Fisher-Rao distance and pullback Hilbert distance between multivariate normal distributions

GSI'23

Frank Nielsen

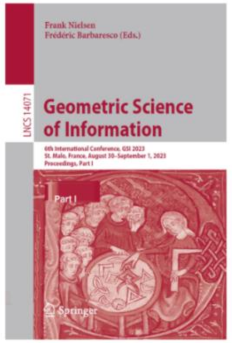
Sony Computer Science Laboratories Inc

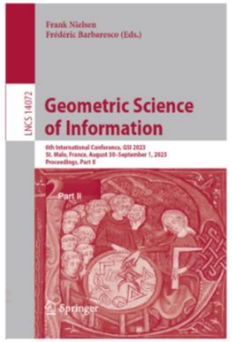
Sony CSL

arXiv:2307.10644

August 2023

References for this talk





- NB: Paper is *not included* in the GSI proceedings

- A Simple Approximation Method for the Fisher–Rao Distance between Multivariate Normal Distributions, Entropy (2023)

- **Fisher-Rao and pullback Hilbert cone distances on the multivariate Gaussian manifold with applications to simplification and quantization of mixtures**, ICML TAG-ML workshop (2023)

Overview and main contributions

- Give details of method [Kobayashi 2023] to calculate the Fisher-Rao geodesics between multivariate normal distributions with boundary conditions

- Report a **guaranteed (1+ε)-approximation** for the Fisher-Rao MVN distance

- Define **a fast metric distance** between d-variate MVNs based on Hilbert projective metric on the SPD cone of dimension d+1: pullback Hilbert distance

Rao distance and Fisher-Rao Riemannian geometry

- Consider a regular **statistical parametric model**: $\{p_{\lambda}: \lambda \in \Lambda\}$, $\dim(\Lambda)=m$

  regular = smooth partial derivatives, $\{\partial_1 p_\lambda, \ldots, \partial_m p_\lambda\}$ linearly independent

  or score functions $\{\partial_1 \log p_\lambda, \ldots, \partial_m \log p_\lambda\}$ defining the tangent plane

- Let the **Fisher information matrix** (FIM) defines the Riemannian metric g

  FIM well-defined, finite, and positive-definite → **Fisher metric tensor**

$$I(\lambda) = \mathrm{Cov}[\nabla \log p_\lambda(x)]$$

- Define the geodesic length as the **Rao distance**

[Atkinson & Mitchell 1981]

$$\rho_{\mathcal{N}}(N(\lambda_1), N(\lambda_2)) = \inf_{\substack{c(t) \\ c(0)=p_{\lambda_1} \\ c(1)=p_{\lambda_2}}} \{\mathrm{Length}(c)\},$$

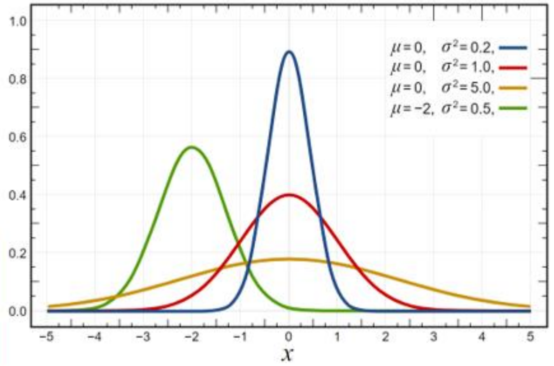
$$\mathrm{Length}(c) = \int_0^1 \sqrt{\langle \dot{c}(t), \dot{c}(t) \rangle_{c(t)}} \mathrm{d}t = \int_0^1 \mathrm{d}s_{\mathcal{N}}(t) \mathrm{d}t = \int_0^1 \|\dot{c}(t)\|_{c(t)} \mathrm{d}t.$$

- By construction, Rao's distance is **invariant to reparameterization**

[Rao 1945]
[Hotelling 1930]

