices

Hatem Hajri, Ioana Ilea, Salem Said, Lionel Bombrun and Yannick Berthoumieu

Abstract

The Riemannian geometry of the space $\mathcal{P}_m$, of $m \times m$ symmetric positive definite matrices, has provided effective tools to the fields of medical imaging, computer vision, and radar signal processing. Still, an open challenge remains, which consists in extending these tools to correctly handle the presence of outliers (or abnormal data), arising from excessive noise or faulty measurements. The present paper tackles this challenge by introducing new probability distributions, called *Riemannian Laplace distributions* on the space $\mathcal{P}_m$. First, it shows that these distributions provide a statistical foundation for the concept of Riemannian median, which offers improved robustness in dealing with outliers (in comparison to the more popular concept of Riemannian centre of mass). Second, it describes an original expectation-maximisation algorithm, for estimating mixtures of Riemannian Laplace distributions. This algorithm is applied to the problem of texture classification, in computer vision, which is considered in the presence of outliers. It is shown to give significantly better performance with respect to other recently proposed approaches.

1 Introduction

Data with values in the space $\mathcal{P}_m$, of $m \times m$ symmetric positive definite matrices, play an essential role in many applications, including medical imaging [1, 2], computer vision [3, 4, 5, 6, 7], and radar signal processing [8, 9]. In these applications, the location where a dataset is centered has a special interest. While several definitions of this location are possible, its meaning as a representative of the set should be clear. Perhaps, the most known and well-used quantity to represent a centre of a dataset is the Fréchet mean. Given a set of points $Y_1, \cdots, Y_n$ in $\mathcal{P}_m$, their Fréchet mean is defined to be

$$\text{Mean}(Y_1, \cdots, Y_n) = \text{argmin}_{Y \in \mathcal{P}_m} \sum_{i=1}^{n} d^2(Y, Y_i) \tag{1}$$

where d is Rao's Riemannian distance on $\mathcal{P}_m$ [10, 11].

Recently, a new distribution on $(\mathcal{P}_m, d)$ has been introduced [12, 13]. This distribution called Riemannian Gaussian distribution depends on two parameters $\bar{Y} \in \mathcal{P}_m$ and $\sigma > 0$. Its density with respect to the Riemannian volume form $dv(Y)$ of $\mathcal{P}_m$ (see formula (13) in Section 2) is

$$\frac{1}{Z_m(\sigma)} \exp\left[-\frac{d^2(Y, \overline{Y})}{2\sigma^2}\right] \tag{2}$$

where $Z_m(\sigma)$ is a normalising factor depending only on σ (and not on $\bar{Y}$). For the Gaussian distribution (2), maximum likelihood estimate (MLE) for the parameter $\bar{Y}$ based on observations $Y_1, \cdots, Y_n$ corresponds to the mean (1). In [13], a detailed study of statistical inference for this distribution was given and then applied to the classification of data in $\mathcal{P}_m$ showing that it yields better performance, in comparison to recent approaches [2].

When a dataset contains extreme values (or outliers), because of the impact of these values on d^2, the mean becomes less useful. It is usually replaced with the Riemannian median

$$\text{Median}(Y_1, \cdots, Y_n) = \text{argmin}_{Y \in \mathcal{P}_m} \sum_{i=1}^{n} d(Y, Y_i) \tag{3}$$

The Gaussian distribution on $\mathcal{P}_m$ also loses its robustness properties. The aim of the present paper is to remedy this problem by introducing the Riemannian Laplace distribution while maintaining the same one to one relation between MLE and the Riemannian median. This will be shown to offer considerable improvement in dealing with outliers.

This paper is organised as follows.

Section 2 reviews the Riemannian geometry of $\mathcal{P}_m$, when this manifold is equipped with the Riemannian metric known as the Rao-Fisher, or affine invariant metric [10, 11]. In particular, it gives analytic expressions for geodesic curves, Riemannian distance and recalls the invariance of Rao's distance under affine transformations.

Section 3 introduces the Laplace distribution $\mathcal{L}(\bar{Y}, \sigma)$ through its probability density function with respect to the volume form $dv(Y)$

$$p(Y|\overline{Y}, \sigma) = \frac{1}{\zeta_m(\sigma)} \exp\left[-\frac{d(Y, \overline{Y})}{\sigma}\right]$$

Here σ lies in an interval $]0, \sigma_{\max}[$ with $\sigma_{\max} < \infty$. This is because the normalising constant $\zeta_m(\sigma)$ becomes infinite for $\sigma \geq \sigma_{\max}$. It will be shown that $\zeta_m(\sigma)$ depend only on σ (and not on $\bar{Y}$) for all $\sigma < \sigma_{\max}$. This important fact leads to simple expressions of MLEs of $\overline{Y}$ and σ. In particular, the MLE of $\bar{Y}$ based on a family of observations

$Y_1, \cdots, Y_N$ sampled from $\mathcal{L}(\bar{Y}, \sigma)$ is given by the median of $Y_1, \cdots, Y_N$ defined by (3) where d is Rao's distance.

Section 4 focuses on mixtures of Riemannian Laplace distributions on $\mathcal{P}_m$. A distribution of this kind has a density

$$p(Y|(\omega_\mu, \overline{Y}_\mu, \sigma_\mu)_{1\leq\mu\leq M}) = \sum_{\mu=1}^{M} \varpi_\mu p(Y|\overline{Y}_\mu, \sigma_\mu) \tag{4}$$

with respect to the volume form $dv(Y)$. Here M is the number of mixture components, $\varpi_\mu > 0, \overline{Y}_\mu \in \mathcal{P}_m, \sigma_\mu > 0$ for all $1 \leq \mu \leq M$ and $\sum_{\mu=1}^{M} \varpi_\mu = 1$. A new EM (expectation maximisation) algorithm which computes maximum likelihood estimates of the mixture parameters $(\varpi_\mu, \bar{Y}_\mu, \sigma_\mu)_{1\leq\mu\leq M}$ is provided. The problem of order selection of the number M in (4) is also discussed and performed using the Bayesian Information criterion (BIC) [14].

Section 5 is an application of the previous material to the classification of data with values in $\mathcal{P}_m$ which contain outliers (abnormal data points). Assume given a training sequence $Y_1, \cdots, Y_n \in \mathcal{P}_m$. Using the EM algorithm developed in Section 4, it is possible to subvidide this sequence into disjoint classes. To classify new data points, a classification rule is proposed. The robustness of this rule lies in the fact that it is based on the distances between new observations and the respective medians of classes instead of the means [13]. This rule will be illustarted by an application to the problem of texture classification in computer vision. The obtained results show improved performance with respect to recent approaches which use the Riemannian Gaussian distribution [13].

2 Riemannian geometry of $\mathcal{P}_m$

This section introduces necessary background on the Riemannian geometry of $\mathcal{P}_m$, the space of symmetric positive definite matrices of size $m \times m$. Precisely, $\mathcal{P}_m$ is equipped with the Riemannian metric known as the affine-invariant metric. First, analytic expressions are recalled for geodesic curves and Riemannian distance. Then, two properties are stated, which are fundamental to the following. These are affine-invariance of the Riemannian distance, and existence and uniqueness of Riemannian medians.

