Gaussian distributions on Riemannian symmetric spaces : statistical learning with structured covariance matrices

Salem Said – Lionel Bombrun – Yannick Berthoumieu

Laboratoire IMS — UMR 5218

Talk based on two papers :

- 2015 : https ://arxiv.org/abs/1507.01760
- 2016 : https ://arxiv.org/abs/1607.06929 . . . both in IEEE Trans. Inf. Theory

---

In questions of Science, the authority of a thousand opinions
is not worth the reasoning of a single individual – Galileo

What is a Gaussian distribution ?

historic point of view : who discovered the Gaussian distribution ?

Statistical inference

Gauss (1809) : maximum likelihood $\Leftrightarrow$ centre of mass

we generalise this definition to Riemannian symmetric spaces

Diffusion process

Laplace (1810) : central limit theorem, random walks, Brownian motion

generalises to any space with a "Laplacian"

Statistical physics

Maxwell (1860) : rotation invariant independent components
velocity distribution in ideal mono-atomic gas

Poincaré (1912) : projection from a uniform distribution on $S^{\infty}(\infty^{1/2})$
extensively developed by Kac and Wiener

Variational definitions

— Information theory : maximum entropy for given dispersion

— Quantum mechanics : equality in Heisenberg inequality

$\rightsquigarrow$ different points of view require different definitions or generalisations

What is a Gaussian distribution ?

historic point of view : who discovered the Gaussian distribution ?

Statistical inference

Gauss (1809) : maximum likelihood ⇔ centre of mass

we generalise this definition to Riemannian symmetric spaces

Diffusion process

Laplace (1810) : central limit theorem, random walks, Brownian motion

generalises to any space with a "Laplacian"

Statistical physics

Maxwell (1860) : rotation invariant independent components
velocity distribution in ideal mono-atomic gas

Poincaré (1912) : projection from a uniform distribution on $S^{\infty}(\infty^{1/2})$
extensively developed by Kac and Wiener

Variational definitions

— Information theory : maximum entropy for given dispersion

— Quantum mechanics : equality in Heisenberg inequality

⇝ different points of view require different definitions or generalisations

The role of invariance

for Gaussian distributions on $\mathbb{R}$ ...

$$p(x\,|\,\bar{x},\sigma) = \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}}}_{\substack{\text{normalising const.}\\ \text{does not depend on } \bar{x}}} \exp\left[-\frac{(x-\bar{x})^2}{2\sigma^2}\right] \implies \ell(\bar{x},\sigma) = -N\log\sigma^2 - \frac{1}{\sigma^2}\underbrace{\sum_{n=1}^{N}(x_n-\bar{x})^2}_{\text{centre of mass problem}}$$

... everything follows from translation invariance

$$\begin{aligned}\text{normalising const.} &= \\ Z(\bar{x},\sigma) &= \underbrace{\int_{-\infty}^{+\infty}\exp\left[-\frac{(x-\bar{x})^2}{2\sigma^2}\right]dx = \int_{-\infty}^{+\infty}\exp\left[-\frac{x^2}{2\sigma^2}\right]dx}_{\text{translation invariant integral !!}} = Z(0,\sigma) \\ &= Z(\sigma)\end{aligned}$$

and we know the Poisson integral ... $= \sqrt{2\pi\sigma^2}$

⇝ replace translation invariance by invariance under a group of isometries

The role of invariance

replace $\mathbb{R}$ with a Riemannian homogeneous space M . . .