Hyperbolic Fisher-Rao Gaussian manifold and partial isometric embedding on the 3D pseudo-sphere

$$\mathcal{P} = \left\{ p_\lambda(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right), \quad \lambda = (\mu, \sigma) \in \mathbb{H} \right\}$$



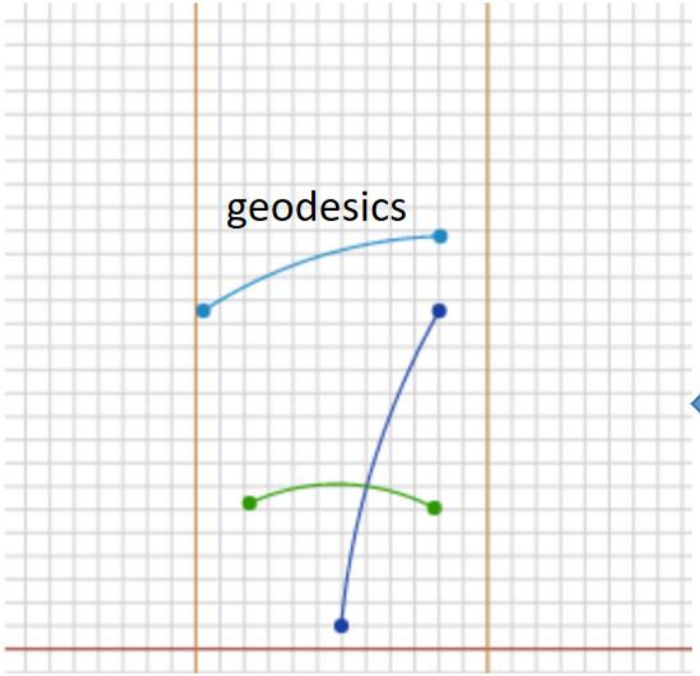
$$I(\mu,\sigma) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix}$$

Fisher-Rao geodesic distance:

$$D_{\mathrm{Rao}}[p_{\mu_1,\sigma_1}, p_{\mu_2,\sigma_2}] = \sqrt{2} \ln \frac{\left\| \left(\frac{\mu_1}{\sqrt{2}}, \sigma_1\right) - \left(\frac{\mu_2}{\sqrt{2}}, -\sigma_2\right) \right\| + \left\| \left(\frac{\mu_1}{\sqrt{2}}, \sigma_1\right) - \left(\frac{\mu_2}{\sqrt{2}}, \sigma_2\right) \right\|}{\left\| \left(\frac{\mu_1}{\sqrt{2}}, \sigma_1\right) - \left(\frac{\mu_2}{\sqrt{2}}, -\sigma_2\right) \right\| - \left\| \left(\frac{\mu_1}{\sqrt{2}}, \sigma_1\right) - \left(\frac{\mu_2}{\sqrt{2}}, \sigma_2\right) \right\|}$$

$$ds_F^2 = \frac{d\mu^2 + 2d\sigma^2}{\sigma^2}$$

Constant curvature -1/2
(= hyperbolic manifold)