The affine-invariant metric, called the Rao-Fisher metric in information geometry, has the following expression

$$g_Y(A, B) = \mathrm{tr}(Y^{-1}AY^{-1}B) \tag{5}$$

where $Y \in \mathcal{P}_m$ and $A, B \in T_Y\mathcal{P}_m$, the tangent space to $\mathcal{P}_m$ at Y, which is identified with the vector space of $m \times m$ symmetric matrices. The Riemannian metric

(5) induces a Riemannian distance on $\mathcal{P}_m$ as follows. The length of a smooth curve $c : [0, 1] \to \mathcal{P}_m$ is given by

$$L(c) = \int_0^1 \sqrt{g_{c(t)}(\dot{c}(t), \dot{c}(t))}\, dt \tag{6}$$

where $\dot{c}(t) = \frac{dc}{dt}$. For $Y, Z \in \mathcal{P}_m$, the Riemannian distance $d(Y, Z)$, called Rao's distance in information geometry, is defined to be

$$d(Y, Z) = \inf \{ L(c), c : [0, 1] \to \mathcal{P}_m \text{ is a smooth curve with } c(0) = Y, c(1) = Z \}.$$

This infimum is achieved by a unique curve $c = \gamma$, called the geodesic connecting Y and Z which has the following equation [10, 15]

$$\gamma(t) = Y^{1/2} \left(Y^{-1/2} Z Y^{-1/2}\right)^t Y^{1/2} \tag{7}$$

Here and throughout the following all matrix functions (for example, square root, logarithm or power) are understood as symmetric matrix functions [16]. By definition, $d(Y, Z)$ coincides with $L(\gamma)$, which turns out to be

$$d^2(Y, Z) = \mathrm{tr}\left[\log(Y^{-1/2} Z Y^{-1/2})\right]^2 \tag{8}$$

Equipped with the affine-invariant metric (5), the space $\mathcal{P}_m$ enjoys two useful properties, which are the following. The first property is invariance under affine transformations [15, 10]. Recall that an affine transformation of $\mathcal{P}_m$ is a mapping $Y \mapsto Y \cdot A$, where A is an invertible real matrix of size $m \times m$,

$$Y \cdot A = A^\dagger\, Y\, A \tag{9}$$

and † denotes the transpose. Denote by $\mathrm{GL}(m)$, the group of $m \times m$ invertible real matrices on $\mathcal{P}_m$. Then the action of $\mathrm{GL}(m)$ on $\mathcal{P}_m$ is transitive. This means that for any $Y, Z \in \mathcal{P}_m$, there exists $A \in \mathrm{GL}(m)$ such that $Y.A = Z$. Moreover the Riemannian distance (8) is invariant by affine transformations in the sense that for all $Y, Z \in \mathcal{P}_m$

$$d(Y, Z) = d(Y \cdot A, Z \cdot A) \tag{10}$$

where $Y \cdot A$ and $Z \cdot A$ are defined by (9). The transitivity of the action (9) and the isometry property (10) make $\mathcal{P}_m$ a Riemannian homogeneous space.

The affine-invariant metric (5) turns $\mathcal{P}_m$ into a Riemannian manifold of negative sectional curvature [17, 10]. As a result, $\mathcal{P}_m$ enjoys the property of existence and uniqueness of Riemannian medians. The Riemannian median of N points $Y_1, \cdots, Y_N \in \mathcal{P}_m$ is defined to be

$$\hat{Y}_N = \mathrm{argmin}_Y \sum_{n=1}^{N} d(Y, Y_n) \tag{11}$$

where $d(Y, Y_n)$ is the Riemannian distance (8). If $Y_1, \cdots, Y_N$ do not belong to the same geodesic, then $\hat{Y}_N$ exists and is unique [18]. More generally, for any probability measure π on $\mathcal{P}_m$, the median of π is defined to be

$$\hat{Y}_\pi = \text{argmin}_Y \int_{\mathcal{P}_m} d(Y, Z) d\pi(Z) \tag{12}$$

Note that (12) reduces to (11) for $\pi = \frac{1}{N} \sum_{n=1}^{N} \delta_{Y_n}$. If the support of π is not carried by a single geodesic, then again $\hat{Y}_\pi$ exists and is unique by the main result of [18].

To end this paragraph, consider the Riemannian volume associated to the affine-invariant Riemannian metric [10]

$$dv(Y) = \det(Y)^{-\frac{m+1}{2}} \prod_{i \leq j} dY_{ij} \tag{13}$$

where the indices denote matrix elements. The Riemannian volume is used to define the integral of a function $f : \mathcal{P}_m \to \mathbb{R}$ as

$$\int_{\mathcal{P}_m} f(Y) dv(Y) = \int \ldots \int f(Y) \, \det(Y)^{-\frac{m+1}{2}} \prod_{i \leq j} dY_{ij} \tag{14}$$

where the integral on the right hand side is a multiple integral over the $m(m+1)/2$ variables, Y_{ij} with $i \leq j$. The integral (14) is invariant under affine transformations. Precisely

$$\int_{\mathcal{P}_m} f(Y \cdot A) dv(Y) = \int_{\mathcal{P}_m} f(Y) dv(Y) \tag{15}$$

where $Y \cdot A$ is the affine transformation given by (9). It takes on a simplified form when $f(Y)$ only depends on the eigenvalues of Y. Precisely, let the spectral decomposition of Y be given by $Y = U^\dagger \, \text{diag}(e^{r_1}, \cdots, e^{r_m}) \, U$, where U is an orthogonal matrix and $e^{r_1}, \cdots, e^{r_m}$ are the eigenvalues of Y. Assume that $f(Y) = f(r_1, \ldots, r_m)$, then the invariant integral (14) reduces to

$$\int_{\mathcal{P}_m} f(Y) dv(Y) = c_m \times \int_{\mathbb{R}^m} f(r_1, \cdots, r_m) \prod_{i<j} \sinh \left(\frac{|r_i - r_j|}{2} \right) dr_1 \cdots dr_m \tag{16}$$

where the constant c_m is given by $c_m = \frac{1}{m!} \times \omega_m \times 8^{\frac{m(m-1)}{4}}, \omega_m = \frac{\pi^{m^2/2}}{\Gamma_m(m/2)}$ and Γ_m is the multivariate Gamma function given in [19].

3 Riemannian Laplace Distribution on $\mathcal{P}_m$

3.1 Definition of $\mathcal{L}(\bar{Y}, \sigma)$

The Riemannian Laplace distribution on $\mathcal{P}_m$ is defined by analogy with the well-known Laplace distribution on $\mathbb{R}$. Recall the density of the Laplace distribution on

$\mathbb{R}$,

$$p(x|\bar{x},\sigma) = \frac{1}{2\sigma} e^{-|x-\bar{x}|/\sigma}$$

where $\bar{x} \in \mathbb{R}$ and $\sigma > 0$. This is a density with respect to the length element dx on $\mathbb{R}$. The density of the Riemannian Laplace distribution on $\mathcal{P}_m$ will be given by

$$p(Y|\,\bar{Y},\sigma) = \frac{1}{\zeta_m(\sigma)} \exp\left[-\frac{d(Y,\bar{Y})}{\sigma}\right] \tag{17}$$