— Lie group G of isometries acts *Transitively* on M

$$g \in G : \qquad \underbrace{d(g \cdot x, g \cdot y) = d(x, y)}_{\text{invariant distance}} \qquad \underbrace{dv(g \cdot x) = dv(x)}_{\text{invariant volume}}$$

. . . everything follows from isometry invariance

$$p(x|\,\bar{x}, \sigma) = \underbrace{\frac{1}{Z(\sigma)}}_{\substack{\text{normalising const.} \\ \text{does not depend on } \bar{x}}} \exp\left[-\frac{d^2(x, \bar{x})}{2\sigma^2}\right] \qquad \text{density w.r.t. } dv(x)$$

---

normalising const. $=$

$$Z(\bar{x}, \sigma) = \underbrace{\int_M \exp\left[-\frac{d^2(x, \bar{x})}{2\sigma^2}\right] dv(x) = \int_M \exp\left[-\frac{d^2(x, o)}{2\sigma^2}\right] dv(x)}_{\text{let } \bar{x} = g \cdot o \text{ and use isometry invariance of the integral}} = Z(o, \sigma)$$

$$= Z(\sigma)$$

but how can this function be computed??

Computing $Z(\sigma)$

M a symmetric space of non-positive curvature ...

$M = G/K$ where G reductive of non-compact type
K compact subgroup

$\theta(g) = (g^{-1})^{\dagger}$ involution of G

$k \cdot o = o$ for $k \in K$

Why the name symmetric space ? $\underbrace{s(g \cdot o) = \theta(g) \cdot o}_{\text{symmetry about } o}$

Polar coordinates $x(a, k) = \exp(\mathrm{Ad}(k)\, a) \cdot o$ $k \in K,\ a \in \mathfrak{a}$ ($\mathfrak{a}$: Cartan subalgebra)

distance to origin $d^2(x, o) = B(a, a)$ $B(a, a) = \mathrm{tr}(a^2)$ (Ad-invariant form)

geodesic through origin $x(t) = x(t\,a, k)$ k, a constant

Rank of M $= \dim \mathfrak{a}$: dimension of maximal flat subspace

---

Expression of $Z(\sigma)$ $Z(\sigma) = \text{Const.} \times \int_{\mathfrak{a}} \exp\left[-\frac{B(a,a)}{2\sigma^2}\right] D(a)\, da$

where $D(a) = \prod_{\lambda > 0} \sinh^{m_\lambda}(|\lambda(a)|)$ $\lambda : \mathfrak{a} \to \mathbb{R}$ positive root of multiplicity m_λ

Statistical inference

M a symmetric space of non-positive curvature . . .

$$\text{log-likelihood function :} \qquad \ell(\bar{x}, \sigma) = -N \log\, Z(\sigma) - \frac{1}{2\sigma^2} \underbrace{\sum_{n=1}^{N} d^2(x_n, \bar{x})}_{\text{centre of mass problem}}$$

—MLE of $\bar{x}$

$$\hat{x}_N = \mathrm{argmin}_{x \in M} \sum_{n=1}^{N} d^2(x_n, x)$$

maximum likelihood $\Leftrightarrow$ centre of mass !!

$\rightsquigarrow$ M has non-positive curvature $\Rightarrow$ existence and uniqueness of centre of mass

—MLE of σ

natural parameter $\quad \eta = -\,1/2\sigma^2$

cumulant g.f. $\quad \psi(\eta) = \log\, Z(\sigma) \quad$ (strictly convex)

$$\hat{\eta}_N = (\psi')^{-1} \left(\tfrac{1}{N} \textstyle\sum_{n=1}^{N} d^2(x_n, \hat{x}_N) \right)$$

Mission accomplished

it is enough to know how to
- compute centre of mass
- compute $\psi(\eta)$

Max. entropy property

A kind of exponential family

natural parameter	cumulant g.f.	suff. statistic	cumulants
$\eta = -1/2\sigma^2$	$\psi(\eta) = \log Z(\sigma)$	$\Delta = d^2(x, \bar{x})$	$\psi'(\eta) = \mathrm{E}(\Delta)$
			$\psi''(\eta) = \mathrm{Var}(\Delta)$
			$\psi^{(n)}(\eta) = \mathrm{K}_n(\Delta)$

Duality and entropy

$$\rho = \mathrm{E}(\Delta) \qquad \underbrace{\psi^*(\rho) = \text{entropy of Gaussian distribution}}_{\text{Legendre transform of } \psi(\eta)}$$

Max. entropy

. . . *the Gaussian distribution is the unique maximum entropy distribution among all distributions on M having centre of mass $\bar{x}$ and dispersion ρ* . . .