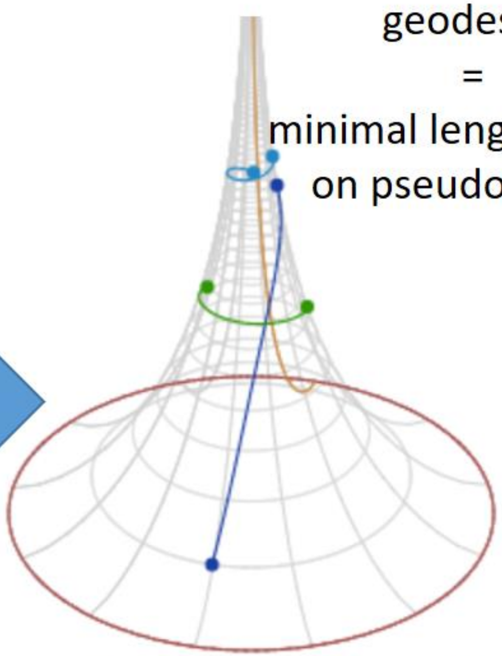


Poincaré upper half-plane

Constant Gaussian negative curvature

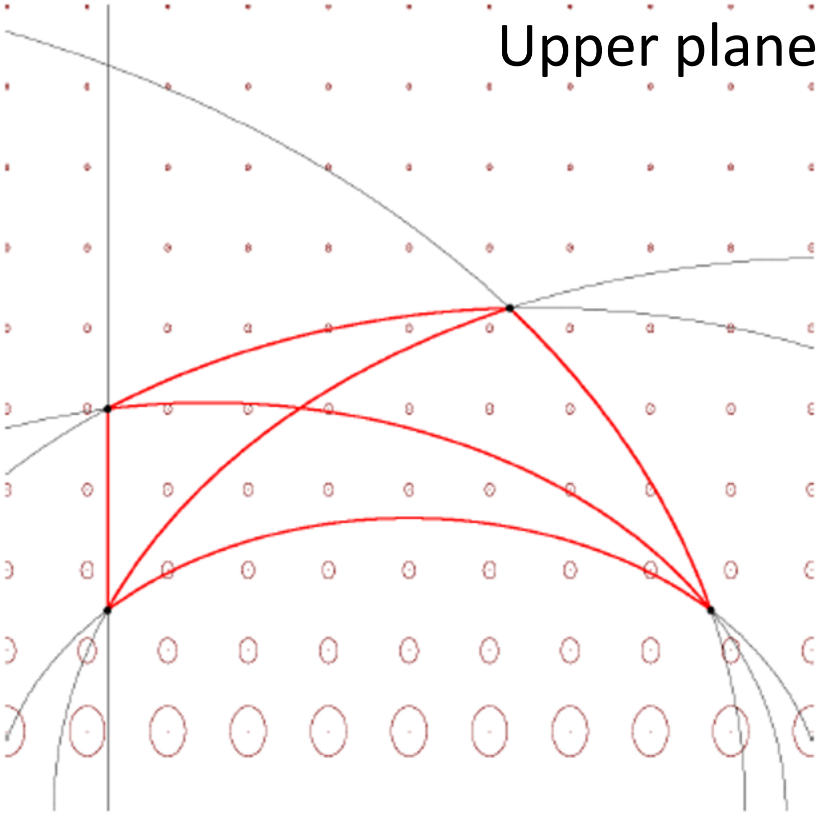
Isometric embedding (partial/periodic)

geodesics = minimal length curves on pseudosphere



Pseudosphere generated by tractrix

Fisher-Rao distance between normal distributions



$$\rho_{\mathcal{N}}(N(\mu_1,\sigma_1^2),N(\mu_2,\sigma_2^2)) = \sqrt{2}\log\left(\frac{1+\Delta(\mu_1,\sigma_1;\mu_2,\sigma_2)}{1-\Delta(\mu_1,\sigma_1;\mu_2,\sigma_2)}\right),$$