Here, $\bar{Y} \in \mathcal{P}_m$, $\sigma > 0$, and the density is with respect to the Riemannian volume element (13) on $\mathcal{P}_m$. The normalising factor $\zeta_m(\sigma)$ appearing in (17) is given by the integral

$$\int_{\mathcal{P}_m} \exp\left[-\frac{d(Y,\bar{Y})}{\sigma}\right] dv(Y)$$

Assume for now this integral is finite for some choice of $\bar{Y}$ and σ. It is possible to show that its value does not depend on $\bar{Y}$. To do so recall that the action of $\mathrm{GL}(m)$ on $\mathcal{P}_m$ is transitive. As a consequence, there exists $A \in \mathcal{P}_m$ such that $\bar{Y} = I.A$ where $I.A$ is defined as in (9). From (10), it follows that $d(Y,\bar{Y}) = d(Y,I.A) = d(Y.A^{-1},I)$. From the invariance property (15)

$$\int_{\mathcal{P}_m} \exp\left[-\frac{d(Y,\bar{Y})}{\sigma}\right] dv(Y) = \int_{\mathcal{P}_m} \exp\left[-\frac{d(Y,I)}{\sigma}\right] dv(Y)$$

The integral on the right does not depend on $\bar{Y}$ which proves the above claim.
The last integral representation and formula (16) lead to the following explicit expression

$$\zeta_m(\sigma) = c_m \times \int_{\mathbb{R}^m} e^{-\frac{|r|}{\sigma}} \prod_{i<j} \sinh\left(\frac{|r_i - r_j|}{2}\right) dr_1 \cdots dr_m \tag{18}$$

where $|r| = (r_1^2 + \cdots + r_2^m)^{\frac{1}{2}}$ and c_m is the same constant as in (16).

A distinctive feature of the Riemannian Laplace distribution on $\mathcal{P}_m$, in comparison to the Laplace distribution on $\mathbb{R}$, is that there exist certain values of σ for which it can not be defined. This is because the integral (18) diverges for certain values of this parameter. This leads to the following definition.

Definition 3.1. *Set $\sigma_m = \sup\{\sigma > 0 : \zeta_m(\sigma) < \infty\}$. Then for $\bar{Y} \in \mathcal{P}_m$ and $\sigma \in (0,\sigma_m)$, the Riemannian Laplace distribution on $\mathcal{P}_m$, denoted by $\mathcal{L}(\bar{Y},\sigma)$, is defined as the probability distribution on $\mathcal{P}_m$ whose density with respect to $dv(Y)$ is given by (17), where $\zeta_m(\sigma)$ is defined by (18).*

As a first example of this definition, consider the case $m = 2$. In this case, the exact value of σ_2 is $\sqrt{2}$. In fact $\zeta_2(\sigma)$ is finite for all $\sigma < \sqrt{2}$ as follows from the inequality

$\sinh(\frac{|r_1-r_2|}{2}) \leq \frac{1}{2}e^{\sqrt{2}|r|}$. In order to check $\zeta_2(\sqrt{2}) = \infty$, it is necessary to show

$$\int_{\mathbb{R}^2} \exp(-\frac{|r|}{\sqrt{2}} + \frac{|r_1 - r_2|}{2})dr_1dr_2 = \infty$$

This is indeed true, because, up to a constant, the last integral is greater than

$$\int_{r_1 \geq -r_2, r_2 \leq 0} \exp\left(-|r| + \frac{(r_1 - r_2)}{\sqrt{2}}\right) dr_1dr_2 \geq \int_{r_1 \geq -r_2, r_2 \leq 0} \exp\left(-\sqrt{2}r_1 - \sqrt{2}r_2\right) dr_1dr_2 = \infty$$

by integrating with respect to r_1 and then r_2.

3.2 Sampling from $\mathcal{L}(\bar{Y}, \sigma)$

The current paragraph presents a general method for sampling from the Laplace distribution $\mathcal{L}(\bar{Y}, \sigma)$. This method relies in part on the transformation properties of $\mathcal{L}(\bar{Y}, \sigma)$, which are given in the following Proposition.

Proposition 3.2. *Let Y be a random variable in $\mathcal{P}_m$. For all $A \in \mathrm{GL}(m)$,*

$$Y \sim \mathcal{L}(\bar{Y}, \sigma) \implies Y \cdot A \sim \mathcal{L}(\bar{Y} \cdot A, \sigma)$$

where $Y \cdot A$ is given by (9). Moreover,

$$Y \sim \mathcal{L}(I, \sigma) \implies Y^{-1} \sim \mathcal{L}(I, \sigma)$$

The proof of this proposition follows that of proposition 5 in [13] and is omitted.

The following algorithm describes how to sample from $\mathcal{L}(\bar{Y}, \sigma)$ where $0 < \sigma < \sigma_m$. For this, it is first required to sample from the density p on $\mathbb{R}^m$ defined by

$$p(r) = \frac{c_m}{\zeta_m(\sigma)} e^{-\frac{|r|}{\sigma}} \prod_{i<j} \sinh\left(\frac{|r_i - r_j|}{2}\right), \quad r = (r_1, \cdots, r_m).$$

This can be done by a usual metropolis algorithm [20].

It is also required to sample from the uniform distribution on $\mathrm{O}(m)$, the group of real orthogonal $m \times m$ matrices. This can be done by generating A an $m \times m$ matrix, whose entries are i.i.d with normal distribution $\mathcal{N}(0, 1)$, then the orthogonal matrix U, in the decomposition $A = UT$ with T upper triangular, is uniformly distributed on $\mathrm{O}(m)$ [19] (p. 70). Sampling from $\mathcal{L}(\bar{Y}, \sigma)$ can now be described as follows.

1. Generate i.i.d samples $(r_1, \cdots, r_m) \in \mathbb{R}^m$ with density p.
2. Generate U from a uniform distribution on $\mathrm{O}(m)$.
3. $X \leftarrow U^\dagger \mathrm{diag}(e^{r_1}, \cdots, e^{r_m})U$.
4. $Y \leftarrow X.\bar{Y}^{\frac{1}{2}}$.

Note that the law of X in step 3 is $\mathcal{L}(I, \sigma)$; the proof of this fact is similar to Proposition 6 in [13]. Finally, the law of Y in step 4 is $\mathcal{L}(I.\bar{Y}^{\frac{1}{2}} = \bar{Y}, \sigma)$ by Proposition (3.2).

3.3 Estimation of $\bar{Y}$ and σ

The current paragraph considers maximum likelihood estimation of the parameters $\bar{Y}$ and σ, based on independent observations $Y_1, \ldots, Y_N$ from the Riemannian Laplace distribution $\mathcal{L}(\bar{Y}, \sigma)$. The main results are contained in Propositions 3.3 and 3.6 below.

Proposition 3.3 states the existence and uniqueness of the maximum likelihood estimates $\hat{Y}_N$ and $\hat{\sigma}_N$ of $\bar{Y}$ and σ. In particular, the maximum likelihood estimate $\hat{Y}_N$ of $\bar{Y}$ is the Riemannian median of $Y_1, \ldots, Y_N$, defined by (11). Numerical computation of $\hat{Y}_N$ will be considered and carried out using a Riemannian sub-gradient descent algorithm [8].

Proposition 3.6 states the convergence of the maximum likelihood estimate $\hat{Y}_N$ to the true value of the parameter $\bar{Y}$. It is based on Lemma 3.5 which states that the parameter $\bar{Y}$ is the Riemannian median of the distribution $\mathcal{L}(\bar{Y}, \sigma)$ in the sense of definition (12).