Some examples

Rank of $M = 1$:

A — Hyperbolic space $\mathcal{H}_n$ $$Z(\sigma) = \mathrm{Vol}\left(S^{n-1}\right) \times \int_0^\infty e^{-\frac{r^2}{2\sigma^2}} \sinh^{n-1}(r)\, dr$$

Rank of $M = 2$:

B — 2×2 real covariance matrices

$$M = GL(2, \mathbb{R})/O(2) \qquad \mathfrak{a} = \{\mathrm{diag}(a_1, a_2) | a_1, a_2 \in \mathbb{R}\}$$

$$B(a, a) = a_1^2 + a_2^2$$

positive roots $\lambda(a) = a_1 - a_2$; $m_\lambda = 1$

$$\Rightarrow \quad Z(\sigma) = \mathrm{Const.} \times \sigma^2 \times e^{\frac{\sigma^2}{4}} \times \mathrm{erf}\left(\frac{\sigma}{2}\right)$$

C — 2×2 complex covariance matrices

$$M = GL(2, \mathbb{C})/U(2) \qquad \mathfrak{a} = \{\mathrm{diag}(a_1, a_2) | a_1, a_2 \in \mathbb{R}\}$$

$$B(a, a) = a_1^2 + a_2^2$$

positive roots $\lambda(a) = a_1 - a_2$; $m_\lambda = 2$

$$\Rightarrow \quad Z(\sigma) = \mathrm{Const.} \times \sigma^2 \times \left(e^{\sigma^2} - 1\right)$$

Some examples

Rank of $M = n$:

D — $n \times n$ real covariance matrices

$$M = GL(n, \mathbb{R})/O(n) \qquad \mathfrak{a} = \{\, \mathrm{diag}\,(a_1, \ldots, a_n) | \, a_i \in \mathbb{R}\,\}$$

$$B(a, a) = a_1^2 + \ldots + a_n^2$$

$$\text{positive roots} \qquad \lambda(a) = a_i - a_j \text{ for } i < j \;;\; m_\lambda = 1$$

$$\Rightarrow \quad Z(\sigma) = \text{Const.} \times \int_{\mathbb{R}^n} \exp\left[-\frac{|a|^2}{2\sigma^2} \right] \prod_{i<j} \sinh(|a_i - a_j|) \, da$$

E — $n \times n$ complex covariance matrices

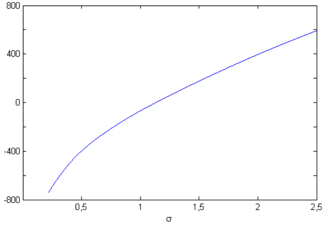
$$M = GL(n, \mathbb{C})/U(n) \qquad \mathfrak{a} = \{\, \mathrm{diag}\,(a_1, \ldots, a_n) | \, a_i \in \mathbb{R}\,\}$$

$$B(a, a) = a_1^2 + \ldots + a_n^2$$

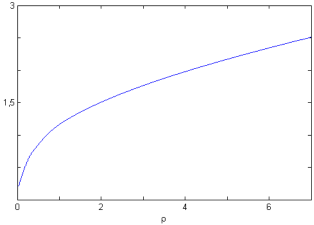
$$\text{positive roots} \qquad \lambda(a) = a_i - a_j \text{ for } i < j \;;\; m_\lambda = 2$$

$$\Rightarrow \quad Z(\sigma) = \text{Const.} \times \int_{\mathbb{R}^n} \exp\left[-\frac{|a|^2}{2\sigma^2} \right] \prod_{i<j} \sinh^2(|a_i - a_j|) \, da$$