$$\Delta(a,b;c,d) = \sqrt{\frac{(c-a)^2+2(d-b)^2}{(c-a)^2+2(d+b)^2}}.$$

When same variance, we have

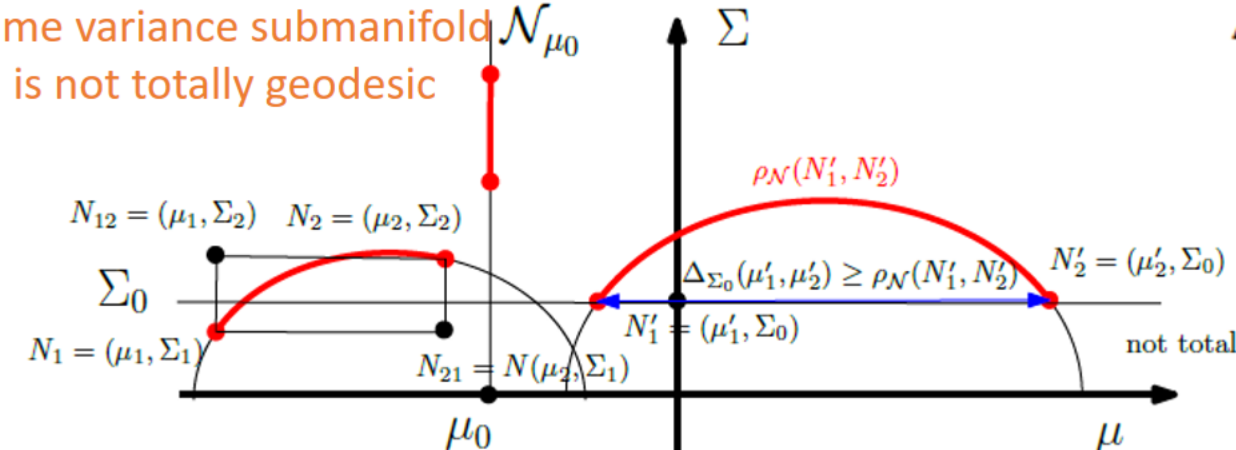
$$\rho_{\mathcal{N}}(N(\mu_1,\sigma_1^2),N(\mu_2,\sigma_2^2)) = \quad h_{\mathrm{FR}}(\Delta_{\sigma^2}(\mu_1,\mu_2))$$

$$\Delta_{\sigma^2}(\mu_1,\mu_2) = \sqrt{(\mu_2-\mu_1)(\sigma^2)-1(\mu_2-\mu_1)} = \frac{|\mu_2-\mu_1|}{\sigma}$$

$$h_{\mathrm{FR}}(u) \quad = \quad \sqrt{2}\,\log\left(\frac{\sqrt{8+u^2}+u}{\sqrt{8+u^2}-u}\right),$$

$$= \quad \sqrt{2}\,\mathrm{arccosh}\left(1+\frac{1}{4}u^2\right).$$

Same variance submanifold is not totally geodesic



not totally geodesic .

[Yoshizawa 1971]

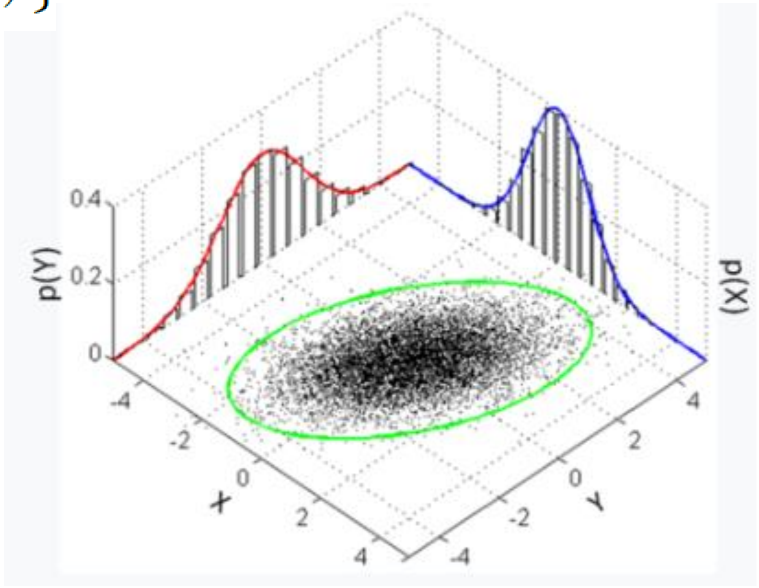
Fisher-Rao geometry: multivariate normals

$$N(\mu,\Sigma) \sim p_{\mu,\Sigma}(x) = \frac{(2\pi)^{-\frac{d}{2}}}{\sqrt{\det(\Sigma)}} \exp\left(-\frac{(x-\mu)^\top \Sigma^{-1}(x-\mu)}{2}\right)$$

$$\mathcal{N}(d) = \{N(\lambda) \ : \ \lambda = (\mu,\Sigma) \in \Lambda(d) = \mathbb{R}^d \times \mathrm{Sym}_+(d,\mathbb{R})\}$$