Proposition 3.3 (MLE and median)**.** *The maximum likelihood estimate of the parameter $\bar{Y}$ is the Riemannian median $\hat{Y}_N$ of $Y_1, \ldots, Y_N$. Moreover, the maximum likelihood estimate of the parameter σ is the solution $\hat{\sigma}_N$ of*

$$\sigma^2 \times \frac{d}{d\sigma} \log \zeta_m(\sigma) = \frac{1}{N} \sum_{n=1}^{N} d(\bar{Y}, Y_n) \tag{19}$$

Both $\hat{Y}_N$ and $\hat{\sigma}_N$ exist and are unique for any realisation of the samples $Y_1, \ldots, Y_N$.

Proof of Proposition 3.3. The log-likelihood function, of the parameters $\bar{Y}$ and σ, can be written as

$$\begin{aligned} \sum_{n=1}^{N} \log\, p(Y_n|\, \bar{Y}, \sigma) &= \sum_{n=1}^{N} \log \left(\frac{1}{\zeta_m(\sigma)} e^{-\frac{d(\bar{Y}, Y_n)}{\sigma}} \right) \\ &= -N \log \zeta_m(\sigma) - \frac{1}{\sigma} \sum_{n=1}^{N} d(\bar{Y}, Y_n) \end{aligned}$$

As the first term in the last expression does not contain $\bar{Y}$,

$$\text{argmax}_{\bar{Y}} \sum_{n=1}^{N} \log\, p(Y_n|\, \bar{Y}, \sigma) = \text{argmin}_{\bar{Y}} \sum_{n=1}^{N} d(\bar{Y}, Y_n)$$

The quantity on the right is exactly $\hat{Y}_N$ by (11). This proves the first claim. Now consider the function

$$F(\eta) = -N \log(\zeta_m(\frac{-1}{\eta})) + \eta \sum_{n=1}^{N} d(\hat{Y}_N, Y_n), \quad \eta < \frac{-1}{\sigma_m}$$

This function is strictly concave, since it is the logarithm of the moment generating function of a positive measure. Note that $\lim_{\eta \to \frac{-1}{\sigma_m}} F(\eta) = -\infty$ and admit for a moment that $\lim_{\eta \to -\infty} F(\eta) = -\infty$. By the strict concavity of F, there exists a unique $\hat{\eta}_N < \frac{-1}{\sigma_m}$ (which is the maximum of F) such that $F'(\hat{\eta}_N) = 0$. It follows that $\hat{\sigma}_N = \frac{-1}{\hat{\eta}_N}$ lies in $(0, \sigma_m)$ and satisfies (19). The uniqueness of $\hat{\sigma}_N$ is a consequence of the uniqueness of $\hat{\eta}_N$. Thus the proof is complete. Now it remains to check that $\lim_{\eta \to -\infty} F(\eta) = -\infty$ or just $\lim_{\sigma \to +\infty} \frac{1}{\sigma} \log(\zeta_m(\frac{1}{\sigma})) = 0$. Clearly

$$\prod_{i<j} \sinh\left(\frac{|r_i - r_j|}{2}\right) \leq A_m e^{B_m |r|}$$

where A_m and B_m are two constants only depending on m. Using this, it follows that

$$\frac{1}{\sigma} \log(\zeta_m(\frac{1}{\sigma})) \leq \frac{1}{\sigma} \log(c_m A_m) + \frac{1}{\sigma} \log\left(\int_{\mathbb{R}^m} \exp((-\sigma + B_m)|r|) dr_1 \cdots dr_m\right) \quad (20)$$

But for some constant C_m only depending on m,

$$\int_{\mathbb{R}^m} \exp((-\sigma + B_m)|r|) dr_1 \cdots dr_m = C_m \int_0^\infty \exp((-\sigma + B_m)u) u^{m-1} du$$

$$\leq (m-1)! C_m \int_0^\infty \exp((-\sigma + B_m + 1)u) du = \frac{(m-1)! C_m}{\sigma - B_m - 1}$$

Combining this bound and (20) yields $\lim_{\sigma \to +\infty} \frac{1}{\sigma} \log(\zeta_m(\frac{1}{\sigma})) = 0$. □

Remark 3.4. *Replacing F in the previous proof with $F(\eta) = -\log(\zeta_m(\frac{-1}{\eta})) + \eta c$ where $c > 0$ shows that the equation*

$$\sigma^2 \times \frac{d}{d\sigma} \log \zeta_m(\sigma) = c$$

has a unique solution $\sigma \in (0, \sigma_m)$. This shows in particular that $\sigma \mapsto \sigma^2 \times \frac{d}{d\sigma} \log \zeta_m(\sigma)$ is a bijection from $(0, \sigma_m)$ to $(0, \infty)$.

Consider now the numerical computation of the maximum likelihood estimates $\hat{Y}_N$ and $\hat{\sigma}_N$ given by Proposition 3.3. Computation of $\hat{Y}_N$ consists in finding the Riemannian median of $Y_1, \ldots, Y_N$, defined by (11). This can be done using the Riemannian sub-gradient descent algorithm of [8]. The k^{th} iteration of this algorithm produces an approximation $\hat{Y}_N^k$ of $\hat{Y}_N$ in the following way.

For $k = 1, 2, \ldots$, let Δ_k be the symmetric matrix

$$\Delta_k = \frac{1}{N} \sum_{n=1}^{N} \frac{\text{Log}_{\hat{Y}_N^{k-1}}(Y_n)}{||\text{Log}_{\hat{Y}_N^{k-1}}(Y_n)||} \quad (21)$$

Here Log is the Riemannian logarithm mapping inverse to the Riemannian exponential mapping

$$\mathrm{Exp}_Y(\Delta) = Y^{1/2}\,\exp\left(Y^{-1/2}\,\Delta\,Y^{-1/2}\right)\,Y^{1/2} \tag{22}$$

and $||\mathrm{Log}_a(b)|| = \sqrt{g_a(b,b)}$. Then, $\hat{Y}_N^k$ is defined to be

$$\hat{Y}_N^k = \mathrm{Exp}_{\hat{Y}_N^{k-1}}(\tau_k\,\Delta_k) \tag{23}$$

where $\tau_k > 0$ is a step size which can be determined using a backtracking procedure.

Computation of $\hat{\sigma}_N$ requires solving a non-linear equation in one variable. This is readily done using Newton's method.

It is shown now that the empirical Riemannian median $\hat{Y}_N$ converges almost surely to the true median $\bar{Y}$. This means that $\hat{Y}_N$ is a consistent estimator of $\bar{Y}$. The proof of this fact requires few notations and a preparatory lemma.

For $\bar{Y} \in \mathcal{P}_m$ and $\sigma \in (0, \sigma_m)$, let

$$\mathcal{E}(Y|\,\bar{Y},\sigma) = \int_{\mathcal{P}_m} d(Y,Z)\,p(Z|\,\bar{Y},\sigma)dv(Z)$$

The following lemma shows how to find $\bar{Y}$ and σ from the function $Y \mapsto \mathcal{E}(Y|\,\bar{Y},\sigma)$.

Lemma 3.5. *For any $\bar{Y} \in \mathcal{P}_m$ and $\sigma \in (0,\sigma_m)$, the following properties hold*
(i) $\bar{Y}$ is given by

$$\bar{Y} = \mathrm{argmin}_Y\,\mathcal{E}(Y|\,\bar{Y},\sigma) \tag{24a}$$

That is $\bar{Y}$ is the Riemannian median of $\mathcal{L}(\bar{Y},\sigma)$.
(ii) σ is given by

$$\sigma = \Phi\left(\mathcal{E}(\bar{Y}|\,\bar{Y},\sigma)\right) \tag{24b}$$

where the function Φ is the inverse function of $\sigma \mapsto \sigma^2 \times d\log\zeta_m(\sigma)/d\sigma$.