Some examples



$n = 20$: graph of normalising const. $Z(\sigma)$



$n = 20$: graph of σ as function of ρ

F — $n \times n$ quaternion covariance matrices

$$M = GL(n, \mathbb{H})/Sp(n) \qquad \mathfrak{a} = \{ \operatorname{diag}(a_1, \ldots, a_n) | \, a_i \in \mathbb{R} \}$$

$$B(a, a) = a_1^2 + \ldots + a_n^2$$

positive roots $\quad \lambda(a) = a_i - a_j$ for $i < j$; $m_\lambda = 4$

$$\Rightarrow \quad Z(\sigma) = \text{Const.} \times \int_{\mathbb{R}^n} \exp\left[-\frac{|a|^2}{2\sigma^2} \right] \prod_{i<j} \sinh^4(|a_i - a_j|) \, da$$

G — Further examples : (Toeplitz, Block-Toeplitz), detailed in 2016 paper (arxiv)

Centre of mass and covariance

Variance function : $$\mathcal{E}(x) = \frac{1}{2}\int_M d^2(x, z)\, p(z|\,\bar{x}, \sigma)\, dv(z)$$

M has non-positive curvature $\Longrightarrow$ $\mathcal{E}$ strictly convex along geodesics

Riemannian gradient : $$\nabla\, \mathcal{E}(x) = -\int_M \mathrm{Log}_x(z)\, p(z|\,\bar{x}, \sigma)\, dv(z)$$

$\bar{x}$ is stationary point :

denote $s : M \longrightarrow M$ the symmetry about $\bar{x}$

for any $x \in M$	$\mathcal{E} \circ s = \mathcal{E}$	(s is an isometry and fixes $\bar{x}$)
then	$\nabla\mathcal{E} \circ s = ds \cdot \nabla\mathcal{E}$	(chain rule)
in particular	$\nabla\mathcal{E}(\bar{x}) = ds \cdot \nabla\mathcal{E}(\bar{x})$	
however	$ds = -\mathrm{Id}$ at $x = \bar{x}$	(s reverses geodesics at $\bar{x}$)

$\rightsquigarrow$ There exists an alternative proof which holds in any homogeneous space (using Fisher identity)

Centre of mass and covariance

Covariance form : $C(u,v) = \int_M \underbrace{\langle u, \mathrm{Log}_{\bar{x}}(z)\rangle \langle \mathrm{Log}_{\bar{x}}(z), v\rangle}_{(\mathrm{Log}_{\bar{x}}(z) \otimes \mathrm{Log}_{\bar{x}}(z))(u,v)} p(z|\bar{x},\sigma)\, dv(z) \qquad u, v \in T_{\bar{x}}M$

Invariance property

$$K_{\bar{x}} = \{ k \in G \mid k \cdot \bar{x} = \bar{x} \}$$

$R : K_{\bar{x}} \longrightarrow O(T_{\bar{x}}M) \qquad R_k = dk|_{\bar{x}}$ isotropy representation

$$C(u,v) = C(R_k \cdot u, R_k \cdot v)$$

$\rightsquigarrow$ De Rham decomposition theorem : $M = M_1 \times \ldots \times M_r$ each M_q irreducible

$u = u_1 + \ldots + u_r \quad v = v_1 + \ldots + v_r \quad u_q, v_q$ tangent to M_q

Schur's lemma $\Rightarrow$ $C(u,v) = \sum_{q=1}^{r} \frac{\psi_q'(\eta)}{\dim M_q} \langle u, v\rangle_{\bar{x}}$ (a diagonal matrix !!)

Alternatively . . .

Fisher information form : $I(u,v) = 4\eta^2 C(u,v)$

Mixtures of Gaussian distributions

$\rightsquigarrow$ to be concrete, (w.l.o.g.), let $M = GL(d, \mathbb{R})/O(d)$

$$\left\{ \begin{array}{c} \text{larga database of} \\ \text{signals or images} \end{array} \right\} \Rightarrow \left\{ \begin{array}{c} \text{statistical population of} \\ \text{covariance matrices} \end{array} \right\} \Rightarrow \left\{ \begin{array}{c} \text{learning model} \\ \text{(sufficiently general)} \end{array} \right\} \Rightarrow \underline{\text{Structure}}$$