Fisher information matrix (vector, matrix):

$$g_{\mathcal{N}}^{\text{Fisher}}(\mu,\Sigma) = \mathrm{Cov}[\nabla \log p_{(\mu,\Sigma)}(x)]$$



Fisher metric tensor:

$$\begin{aligned} g_{(\mu,\Sigma)}^{\text{Fisher}}((v_1,V_1),(v_2,V_2)) &= \langle (v_1,V_1),(v_2,V_2)\rangle_{(\mu,\Sigma)}, \\ &= [v_1]^\top \Sigma^{-1}[v_2] + \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}[V_1]\Sigma^{-1}[V_2]\right). \end{aligned}$$

v= vector space R^d
V=Symmetric matrix vector space

Length element:

$$\begin{aligned} ds_{\mathcal{N}}^2(\mu,\Sigma) &= \begin{bmatrix} d\mu \\ d\Sigma \end{bmatrix}^\top I(\mu,\Sigma) \begin{bmatrix} d\mu \\ d\Sigma \end{bmatrix}, \\ &= d\mu^\top \Sigma^{-1} d\mu + \frac{1}{2}\mathrm{tr}\left(\left(\Sigma^{-1} d\Sigma\right)^2\right). \end{aligned}$$

[Skovgaard 1984]

Non-constant sectional curvatures which can also be positive, not NPC space (d>1)

Invariance under action of the positive affine group

- Length element/Rao distance is **invariant** under the action of the **positive affine group** (a,A):

$$\mathrm{Aff}_+(d,\mathbb{R}) := \{(a,A) \ : \ a \in \mathbb{R}^d, A \in \mathrm{GL}_+(d,\mathbb{R})\}$$

$$(a_1,A_1).(a_2,A_2) = (a_1 + A_1 a_2, A_1 A_2)$$

$$(a,A)^{-1} = (-A^{-1}a, A^{-1}))$$

Matrix group: $M_{(a,A)} := \begin{bmatrix} A & a \\ 0 & 1 \end{bmatrix}$

$$\rho_{\mathcal{N}}(N(A\mu_1 + a, A\Sigma_1 A^\top), N(A\mu_2 + a, A\Sigma_2 A^\top)) = \rho_{\mathcal{N}}(N(\mu_1,\Sigma_1), N(\mu_2,\Sigma_2)).$$

- Thus we may always consider one normal distribution is the **standard normal distribution** $\mathbf{N_{std}}$

$$\begin{aligned} \rho_{\mathcal{N}}(N(\mu_1,\Sigma_1), N(\mu_2,\Sigma_2)) &= \rho_{\mathcal{N}}\left(N_{\mathrm{std}}, N\left(\Sigma_1^{-\frac{1}{2}}(\mu_2-\mu_1), \Sigma_1^{-\frac{1}{2}}\Sigma_2\Sigma_1^{-\frac{1}{2}}\right)\right), \\ &= \rho_{\mathcal{N}}\left(N\left(\Sigma_2^{-\frac{1}{2}}(\mu_1-\mu_2), \Sigma_2^{-\frac{1}{2}}\Sigma_1\Sigma_2^{-\frac{1}{2}}\right), N_{\mathrm{std}}\right), \end{aligned}$$