Proof. (i) Let $\mathcal{E}(Y) = \mathcal{E}(Y|\,\bar{Y},\sigma)$. According to Theorem 2.1 in [18], this function has a unique global minimum which is also a unique stationary point. Thus, to prove that $\bar{Y}$ is the minimum point of $\mathcal{E}$, it will suffice to check that for any geodesic γ starting from $\bar{Y}$, $\frac{d}{dt}|_{t=0}\mathcal{E}(\gamma(t)) = 0$ [21] (p. 76). Note that

$$\frac{d}{dt}|_{t=0}\mathcal{E}(\gamma(t)) = \int_{\mathcal{P}_m}\frac{d}{dt}|_{t=0}d(\gamma(t),Z)\,p(Z|\,\bar{Y},\sigma)dv(Z) \tag{25}$$

where for all $Z \neq \bar{Y}$ [22]

$$\frac{d}{dt}|_{t=0}d(\gamma(t),Z) = -g_{\bar{Y}}(\log_{\bar{Y}}(Z),\gamma'(0))d(\bar{Y},Z)^{-1}$$

The integral in (25) is, up to a constant,

$$\frac{d}{dt}|_{t=0} \int_{\mathcal{P}_m} p(Z|\,\gamma(t),\sigma)dv(Z) = 0$$

since $\int_{\mathcal{P}_m} p(Z|\,\gamma(t),\sigma)dv(Z) = 1$.

The proof of part (ii) is based on the same techniques as (i) by differentiating the last integral with respect to σ. □

Proposition 3.6 (Consistency of $\hat{Y}_N$)**.** *Let $Y_1, Y_2, \cdots$ be independent samples from a Laplace distribution $G(\bar{Y}, \sigma)$. The empirical median $\hat{Y}_N$ of $Y_1, \ldots, Y_N$ converges almost surely to $\bar{Y}$, as $N \to \infty$.*

Proof. Corollary 3.5 in [23] (p. 49) states that if (Y_n) is a sequence of i.i.d random variables on $\mathcal{P}_m$ with law π, then the Riemannian median $\hat{Y}_N$ of $Y_1, \cdots, Y_N$ converges almost surely as $N \to \infty$ to $\hat{Y}_\pi$, the Riemannian median of π defined by (12). Applying this result to $\pi = \mathcal{L}(\bar{Y}, \sigma)$ and using $\hat{Y}_\pi = \bar{Y}$ which follows from item (i) of Lemma 3.5 shows that $\hat{Y}_N$ converges almost surely to $\bar{Y}$. □

4 Mixtures of Laplace distributions

The present section introduces the class of probability distributions which are finite mixtures of Riemannian Laplace distributions on $\mathcal{P}_m$. These constitute the main theoretical tool, to be used for the target application of the present paper, namely the problem of texture classification in computer vision, which will be treated in Section 5.

A mixture of Riemannian Laplace distributions is a probability distribution on $\mathcal{P}_m$, whose density with respect to the Riemannian volume element (13) has the following expression

$$p(Y|(\varpi_\mu, \bar{Y}_\mu, \sigma_\mu)_{1\leq\mu\leq M}) = \sum_{\mu=1}^{M} \varpi_\mu \times p(Y|\,\bar{Y}_\mu, \sigma_\mu) \tag{26}$$

where ϖ_μ are nonzero weights, whose sum is equal to 1, $\bar{Y}_\mu \in \mathcal{P}_m$ and $\sigma_\mu \in (0, \sigma_m)$ for all $1 \leq \mu \leq M$ and the parameter M is called the number of mixture components.

Paragraph 4.1 describes a new EM (expectation maximisation) algorithm, which computes the maximum likelihood estimates of the mixture parameters $(\varpi_\mu, \bar{Y}_\mu, \sigma_\mu)_{1\leq\mu\leq M}$, based on independent observations $Y_1, \ldots, Y_N$ from the mixture distribution (26).

Paragraph 4.2 considers the problem of order selection for mixtures of Riemannian Laplace distributions. Precisely, this consists in finding the number M of mixture components in (26) which realises the best representation of a given set of data $Y_1, \ldots, Y_N$.

This problem is solved by computing the BIC criterion, which is here found in explicit form for the case of mixtures of Riemannian Laplace distributions on $\mathcal{P}_m$.

4.1 Estimation of the mixture parameters

In this section $Y_1, \dots, Y_N$ are i.i.d samples from (26). Based on these observations, an EM algorithm is proposed to estimate $(\varpi_\mu, \bar{Y}_\mu, \sigma_\mu)_{1\leq\mu\leq M}$. The derivation of this algorithm can be carried out similarly to [13].

To explain how this algorithm works, define for all $\vartheta = \{(\varpi_\mu, \bar{Y}_\mu, \sigma_\mu)\}$,

$$\omega_\mu(Y_n, \vartheta) = \frac{\varpi_\mu \times p(Y_n|\, \bar{Y}_\mu, \sigma_\mu)}{\sum_{s=1}^{M} \varpi_s \times p(Y_n|\, \bar{Y}_s, \sigma_s)}, \qquad N_\mu(\vartheta) = \sum_{n=1}^{N} \omega_\mu(Y_n) \tag{27}$$

The algorithm iteratively updates $\hat{\vartheta} = \{(\hat{\varpi}_\mu, \hat{Y}_\mu, \hat{\sigma}_\mu)\}$, which is an approximation of the maximum likelihood estimate of the mixture parameters $\vartheta = (\varpi_\mu, \bar{Y}_\mu, \sigma_\mu)$ as follows.

► Update for $\hat{\varpi}_\mu$: Based on the current value of $\hat{\vartheta}$, assign to $\hat{\varpi}_\mu$ the new value $\hat{\varpi}_\mu = N_\mu(\hat{\vartheta})\Big/N$.

► Update for $\hat{Y}_\mu$: Based on the current value of $\hat{\vartheta}$, assign to $\hat{Y}_\mu$ the value

$$\hat{Y}_\mu = \text{argmin}_Y \sum_{n=1}^{N} \omega_\mu(Y_n, \hat{\vartheta})\, d(Y, Y_n) \tag{28}$$

► Update for $\hat{\sigma}_\mu$: Based on the current value of $\hat{\vartheta}$, assign to $\hat{\sigma}_\mu$ the new value

$$\hat{\sigma}_\mu = \Phi\big(\, {N_\mu^{-1}}(\hat{\vartheta}) \times \textstyle\sum_{n=1}^{N} \omega_\mu(Y_n, \hat{\vartheta})\, d(\hat{Y}_\mu, Y_n)\big) \tag{29}$$

where the function Φ is defined in Proposition 3.5.

These three update rules should be performed in the above order. Realisation of the update rules for $\hat{\varpi}_\mu$ and $\hat{\sigma}_\mu$ is straightforward. The update rule for $\hat{Y}_\mu$ is realised using a slight modification of the sub-gradient descent algorithm described in paragraph 3.2. More precisely, the factor $1/N$ appearing in (21) is only replaced with $\omega_\mu(Y_n, \hat{\vartheta})$ at each iteration.

