ON THE USE OF GEODESIC TRIANGLES BETWEEN GAUSSIAN DISTRIBUTIONS FOR CLASSIFICATION PROBLEMS

A. Collas[1], *F. Bouchard*[2], *G. Ginolhac*[3], *A. Breloy*[4], *C. Ren*[1], *J.-P. Ovarlez*[1,5]

[1]SONDRA, CentraleSupélec, Université Paris-Saclay
[2]CNRS, L2S, CentraleSupélec, Université Paris-Saclay
[3]LISTIC, Université Savoie Mont Blanc
[4]LEME, Université Paris Nanterre
[5]DEMR, ONERA, Université Paris-Saclay

1. INTRODUCTION

This paper presents a new classification framework for both first and second order statistics, *i.e* mean/location and covariance matrix. In the last decade, several covariance matrix classification algorithms leveraging the Riemannian geometry of symmetric positive definite matrices have been developed. However, their underlying statistical model assumes a zero mean. This is of course damaging for applications where the mean is a discriminative feature. Unfortunately, the distance associated to the Riemannian geometry for both mean and covariance matrix remains unknown. Leveraging previous works on geodesic triangles, we propose two affine invariant divergences that use both statistics. Then, the associated Riemannian center of mass can be computed using optimization on Riemannian manifolds. Finally, a divergence based *Nearest centroid classifier*, applied on the crop classification dataset *Breizhcrops* [1] (see Figure 1), shows the interest of the proposed framework. For the sake of brevity, only the main ideas and results are presented.

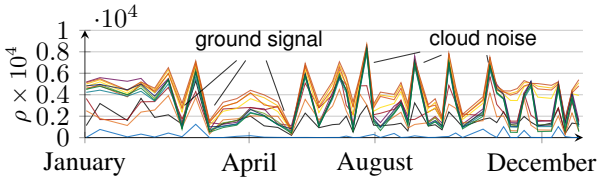

Fig. 1. A time series of meadows from the *Breizhcrops* dataset.

2. INFORMATION GEOMETRY OF THE MULTIVARIATE GAUSSIAN DISTRIBUTION

Let a set of n data points $\boldsymbol{x}_i \in \mathbb{R}^p$ sampled from a random variable $\boldsymbol{x}$ following a Gaussian distribution

$$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{1}$$

The parameters $\boldsymbol{\mu} \in \mathbb{R}^p$ and $\boldsymbol{\Sigma} \in \mathcal{S}_p^{++}$ (set of symmetric positive definite matrices) are the location and covariance matrix respectively. The maximum likelihood estimators are the well known sample mean and sample covariance matrix (SCM)

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i, \quad \hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})^T. \tag{2}$$

Then, $\mathcal{N}^p$ is turned into a Riemannian manifold. The tangent space $T_\upsilon\mathcal{N}^p$ of $\mathcal{N}^p$ at υ is identified to the product space $\mathbb{R}^p \times \mathcal{S}_p$ with $\mathcal{S}_p$ the set of symmetric matrices. Moreover, $\mathcal{N}^p$ is equipped with the Fisher information metric (FIM). Let $\xi = (\boldsymbol{\xi}_{\boldsymbol{\mu}}, \boldsymbol{\xi}_{\boldsymbol{\Sigma}}), \eta = (\boldsymbol{\eta}_{\boldsymbol{\mu}}, \boldsymbol{\eta}_{\boldsymbol{\Sigma}}) \in T_\upsilon\mathcal{N}^p$, this metric writes

$$\langle \xi, \eta \rangle_\upsilon^{\mathcal{N}^p} = \boldsymbol{\xi}_{\boldsymbol{\mu}}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\eta}_{\boldsymbol{\mu}} + \frac{1}{2}\operatorname{Tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\xi}_{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}_{\boldsymbol{\Sigma}}). \tag{3}$$

Remarkably, it is invariant under affine transformations. Given $\boldsymbol{A} \in \mathbb{R}^{p\times p}$ invertible and $\boldsymbol{\mu}_0 \in \mathbb{R}^p$, we verify that

$$\langle \mathrm{D}\,\phi(\upsilon)[\xi], \mathrm{D}\,\phi(\upsilon)[\eta] \rangle_{\phi(\upsilon)}^{\mathcal{N}^p} = \langle \xi, \eta \rangle_\upsilon^{\mathcal{N}^p}, \tag{4}$$

where the affine transformation writes $\phi(\upsilon) = (\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{\mu}_0, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$ with $\mathrm{D}\,\phi(\upsilon)[\xi]$ being the directional derivative of ϕ at υ in the direction ξ. Unfortunately, the geodesic between $\upsilon_1 = (\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\upsilon_2 = (\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ is unknown in general. However, geodesic triangles can be established which is enough to do *machine learning* on $\mathcal{N}^p$.