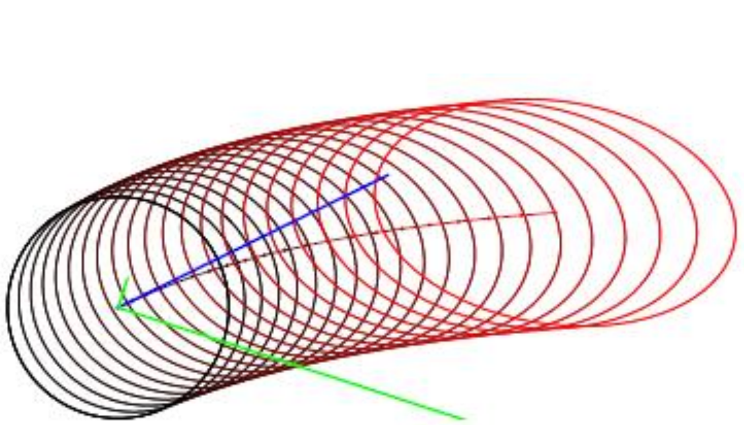
Geodesic equation

In general, geodesic wrt Levi-Civita connection $$\frac{d^2\theta_k}{dt^2} + \sum_{i=1}^{p}\sum_{j=1}^{p}\Gamma_{ij}^{k}\frac{d\theta_i}{dt}\frac{d\theta_j}{dt} = 0$$

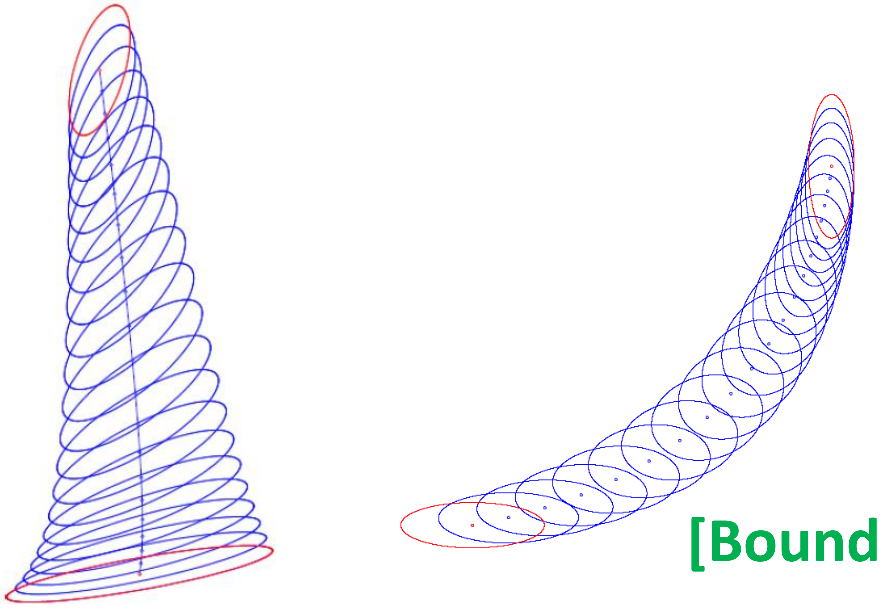
using (vector, Matrix) parameterization:

Second-order ODE: $$\begin{cases} \ddot{\mu} - \dot{\Sigma}\Sigma^{-1}\dot{\mu} & = & 0, \\ \ddot{\Sigma} + \dot{\mu}\dot{\mu}^\top - \dot{\Sigma}\Sigma^{-1}\dot{\Sigma} & = & 0. \end{cases}$$

- Consider either **initial value conditions** or **boundary value conditions** of ODE



[Initial values]



[Boundary values]

- Once the geodesics are known, **integrate length elements to get Rao distance**

Geodesic solution: Initial value condition (N_0=N_{std}) indirect solution $(v, V) \in T_{(\mu,\Sigma)} \in \mathbb{R}^d \times \mathrm{Sym}(d, \mathbb{R})$

- Manipulate geodesic equation **algebraically** $\left\{ \begin{array}{rcl} \ddot{\mu} - \dot{\Sigma}\Sigma^{-1}\dot{\mu} & = & 0, \\ \ddot{\Sigma} + \dot{\mu}\dot{\mu}^\top - \dot{\Sigma}\Sigma^{-1}\dot{\Sigma} & = & 0. \end{array} \right.$