In practice, the initial conditions $(\hat{\varpi}_{\mu_0}, \hat{Y}_{\mu_0}, \hat{\sigma}_{\mu_0})$ in this algorithm were chosen in the following way. The weights (ϖ_{μ_0}) are uniform and equal to $1/M$, $(\hat{Y}_{\mu_0})$ are M different observations from the set $\{Y_1, .., Y_N\}$ chosen randomly and $(\hat{\sigma}_{\mu_0})$ is computed from (ϖ_{μ_0}) and $(\hat{Y}_{\mu_0})$ according to the rule (29). Since the convergence of the algorithm depends on the initial conditions, the EM algorithm is run several times and the best result is retained, i.e., the one maximizing the log-likelihood function.

4.2 The Bayesian Information Criterion

The BIC (Bayesian Information criterion) was introduced by Schwarz to find the appropriate dimension of a model that will fit a given set of observations [14]. Since then, BIC has been used in many Bayesian modeling problems where priors are hard to set precisely. In large-sample settings, the fitted model favored by BIC ideally corresponds to the candidate model which is a posteriori most probable; i.e., the model which is rendered most plausible by the data at hand. One of the main features of the BIC is its easy computation since it is only based on the empirical log-likelihood function.

Given a set of observations $\{Y_1, \cdots, Y_N\}$ arising from (26) where M is unknown, the BIC consists in choosing the parameter

$$\bar{M} = \text{argmax}_M BIC(M)$$

where

$$BIC(M) = LL - \frac{1}{2} \times DF \times \log(N) \tag{30}$$

Here LL is the log-likelihood given by

$$LL = \sum_{n=1}^{N} \log \left(\sum_{k=1}^{M} \hat{\varpi}_k p(Y_n | \hat{Y}_k, \hat{\sigma}_k) \right) \tag{31}$$

and DF is the number of degrees of freedom of the statistical model

$$DF = M \times \frac{m(m+1)}{2} + M + M - 1 \tag{32}$$

In formula (31), $(\hat{\varpi}_k, \hat{Y}_k, \hat{\sigma}_k)_{1 \leq k \leq M}$ are obtained from an EM algorithm as stated in paragraph 4.1 assuming the exact dimension is M. Finally note that in formula (32), $M \times \frac{m(m+1)}{2}$ (respectively M and $M-1$) corresponds to the number of degrees of freedom associated to $(\hat{Y}_k)_{1 \leq k \leq M}$ (respectively $(\hat{\sigma}_k)_{1 \leq k \leq M}$ and $(\hat{\varpi}_k)_{1 \leq k \leq M}$).

5 Application to classsification of data on $\mathcal{P}_m$

The present section considers the application of Riemannian Laplace distributions to the classification of data in $\mathcal{P}_m$. It gives an original Laplace classification rule, which can be used to carry out the task of classification, even in the presence of outliers. It also applies this classification rule to the problem of texture classification in computer vision, showing that it leads to improved results in comparison with recent literature.

Paragraph 5.1 considers, from the point of view of statistical learning, the classification of data with values in $\mathcal{P}_m$. Given data points $Y_1, \cdots, Y_N \in \mathcal{P}_m$, this proceeds

in two steps, called the learning phase and the classification phase respectively. The learning phase uncovers the class structure of the data, by estimating a mixture model using the EM algorithm developed in paragraph 4.1. Once training is accomplished, data points are subdivided into disjoint classes. Classification consists in associating each new data point to the most suitable class. For this, a new classification rule will be established and shown to be optimal.

Paragraph 5.2 is the implementation of the Laplace classification rule together with the BIC criterion to texture classification in computer vision. It highlights the advantage of the Laplace distribution in presence of outliers and shows its better performance compared with recent approaches.

5.1 Classification using mixtures of Laplace distributions

Assume given a set of training data $Y_1, \cdots, Y_N$. These are now modeled as a realisation of a mixture modeled as a realisation of a mixture of Laplace distributions

$$p(Y) = \sum_{\mu=1}^{M} \varpi_\mu \times p(Y|\, \bar{Y}_\mu, \sigma_\mu) \tag{33}$$

In this paragraph, the order M in (33) is considered as known. The training phase of this data consists in learning its structure as a family of M disjoint classes $C_\mu, \mu = 1, \cdots, M$. To be more precise, depending on the family (ϖ_μ), some of these classes may be empty. Training is done by applying the EM algorithm described in paragraph 4.1. As a result each class C_μ is represented by a triple $(\hat{\varpi}_\mu, \hat{Y}_\mu, \hat{\sigma}_\mu)$ corresponding to maximum likelihood estimates of $(\varpi_\mu, Y_\mu, \sigma_\mu)$. Each observation Y_n is now associated to the class C_{μ^*} where $\mu^* = \text{argmax}_\mu \omega(Y_n, \hat{\nu})$ (recall the definition from (27)). In this way $\{Y_1, \cdots, Y_N\}$ is subdivided into M disjoint classes.

The classification phase requires a classification rule. Following [13], the optimal rule (in the sense of a Bayesian risk criterion given in [24]) consists in associating any new data Y_t, to the class C_{μ^*} where

$$\mu^* = \text{argmax}_\mu \left\{ \hat{N}_\mu \times p(Y_t|\hat{Y}_\mu, \hat{\sigma}_\mu) \right\} \tag{34}$$

Here $\hat{N}_\mu$ is the number of elements in C_μ. Replacing $\hat{N}_\mu$ with $N \times \hat{\varpi}_\mu$, (34) becomes $\text{argmax}_\mu\, \hat{\varpi}_\mu \times p(Y_t|\hat{Y}_\mu, \hat{\sigma}_\mu)$. Note that when the weights ϖ_μ in (33) are assumed to be equal, this rule reduces to maximum likelihood classification rule $\max_\mu\, p(Y_t|\hat{Y}_\mu, \hat{\sigma}_\mu)$. A quick look at the expression (17) shows that (34) can also be expressed as

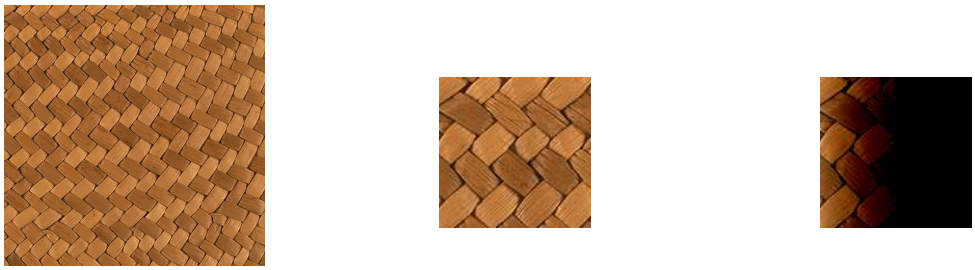
$$\mu^* = \text{argmin}_\mu \left\{ -\log\, \hat{\varpi}_\mu + \log\, \zeta(\hat{\sigma}_\mu) + \frac{d(Y_t\,, \hat{Y}_\mu)}{\hat{\sigma}_\mu} \right\} \tag{35}$$

The rule (35) will be called Laplace classification rule. It favours clusters C_μ having a larger number of data points (the minumum contains $-\log\,\widehat{\varpi}_\mu$), or a smaller dispersion away from the median (the minimum contains $\log\,\zeta(\hat{\sigma}_\mu)$). When choosing between two clusters with the same number of points and the same dispersion, this rule favours the one whose median is closer to Y_t . If the number of data points inside clusters and the respective dispersions are neglected, then (35) reduces to the nearest neighbor rule involving only the Riemannian distance introduced in [2].