3. DIVERGENCES

Geodesic triangles between υ_1 and υ_2 can be derived. Indeed, by carefully choosing intermediate points υ, geodesics are obtained between υ_1 and υ and then between υ and υ_2. Hence, we get geodesic triangles $\upsilon_1 \to \upsilon \to \upsilon_2$. The squared arclength of one of these geodesic triangles is then measured to get a divergence denoted $\delta^2_{\mathcal{N}^p}$. By construction, it is invariant by affine transformation,

$$\delta^2_{\mathcal{N}^p}(\phi(\upsilon_1), \phi(\upsilon_2)) = \delta^2_{\mathcal{N}^p}(\upsilon_1, \upsilon_2). \tag{5}$$

Two divergences are proposed in Corollaries 1 and 2.

Corollary 1 (Divergence $\delta^2_{c,\mathcal{N}^p}$)**.** *A separable and invariant by affine transformation divergence on $\mathcal{N}^p$ is*

$$\delta^2_{c,\mathcal{N}^p}(\upsilon_1,\upsilon_2) = 2\,\mathrm{acosh}\left(\frac{c^{-\frac{1}{2}}}{2}\left(c+1+\frac{1}{2}\Delta\boldsymbol{\mu}^T\boldsymbol{\Sigma}_1^{-1}\Delta\boldsymbol{\mu}\right)\right)^2 + \frac{(p-1)}{2}\log(c)^2 + \frac{1}{2}\left\|\log\left(c\boldsymbol{\Sigma}_2^{-\frac{1}{2}}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-\frac{1}{2}}\right)\right\|_2^2.$$

where $\Delta\boldsymbol{\mu} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$ and $c = \left|\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2\right|^{\frac{1}{p}}$.

Corollary 2 (Divergence $\delta^2_{\perp,\mathcal{N}^p}$)**.** *A separable and invariant by affine transformation divergence on $\mathcal{N}^p$ is*

$$\delta^2_{\perp,\mathcal{N}^p}(\upsilon_1,\upsilon_2) = \frac{1}{2}\left[\mathrm{acosh}\left(1+\Delta\boldsymbol{\mu}^T\boldsymbol{\Sigma}_1^{-1}\Delta\boldsymbol{\mu}\right)^2 + \left\|\log\left(\boldsymbol{\Sigma}_2^{-\frac{1}{2}}\left(\boldsymbol{\Sigma}_1+\frac{1}{2}\Delta\boldsymbol{\mu}\Delta\boldsymbol{\mu}^T\right)\boldsymbol{\Sigma}_2^{-\frac{1}{2}}\right)\right\|_2^2\right].$$

4. ESTIMATION OF CENTERS OF MASS

Some important algorithms in machine learning require the computation the center of mass of a set of points $S = \{\upsilon_i\}_{i=1}^M \subset \mathcal{N}^p$. This center is associated to a proximity measure which in our case is one of the divergences, $\delta^2_{\mathcal{N}^p}$, defined in Section 3. The Riemannian center of mass $\upsilon^\star$ is defined as the minimizer of the variance of S

$$\upsilon^\star = \underset{\upsilon\in\mathcal{N}^p}{\arg\min}\ \frac{1}{2M}\sum_{i=1}^M \delta^2_{\mathcal{N}^p}(\upsilon,\upsilon_i). \tag{6}$$

This cost function can be optimized using gradient based algorithms on $\mathcal{N}^p$.

5. APPLICATION

In this Section, we provide an application of the divergences presented earlier on the large-scale satellite image time series dataset for crop type mapping called *Breizhcrops* [1].