- **natural parameterization** of the exponential family of MVNs: $\left(\xi = \Sigma^{-1}\mu, \Xi = \Sigma^{-1}\right)$
- Consider the **matrix exponential** (a la "**symmetric homogeneous space**") of **(2d+1)x(2d+1) matrices** to solve geodesics with initial values

$$A = \begin{bmatrix} -V & v & 0 \\ v^\top & 0 & -v^\top \\ 0 & -v^\top & V \end{bmatrix} \in \mathbb{P}(2d+1)$$

[Eriksen 1987]

Compute **matrix exponential**:

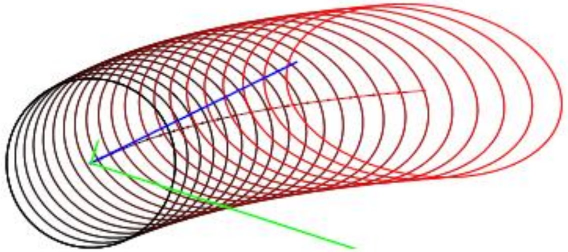
$\exp(tA)$ → retrieve natural parameters + convert to ordinary parameterization

$$\Xi(t) = [\exp(tA)]_{1:d,1:d}, \quad \xi(t) = [\exp(tA)]_{1:d,d+1}$$

$$\Sigma(t) = \Xi^{-1}(t), \quad \mu(t) = \Sigma(t)\xi(t)$$

Fisher-Rao geodesics from multivariate normals with initial value conditions (direct solution)

Geodesic equation: $\left\{\begin{array}{lcl} \ddot{\mu} - \dot{\Sigma}\Sigma^{-1}\dot{\mu} & = & 0, \\ \ddot{\Sigma} + \dot{\mu}\dot{\mu}^\top - \dot{\Sigma}\Sigma^{-1}\dot{\Sigma} & = & 0. \end{array}\right.$



When initial value conditions $(a = \dot{\xi}(0), B = \dot{\Xi}(0))$ are given, the geodesics are known in **closed-form** using the **natural parameters** $(\xi = \Sigma^{-1}\mu, \Xi = \Sigma^{-1})$

$$\boxed{\begin{array}{lcl} \Xi(t) & = & \Xi(0)^{\frac{1}{2}}\, R(t)R(t)^\top\, \Xi(0)^{\frac{1}{2}}, \\ \xi(t) & = & 2\Xi(0)^{\frac{1}{2}}\, R(t)\mathrm{Sinh}\left(\frac{1}{2}Gt\right) G^\dagger a + \Xi(t)\Xi^{-1}(0)\xi(0), \end{array}}$$

[Calvo & Oller 1991]

with $R(t) = \mathrm{Cosh}\left(\frac{1}{2}Gt\right) - BG^\dagger\mathrm{Sinh}\left(\frac{1}{2}Gt\right)$ and matrix pseudo-inverse $G^\dagger = (G^\top G)^{-1}G^\top$

and matrix hyperbolic functions for M= $O\,\mathrm{diag}(\lambda_1, \ldots, \lambda_d)\,O^\top$

$$\begin{array}{lcl} \mathrm{Sinh}(M) & = & O\,\mathrm{diag}(\sinh(\lambda_1), \ldots, \sinh(\lambda_d))\,O^\top \\ \mathrm{Cosh}(M) & = & O\,\mathrm{diag}(\cosh(\lambda_1), \ldots, \cosh(\lambda_d))\,O^\top \end{array}$$

Special case: Centered multivariate normals Closed form geodesics and Fisher-Rao distances

- Submanifold of MVNs with constant mean is **totally geodesic**

[James 1973]