5.2 Application to texture classification

This paragraph presents an application of the mixture of Laplace distributions to the context of texture classification on the MIT Vision Texture (VisTex) database [25]. The purpose of this experiment is to classify the textures, by taking into consideration the within-class diversity. In addition, the influence of outliers on the classification performances is analyzed. The obtained results for the Riemannian Laplace distribution (RLD) are compared to those given by the Riemannian Gaussian distribution (RGD) [13].

The VisTex database contains 40 images, considered as being 40 different texture classes. The database used for the experiment is obtained after several steps. First of all, each texture is decomposed in 169 patches of 128×128 pixels, with an overlap of 32 pixels, giving a total number of 6760 textured patches. Next, some patches are corrupted, in order to introduce abnormal data into the dataset. Therefore, their intensity is modified by applying a gradient of luminosity. For each class, between 0 and 60 patches are modified in order to become outliers. An example of a VisTex texture with one of its patches and an outlier patch are shown in Figure 1.



(a) (b) (c)

Figure 1: *Example of a texture from the VisTex database (a), one of its patches (b) and the corresponding outlier (c).*

Once the database is built, it is 15 times equally and randomly divided in order to obtain the training and the testing sets that are further used in the supervised classification algorithm. Then, for each patch in both databases a feature vector has to be computed. The luminance channel is first extracted, and then normalised in intensity. The grayscale patches are filtered using the stationary wavelet transform Daubechies db4 filter, with 2 scales and 3 orientations. To model the wavelet subbands, various stochastic models have been proposed in the literature. Among them, the univariate generalized Gaussian distribution has been found to accurately model the empirical histogram of wavelet subbands [26]. Recently, it has been proposed to model the spatial dependency of wavelet coefficients. To this aim, the wavelet coefficients located in a $p \times q$ spatial neighborhood of the current spatial position are clustered in a random vector. The realisations of these vectors can be further modeled by elliptical distributions [27, 28], copula based models [30, 29], etc. In this paper, the wavelet coefficients are considered as being realisations of zero-mean multivariate Gaussian distributions. In addition, for this experiment the spatial information is captured by using a vertical (2×1) and a horizontal (1×2) neighborhood. Next, the 2×2 sample covariance matrices are estimated for each wavelet subband and each neighborhood. Finally, each patch is represented by a set of $F = 12$ covariance matrices (2 scales $\times$ 3 orientations $\times$ 2 neighborhoods) denoted $Y = [Y_1, \cdots, Y_F]$.

The estimated covariance matrices are elements of $\mathcal{P}_m$, with $m = 2$ and therefore they can be modeled by Riemannian Laplace distributions. More precisely, in order to take into consideration the within-class diversity, each class in the training set is viewed as a realisation of a mixture of Riemannian Laplace distributions (equation (26)) with M mixture components, characterized by $(\varpi_\mu, \bar{Y}_{\mu,f}, \sigma_{\mu,f})$, having $\bar{Y}_{\mu,f} \in \mathcal{P}_2$, with $\mu = 1, \cdots, M$ and $f = 1, \cdots, F$. Since the subbands are assumed to be independent, the probability density function is given by:

$$p(Y|(\varpi_\mu, \bar{Y}_{\mu,f}, \sigma_{\mu,f})_{1\leq\mu\leq M, 1\leq f\leq F}) = \sum_{\mu=1}^{M} \varpi_\mu \prod_{f=1}^{F} p(Y_f|\, \bar{Y}_{\mu,f}, \sigma_{\mu,f}). \tag{36}$$

The learning step of the classification is performed using the EM algorithm presented in paragraph 4 and the number of mixture components is determined using the BIC criterion recalled in (30). Note that for the considered model given in (36), the degree of freedom is expressed as

$$DF = M - 1 + M \times F \times \left(\frac{m(m+1)}{2} + 1\right) \tag{37}$$

since one centroid and one dispersion parameter should be estimated per feature and per component of the mixture model. In practice, the number of mixture components M varies between 2 and 5, and the M yielding to the highest BIC criterion is retained.

As mentioned earlier, the EM algorithm is sensitive to the initial conditions. In order to minimize this influence, for this experiment the EM algorithm is repeated 10 times and the result maximizing the log-likelihood function is retained. Finally, the classification is performed by assigning each element $Y_t \in \mathcal{P}_2$ in the testing set to the class of the closest cluster μ^*, given by

$$\mu^* = \text{argmin}_\mu \left\{ -\log\, \hat{\varpi}_\mu + \sum_{f=1}^{F} \log\, \zeta(\hat{\sigma}_{\mu,f}) + \sum_{f=1}^{F} \frac{d(Y_t\,, \hat{Y}_{\mu,f})}{\hat{\sigma}_{\mu,f}} \right\} \tag{38}$$

This expression is obtained starting from (35) and (36), knowing that F features are extracted for each patch.

The classification results of the proposed model (solid red line), expressed in terms of overall accuracy, shown in Figure 2, are compared with those given by a fixed number of mixture components (that is for $M = 3$, dashed red line) and with those given when the within-class diversity is not considered (that is for $M = 1$, dotted red line). In addition, the classification performances given by the RGD model (displayed in black) proposed in [13] are also considered. For this model, the number of mixture components is first computed using the BIC, and next it is fixed to $M = 3$ and $M = 1$. For all the considered methods, the classification rate is given as a function of the number of outliers, that varies between 0 and 60 for each class.

First, the influence of abnormal data on the RGD and RLD models is analyzed as the number of outlier patches increases. It is shown that the RLD gives progressively better results than the RGD. Second, the number of mixture components is considered. It can be noticed that the results are improved by using mixture distributions joint with the BIC criterion for choosing the suitable number of clusters. In conclusion, the mixture of RLDs combined with the BIC criterion to estimate the best number of mixtures components can minimize the influence of abnormal samples present in the dataset, illustrating the relevance of the proposed method.

6 Conclusions

Motivated by the problem of outliers in statistical data, this paper introduces a new distribution on the space $\mathcal{P}_m$ of $m \times m$ symmetric positive definite matrices called Riemannian Laplace distribution. Denoted throughout the paper by $\mathcal{L}(\bar{Y}, \sigma)$, where $\bar{Y} \in \mathcal{P}_m$ and $\sigma > 0$ are the indexing parameters, this distribution may be thought of as specifying the law of a family of observations on $\mathcal{P}_m$ concentrated around the location $\bar{Y}$ and having dispersion σ. If d denotes Rao's distance on $\mathcal{P}_m$ and $dv(Y)$ its associated volum form, the density of $\mathcal{L}(\bar{Y}, \sigma)$ with respect to $dv(Y)$ is proportional to $\exp(-\frac{d(Y,\bar{Y}}{\sigma}))$. Interestingly, the normalising constant depends only on σ (and not