— Learning model :

$$\underbrace{p(x)}_{\substack{\text{mixture distribution :} \\ \text{model of a generic population}}} = \sum_{\kappa=1}^{K} \omega_\kappa \times \underbrace{p(x|\,\bar{x}_\kappa\,,\sigma_\kappa)}_{\substack{\text{Gaussian distribution :} \\ \text{MaxEnt. model of a cluster}}}$$

— Learning problem :

real density $q(x)$	learned density $p_*(x)$
hopelessly complicated	best approximation of $q(x)$ within learning model

$$p_* = \operatorname{argmin}_p D(\,q\,\|\,p\,)$$

The EM algorithm

— Empirical cost function : based on data $x_1, \ldots, x_N$

$$D(\, q \,\|\, p\,) = \int_M q(x) \log\left(\frac{q(x)}{p(x)}\right) dv(x) \approx \frac{1}{N}\sum_{n=1}^{N} \log\, q\,(x_n) - \underbrace{\frac{1}{N}\sum_{n=1}^{N} \log\, p\,(x_n)}$$

⇒ max. likelihood "as if data were independent"

— EM is a usual solution :

E step — compute conditional weights $\pi_\kappa(x_n)$ % $\pi_\kappa(x_n) = \pi_\kappa(x_n | \hat{p})$

M step — $\hat{\omega}_\kappa =$ (usual formula)

$\hat{x}_\kappa = \text{argmin}_x \sum_{n=1}^{N} \pi_\kappa(x_n)\, d^{\,2}(x_n, x)$

$\hat{\eta}_\kappa = (\psi')^{-1}(\ldots)$

— In practice : somewhat difficult to exploit!!

a) linear convergence, gets trapped in local max. or saddlepoint

b) stores and processes all data points ("big data" problem)

— Ongoing work :

Stochastic EM : SEM, SAEM, . . . , overcomes both these problems

“one-pass” EM

— Meaning of one-pass : each data point x_n is treated only once, then forgotten

the asymptotic performance must be the same as MLE

— Examples of efficient one-pass algorithms :

stoch. Newton method ; natural gradient ; averaged stoch. gradient

---

Parameter space

$$\Theta = \left\{ \theta = \begin{pmatrix} s \\ (\bar{x}_\kappa) \\ (\eta_\kappa) \end{pmatrix} ; \begin{matrix} s \in S^{K-1} \text{ (unit sphere)} \\ \bar{x}_\kappa \in M \\ \eta_\kappa < 0 \end{matrix} \right\} \cong S^{K-1} \times M^K \times \mathbb{R}^K$$

Where does the sphere come from ?

$$s = (s_1, \ldots, s_K) \qquad s_\kappa^2 = \omega_\kappa$$

a usual replacement !!

⇝ it is necessary to compute the Fisher information of Θ

Natural gradient

How to achieve efficiency ??

$$\hat{\theta}_{n+1} = \mathrm{Exp}_{\hat{\theta}_n}\left[\gamma_{n+1} A_{\hat{\theta}_n} \cdot u(x_{n+1})\right]$$

Exp : from a natural connection

γ_{n+1} : step size

$A_\theta : T^*_\theta \Theta \longrightarrow T_\theta \Theta$

$u(x_{n+1}) = d\ell_{\mathrm{mixture}}(x_{n+1} \,|\, \hat{\theta}_n)$

$\rightsquigarrow$ A_θ = inverse of Fisher information

another possibility would be inverse of Hessian (tractable)

Example of A_θ

$K = 1$ and $M = GL(d, \mathbb{R})/O(d)$

$$\Theta = \mathbb{R} \times \mathcal{P}_d$$

$\rightsquigarrow$ De Rham decomposition : $\mathcal{P}_d = \mathbb{R} \times S\mathcal{P}_d$

$$x \longmapsto (t, s) \qquad t = \log\det(x) \qquad s = e^{-t/d}\, x$$

decomposition of tangent space :

$$\left\{\begin{array}{l} v \in T_x\mathcal{P}_n \\ v = x\left(s^{-1} v_2 + (1/d)\, v_1\right) \end{array}\right\} \longleftrightarrow \left\{\begin{array}{l} v_1 = \mathrm{tr}\left(x^{-1} v\right) \\ v_2 = \ldots \end{array}\right\}$$

