Non supervised classification in the space of SPD matrices

Salem Said – Lionel Bombrun – Yannick Berthoumieu

Laboratoire IMS CNRS UMR 5218 – Université de Bordeaux

29 October 2015
Context of our work

Our project: *Statistical learning in the space of SPD matrices*

Our team: 3 members of IMS laboratory + 2 post docs (Hatem Hajri, Paolo Zanini)

Target applications: remote sensing, radar signal processing, Neuroscience (BCI)

Our partners: IMB (Marc Arnaudon + PhD student), Gipsa-lab, Ecole des Mines

Our recent work

- Riemannian Gaussian distributions on the space of SPD matrices (in review, IEEE IT)

Some of our problems:

Given a population of SPD matrices (any size or structure)

- Non-supervised learning of its class structure
- Semi-parametric learning of its density

Please look up our paper on Arxiv :-)

http://arxiv.org/abs/1507.01760
Geometric tools

Statistical manifold: \( \Theta = \text{SPD, Toeplitz, Block-Toeplitz, etc, matrices} \)

Hessian or Fisher metric:

\[
ds^2(\theta) = \text{Hess}_\Phi (d\theta, d\theta) \quad \Phi \text{ model entropy}
\]

\( \Theta \) becomes a Riemannian homogeneous space of negative curvature‼

Example: 2 \times 2 correlation (baby Toeplitz)

\[
\Theta = \left\{ \begin{pmatrix} 1 & \theta \\ \theta^* & 1 \end{pmatrix} \bigg| |\theta| < 1 \right\} \quad \Phi(\theta) = -\log[1 - |\theta|^2]
\]

\[
\Rightarrow ds^2(\theta) = \frac{|d\theta|^2}{[1 - |\theta|^2]^2} \quad \text{Poincaré disc model}
\]

Why do we use this?

– Suitable mathematical properties
– Relation to entropy or “information”
– Often leads to excellent performance

First place in IEEE BCI challenge
Contribution

I-Introduction of Riemannian Gaussian distributions

A statistical model of a class/cluster:

\[ p(\theta | \bar{\theta}, \sigma) = \frac{Z^{-1}(\sigma)}{Z(\sigma)} \times \exp \left[ -\frac{d^2(\theta, \bar{\theta})}{2\sigma^2} \right] \]

Expression unknown in the literature

\( d(\theta, \bar{\theta}) \) Riemannian distance

Computing \( Z(\sigma) \)

Case where \( \Theta \) is the space of \( m \times m \) covariance matrices

\[ Z(\sigma) = \int_{\Theta} \exp \left[ -\frac{d^2(\theta, \bar{\theta})}{2\sigma^2} \right] d\nu(\theta) \]

\[ d^2(\theta, \bar{\theta}) = \text{tr} \left[ \log \left( \theta^{-1} \bar{\theta} \right) \right]^2 \]

\[ d\nu(\theta) = \det(\theta)^{-\frac{m+1}{2}} \prod_{i<j} d\theta_{ij} \]

We have numerical tables of \( Z(\sigma) \)

10 \times 10 matrices

20 \times 20 matrices
Contribution

II – Statistical significance of the centre of mass

A statistical model of a class/cluster:

\[ C = \{\theta_1, \ldots, \theta_N\} \implies \hat{\theta}, \hat{\sigma} \quad (\text{estimates of parameters } \bar{\theta}, \sigma) \]

Log-likelihood function

\[
L(\theta, \sigma) = -N \log Z(\sigma) - \frac{1}{2\sigma^2} \sum_{n=1}^{N} d^2(\theta_n, \theta)
\]

Computing the estimates:

\[
\hat{\theta} = \arg\min_{\theta} \sum_{n=1}^{N} d^2(\theta_n, \theta) \implies \text{centre of mass of cluster } C
\]

\[
\hat{\sigma} = F(\sum_{n=1}^{N} d^2(\theta_n, \hat{\theta})) \quad F \text{ strictly increasing}
\]

\[\implies \text{hypothesis tests, confidence regions, etc, for the centre of mass}\]
Contribution

III – A new aproach to non-supervised classification

Estimating mixtures of Riemannian Gaussian distributions:

Mixture model
\[ p(\theta) = \sum_{k=1}^{K} \omega_k \times p(\theta | \bar{\theta}_k, \sigma_k) \]

EM algorithm

- E step : compute conditional weights \( \lambda_k(n) \)
- M step :
  \[ \hat{\theta}_k = \arg\min_{\theta} \sum_{n=1}^{N} \lambda_k(n) \times d^2(\theta_n, \theta) \]
  \( \hat{\sigma}_k = F(\ldots) \)

New classification rule:
Choose class \( C_k \) in order to minimise

\[ (* ) \quad - \log \hat{\omega}_k + \log Z(\hat{\sigma}_k) + \frac{1}{2\sigma^2_k} d^2(\theta_t, \hat{\theta}_k) \]

\( \Rightarrow \) Only the third term is used in the literature
Numerical experiment