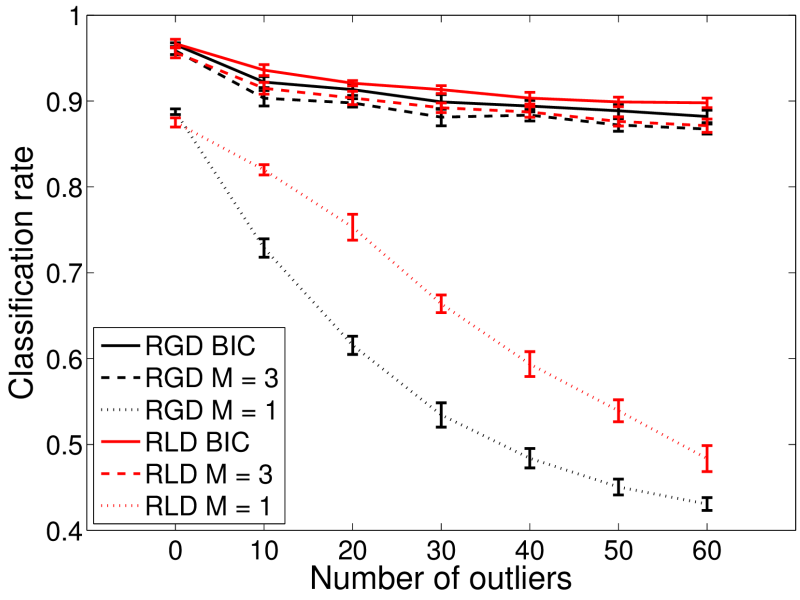


Figure 2: *Classification results.*

on $\bar{Y}$). This allows us to deduce exact expressions for maximum likelihood estimates of $\bar{Y}$ and σ relying on the Riemannian median on $\mathcal{P}_m$. These estimates are also computed numerically by means of sub-gradient algorithms. Estimation of parameters in mixture models of Laplace distributions are also considered and performed using a new expectation-maximisation algorithm. Finally, the main theoretical results are illustrated by an application to texture classification. The proposed experiment consists in introducing abnormal data (outliers) into a set of images from the Vistex database and analyzing their influences on the classification performances. Each image is characterized by a set of 2×2 covariance matrices modeled as mixtures of Riemannian Laplace distributions in the space $\mathcal{P}_2$. The number of mixtures, is estimated using the BIC criterion. The obtained results are compared with those given by the Riemannian Gaussian distribution, showing the better performance of the proposed method.

Acknowledgments: 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] Pennec, X.; Fillard, P.; Ayache, N. A Riemannian framework for tensor computing. Int. J. Comput. Vision, 2006, v. 66, n. 1, 41-66.

[2] Barachant, A.; Bonnet, S.; Congedo, M.; Jutten, C. Multiclass Brain–Computer Interface Classification by Riemannian Geometry. IEEE Trans. Biomed. Eng, 2012, v. 59, n. 4, 920-928.

[3] Jayasumana, S.; Hartley, R.; Salzmann, M.; Li, H.; Harandi, M. Kernel Methods on the Riemannian Manifold of Symmetric Positive Definite Matrices. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, 73-80.

[4] Zheng, L.; Qiu, G.; Huang, J.; Duan, J. Fast and accurate Nearest Neighbor search in the manifolds of symmetric positive definite matrices. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, 3804-3808.

[5] Dong, G.; Kuang, G. Target recognition in SAR images via classification on Riemannian manifolds. IEEE Geoscie. Remote Sens. Lett. 2015, volume 21, n. 1, 199-203.

[6] Tuzel, O.; Porikli, F.; Meer, P. Pedestrian detection via classification on Riemannian manifolds. IEEE Trans. Pattern Anal. and Mach. Intell. 2008, v. 30, n. 10, 1713-1727.

[7] Caseiro, R.; Henriques, J. F.; Martins, P.; Batista, J. A nonparametric Riemannian framework on tensor field with application to foreground segmentation. Pattern Recognition, 2012, v. 45, n. 11, 3997 - 4017.

[8] Arnaudon, M.; Barbaresco, F.; Yang, L. Riemannian Medians and Means With Applications to Radar Signal Processing. IEEE J. Sel. Topics Signal Process, 2013, v. 7, n. 4, 595-604.

[9] Arnaudon, M.; Yang, L.; Barbaresco, F. Stochastic algorithms for computing p-means of probability measures, Geometry of Radar Toeplitz covariance matrices and applications to HR Doppler processing. International Radar Symposium (IRS), 2011, 651-656.

[10] Terras, A. Harmonic analysis on symmetric spaces and applications, Vol. II. Springer-Verlag, New York, 1988.

[11] Atkinson, C.; Mitchell, A. Rao's distance measure. *Sankhya Ser. A*, 1981, 43, 345-365.

[12] Guang, C.; Baba C. V. A Novel Dynamic System in the Space of SPD Matrices with Applications to Appearance Tracking. SIAM J. Imaging Sciences, 2013, v. 6, n. 1, 592–615.

[13] Said, S.; Bombrun, L.; Berthoumieu, Y.; H. Manton, J. Riemannian Gaussian distributions on the space of symmetric positive definite matrices. Available on arxiv via http://arxiv.org/abs/1507.01760, 2015.

[14] Schwarz, G. Estimating the Dimension of a Model. Ann. Stat. 1978, v. 6, n. 2, 461–464.

[15] Maass, H. Siegel's modular forms and Dirichlet series. Lecture Notes in Mathematics, Vol. 216. Springer-Verlag, Berlin-New York, 1971.

[16] Higham, N. J. Functions of matrices, Theory and computation. Society for Industrial and Applied Mathematics, 2008.

[17] Helgason, S. Differential geometry, Lie groups, and symmetric spaces. American Mathematical Society, 2001.

[18] Afsari, B. Riemannian L^p center of mass: existence, uniqueness and convexity. Proc. Amer. Math. Soc. 2011, v. 139, n. 2, 655-673.

[19] Muirhead, R. J. Aspects of multivariate statistical theory. John Wiley & Sons, New York, 1982.

[20] Robert, C. P.; Casella, G. Monte Carlo Statistical Methods. Springer-Verlag, 2004.

[21] Udriste, C. Convex functions and optimization methods on Riemannian manifolds. Mathematics and its applications, Kluwer Academic Publishers, 1994.

[22] Chavel, I. Riemannian geometry, a modern introduction. Cambridge University Press, 2006.

[23] L. Yang. Médianes de mesures de probabilité dans les variétés riemanniennes et applications à la détection de cibles radar. Thèse de doctorat de l'université de Poitiers, 2011.

[24] Hastie, T.; Tibshirani, R.; Friedman, J. The elements of statistical learning: data mining, inference, and prediction. Springer, 2009.

[25] VisTex: Vision Texture Database. MIT Media Lab Vision and Modeling group, http://vismod.media.mit.edu/pub/, 1995.

[26] Do, M.N.; Vetterli, M. Wavelet-Based Texture Retrieval Using Generalized Gaussian Density and Kullback-Leibler Distance. IEEE Trans. on Image Process. 2002, v. 11, 146–158.

[27] Bombrun, L.; Berthoumieu, Y.; Lasmar, N. E.; Verdoolaege, G. Mutlivariate Texture Retrieval Using the Geodesic Distance Between Elliptically Distributed Random Variables. EEE ICIP, 2011, 3637-3640.

[28] Verdoolaege, G.; Scheunders, P. On the Geometry of Multivariate Generalized Gaussian Models. Journal of Mathematical Imaging and Vision, 2012, v. 43, n. 3, 180-193.

[29] Stitou, Y.; Lasmar, N. E.; Berthoumieu, Y. Copulas based Multivariate Gamma Modeling for Texture Classification. IEEE International Conference on Acoustic Speech and Signal Processing, 2009, 1045-1048.

[30] Kwitt, R.; Uhl, A. Lightweight Probabilistic Texture Retrieval. IEEE Trans. on Image Process. 2010, v. 19, n. 1, 241-253.