Natural gradient

Fisher information

$$I_\theta\, v = \begin{pmatrix} \psi''(\eta) & & \\ & \phi_1(\eta) & \\ & & \phi_2(\eta) \end{pmatrix} \underbrace{\begin{pmatrix} v_\eta \\ v_1 \\ v_2 \end{pmatrix}}_{\substack{v = (v_\eta\,,\, v) \\ \text{tangent vector at } \theta}} \qquad \theta = (\sigma\,, x)$$

— Notation

$$\psi = \psi_1 + \psi_2 \qquad \psi(\eta) = \log Z(\sigma) \qquad \text{slide no. 7}$$

$$\psi_1(\eta) \sim \log(\sigma)$$

$$\phi_a(\eta) = 4\eta^2\ \psi_a'(\eta)/d_a \qquad \text{slide no. 8}$$

$$d_1 = 1\,;\ d_2 = \tfrac{d(d+1)}{2} - 1$$

$\rightsquigarrow\ A_\theta\ =\ \text{inverse of } I_\theta$

— Score form

$$u(x) = \underbrace{d\,\big(\,\eta\, d^2(x\,,\bar{x}) - \psi(\eta)\big)}_{d\ell(x|\,\theta)} \qquad \ldots \text{ usual calculations}$$

The algorithm

developed by post-doc Paolo Zanini

$$
\text{to process } x_{n+1}: \quad
\begin{aligned}
\hat{\eta}_{n+1} &= \hat{\eta}_n + \frac{\gamma_{n+1}}{\psi''(\hat{\eta}_n)} \left(d^2(x_{n+1}, \hat{x}_n) - \psi'(\hat{\eta}_n) \right) \\
\hat{t}_{n+1} &= \hat{t}_n + \gamma_{n+1} \left(t_{n+1} - \hat{t}_n \right) \\
\hat{s}_{n+1} &= \operatorname{Exp}_{\hat{s}_n} \left[\frac{\gamma_{n+1}}{\phi_2(\hat{\eta}_n)} \operatorname{Log}_{\hat{s}_n} s_{n+1} \right] \\
\hat{x}_{n+1} &= e^{\hat{t}_{n+1}/d} \, \hat{s}_{n+1}
\end{aligned}
$$

— Preliminary results

$$\sqrt{n} \left(\hat{\eta}_n - \eta \right) \Longrightarrow N\left(0, \frac{1}{\psi''(\eta)}\right)$$ efficient

$$\sqrt{n} \left(\hat{t}_n - t \right) \Longrightarrow N\left(0, \sigma^2\right)$$ efficient

$$\sqrt{n} \operatorname{Log}_s \hat{s}_n \Longrightarrow \ldots$$ we don't know yet!!

Summary ...

– Gaussian distributions give a statistical foundation to Riemannian centre of mass

– they can be defined on any Riemannian symmetric space of non-positive curvature

– in particular, this includes many important spaces of (structured) covariance matrices

– they have deeper connections with information geometry (not mentioned here)

– a Gaussian distribution is a maximum entropy model of a “cluster” in a manifold

– they provide a learning paradigm where any density on a manifold is a mixture of clusters

– estimating such mixtures is possible in principle using an expectation-maximisation algorithm

– however, this does not realistically apply to high dimensional big data : current difficulty !!

– to overcome this, we wish to consider stochastic or “one-pass” versions of EM

– the figure of merit is taken to be a form of consistency (minimum asymptotic covariance)

– this has lead us to consider the Riemannian geometry of the space of Gaussian distributions

– the model is nice but the applications still need to mature ... Thank you for your attention !!