Texture classification in the VisTex image database

Classification performance

<table>
<thead>
<tr>
<th>Classification rule</th>
<th>Overall performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>New rule (*)</td>
<td>94.3 ± 0.4 %</td>
</tr>
<tr>
<td>Nearest centre of mass</td>
<td>92.1 ± 0.5 %</td>
</tr>
<tr>
<td>Wishart classifier</td>
<td>89.7 ± 0.8 %</td>
</tr>
</tbody>
</table>

Improvement in comparison with state of the art
2 × 2 SPD matrices

From SPD matrices to Poincaré disc:

\[
2 \times 2 \text{ SPD matrix } \quad Y = e^t \times \begin{pmatrix}
z + x & y \\
y & z - x
\end{pmatrix},
\]

\[
t = \log \det(Y)
\]

\[
z^2 - y^2 - x^2 = 1
\]

equation of a hyperboloid

Stereographic projection

\[
\begin{align*}
z &= \cosh r \\
y &= \sinh r \sin \varphi \\
x &= \sinh r \cos \varphi
\end{align*}
\]

\[
\implies \theta = \tanh \frac{r}{2} \times e^{i\varphi}
\]

Meaning of this construction:

spectral decomposition

\[
Y = \begin{pmatrix}
\cos \frac{\varphi}{2} & \sin \frac{\varphi}{2} \\
-\sin \frac{\varphi}{2} & \cos \frac{\varphi}{2}
\end{pmatrix} \begin{pmatrix}
e^{t+r} & \\
e^{t-r}
\end{pmatrix} \begin{pmatrix}
\cos \frac{\varphi}{2} & -\sin \frac{\varphi}{2} \\
\sin \frac{\varphi}{2} & \cos \frac{\varphi}{2}
\end{pmatrix}
\]
Riemannian geometry

Affine invariant metric:

\[ ds^2(Y) = \frac{1}{2} \text{tr} \left[ Y^{-1} dY \right]^2 = dt^2 + dr^2 + \sinh^2(r) d\varphi^2 \]

As a Hessian metric:

\[ \Phi(T, \theta) = -\log(T) - \log[1 - |\theta|^2] \quad \text{where} \quad T = e^t \]

This metric has negative sectional curvature \( \implies \) existence and uniqueness of centres of mass

Meaning of affine invariance (Riemannian homogeneous space):

affine transformation (\( A \) invertible) \( Y \mapsto A \cdot Y = A \cdot Y A^\dagger \)

transitive action (homogeneous space)

isometric action \( ds^2(Z) = ds^2(Y) \)
Riemannian Gaussian distributions

Gaussian distribution: maximum likelihood $\iff$ centre of mass

Probability density of $G(\bar{Y}, \sigma)$ 
\[ p(Y | \bar{Y}, \sigma) \propto \exp \left[ -\frac{d^2(Y, \bar{Y})}{2\sigma^2} \right] \]

$d(Y, \bar{Y})$ Riemannian distance

How to normalize:
\[ \int \exp \left[ -\frac{d^2(Y, \bar{Y})}{2\sigma^2} \right] dv(Y) \quad dv(Y) = \det(Y)^{-3/2} dY_{11} dY_{12} dY_{22} \]

Riemannian homogeneous space property: enough to consider $\bar{Y} = I$

integral $Z(\sigma) = \int_{-\infty}^{\infty} \int_{0}^{\infty} \int_{0}^{2\pi} \exp \left[ -(t^2 + r^2)/2\sigma^2 \right] \sinh(r) dt dr d\phi$

Analytic formula:
\[ Z(\sigma) = 8\pi^2 \times \sigma^2 \times e^{\frac{\sigma^2}{2}} \times \text{erf} \left( \frac{\sigma}{\sqrt{2}} \right) \]

Transformation property:
\[ Y \sim G(\bar{Y}, \sigma) \implies A \cdot Y \sim G(A \cdot \bar{Y}, \sigma) \]
Riemannian Gaussian distributions

Joint density of \((t, r, \varphi)\) under \(G(I, \sigma)\):

these three coordinates are independent

\[
p(t, r, \varphi) \propto \frac{1}{\pi} \times e^{-t^2/2\sigma^2} \times e^{-r^2/2\sigma^2} \sinh(r)
\]

Generating \(Y \sim G(\bar{Y}, \sigma)\):

<table>
<thead>
<tr>
<th>if (\bar{Y} = I)</th>
<th>if (\bar{Y} \neq I)</th>
</tr>
</thead>
<tbody>
<tr>
<td>– generate uniform (\varphi)</td>
<td>– generate (Y \sim G(I, \sigma))</td>
</tr>
<tr>
<td>– generate normal (t)</td>
<td>– apply affine transformation (Y \mapsto \bar{Y}^{1/2} \cdot Y)</td>
</tr>
<tr>
<td>– generate (r) from above density</td>
<td></td>
</tr>
<tr>
<td>– replace in spectral decomposition</td>
<td></td>
</tr>
</tbody>
</table>
Questions?

...and so, in conclusion, the proposed method...

THANK GOODNESS, ALMOST OVER... HOPEFULLY I DIDN'T BORE THEM TO TEARS.

...thank you, you've been a great audience...

OK, THE OBLIGATORY CALL FOR QUESTIONS AND I AM DONE...

any ques...

...tions?

OR NOT.

www.phdcomics.com