More specifically, for each crop $n = 45$ observations $\boldsymbol{x}_i \in \mathbb{R}^p$ are measured over time. Each $\boldsymbol{x}_i$ contains measurements of reflectance of $p = 13$ spectral bands. Then, these measurements are concatenated into one batch $\boldsymbol{X}_j = [\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n] \in \mathbb{R}^{p\times n}$. Hence, we get one matrix $\boldsymbol{X}_j$ per crop and each one belongs to an unknown class $y \in [\![1,K]\!]$. These $K = 9$ classes represent crop types such as nuts, barley or wheat. The data are divided into a training set and a test set with 485 649 and 122 614 batches respectively. All data are centered using the global mean. For simplicity, the matrix $\boldsymbol{X}_j$ is denoted $\boldsymbol{X}$ in the following.

To classify these crops, we apply a *Nearest centroid classifier* algorithm on descriptors. This classification algorithm works in three steps: *(i)* For each batch $\boldsymbol{X}$, a descriptor is computed (*e.g* the sample mean or the SCM (2)). *(ii)* Then, on the training set, the center of mass of the descriptors of each class is computed. *(iii)* Finally, on the test set, each descriptor is associated to the nearest center of mass. Thus, we get a classification of the $\boldsymbol{X}$. The different descriptors used in the application are the following. The first two descriptors are the batches themselves $\boldsymbol{X}$ and their sample means $\hat{\boldsymbol{\mu}}$ (2). Their associated geometry is the Euclidean one with the Frobenius distance. Thus, the center of mass is the classical element-wise arithmetic mean. Then, the next two estimators are the SCMs $\hat{\boldsymbol{\Sigma}}$ (2) with location assumed to be known or not. In the case of known location, the SCM is simply estimated as $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i{}^T$. The associated Riemannian geometry is $\mathcal{S}_p^{++}$. Finally, the last two descriptors use both sample mean and SCM, $(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}})$ from (2). These estimators are used with the geometry $\mathcal{N}^p$ and the two divergences $\delta^2_{c,\mathcal{N}^p}$ and $\delta^2_{\perp,\mathcal{N}^p}$ presented in Corollaries 1 and 2 respectively.

Estimator of $\boldsymbol{X}$	Geometry	OA (%)	AA (%)
$\boldsymbol{X}$	$\mathbb{R}^{p\times n}$	10.1	18.5
$\hat{\boldsymbol{\mu}}$	$\mathbb{R}^p$	13.2	14.8
$\hat{\boldsymbol{\Sigma}}$, ($\boldsymbol{\mu}$ known)	$\mathcal{S}_p^{++}$	43.9	28.1
$\hat{\boldsymbol{\Sigma}}$, ($\boldsymbol{\mu}$ unknown)	$\mathcal{S}_p^{++}$	46.7	30.1
$(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}})$	$\mathcal{N}^p$ with $\delta^2_{c,\mathcal{N}^p}$	**54.3**	**37.0**
$(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}})$	$\mathcal{N}^p$ with $\delta^2_{\perp,\mathcal{N}^p}$	53.3	35.7

Table 1. Performance of the different estimators and Riemannian geometries on the *Breizhcrops* dataset [1]. OA = Overall Accuracy, AA = Average Accuracy.

Table 1 presents the Overall Accuracy and Average Accuracy of the different descriptors and geometries used in the *Nearest centroid classifier*. Estimators using $\hat{\boldsymbol{\Sigma}}$ along with the FIM clearly outperform the others. Also, the three estimators assuming $\boldsymbol{\mu}$ is unknown perform better than the others. This shows the interest of not considering $\boldsymbol{\mu} = \mathbf{0}$ for such applications, even if the global mean has been subtracted in a preprocessing step. Finally, using the divergences proposed in Corollaries 1 and 2 with their Riemannian centers of mass greatly improves both Overall Accuracy and Average Accuracy. These results confirm the interest of geodesic triangles when the distance associated to the FIM is not available in closed form.

6. REFERENCES

[1] M. Rußwurm, C. Pelletier, M. Zollner, S. Lefèvre, and M. Körner, "Breizhcrops: A time series dataset for crop type mapping," *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences ISPRS (2020)*, 2020.