- **Rao geodesics**:

$$\gamma_{\mathrm{FR}}^{\mathcal{N}}(N_0, N_1; t) = N(\mu, \Sigma_t)$$

$$\Sigma_t = \Sigma_0^{\frac{1}{2}} \left(\Sigma_0^{-\frac{1}{2}} \Sigma_1 \Sigma_0^{-\frac{1}{2}}\right)^t \Sigma_0^{\frac{1}{2}}$$

- **Rao distance**:

$$\rho_{\mathcal{N}_\mu}(N_0, N_1) = \sqrt{\frac{1}{2} \sum_{i=1}^{d} \log^2 \lambda_i\left(\Sigma_0^{-\frac{1}{2}} \Sigma_1 \Sigma_0^{-\frac{1}{2}}\right)}$$

- Require to compute all eigenvalues (costly)
- Because of sum of $\log^2$, $\rho(P_1,P_2)=\rho(P_1^{-1},P_2^{-1})$: **invariant to matrix inversion**

Riemanian geometry of the SPD cone (trace metric)

Trace metric: $\langle A, B\rangle_P = \mathrm{tr}(P^{-1}AP^{-1}B)$

related to Fisher information of centered normal/Wishart

$$I_F(\Sigma) = \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\Sigma^{-1}\right) \qquad I_F(V) = \frac{1}{2}n\mathrm{tr}\left(V^{-1}V^{-1}\right)$$

Length element: $\mathrm{d}s_P^2 = \mathrm{tr}(P^{-1}\mathrm{d}P\; P^{-1}\mathrm{d}P)$

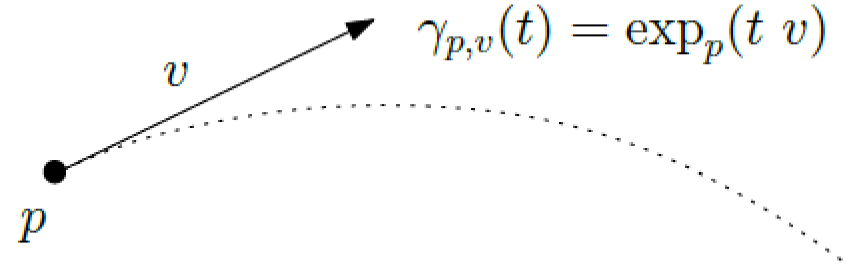
$$\text{Invariance: } \mathrm{d}s^2_{CPC^\top} = \mathrm{d}s^2_P,\ \mathrm{d}s^2_{P^{-1}} = \mathrm{d}s^2_P$$

Geodesic equation: $\ddot{P} - \dot{P}P^{-1}\dot{P} = 0$

Initial value conditions:

$$P(0) = P \text{ and } \dot{P}(0) = S$$

$$P(t) = P^{\frac{1}{2}}\exp(tP^{-\frac{1}{2}}SP^{-\frac{1}{2}})P^{\frac{1}{2}}$$

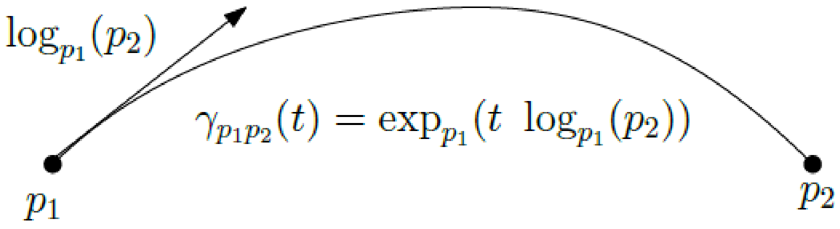


Geodesic wrt. initial conditions

Boundary value conditions:

$$P(0) = P_1,\ P(1) = P_2$$

$$P(t) = P_1^{\frac{1}{2}}\exp(t\mathrm{Log}(P_1^{-\frac{1}{2}}P_2P_1^{-\frac{1}{2}}))P_1^{\frac{1}{2}}$$



Geodesic wrt. boundary conditions

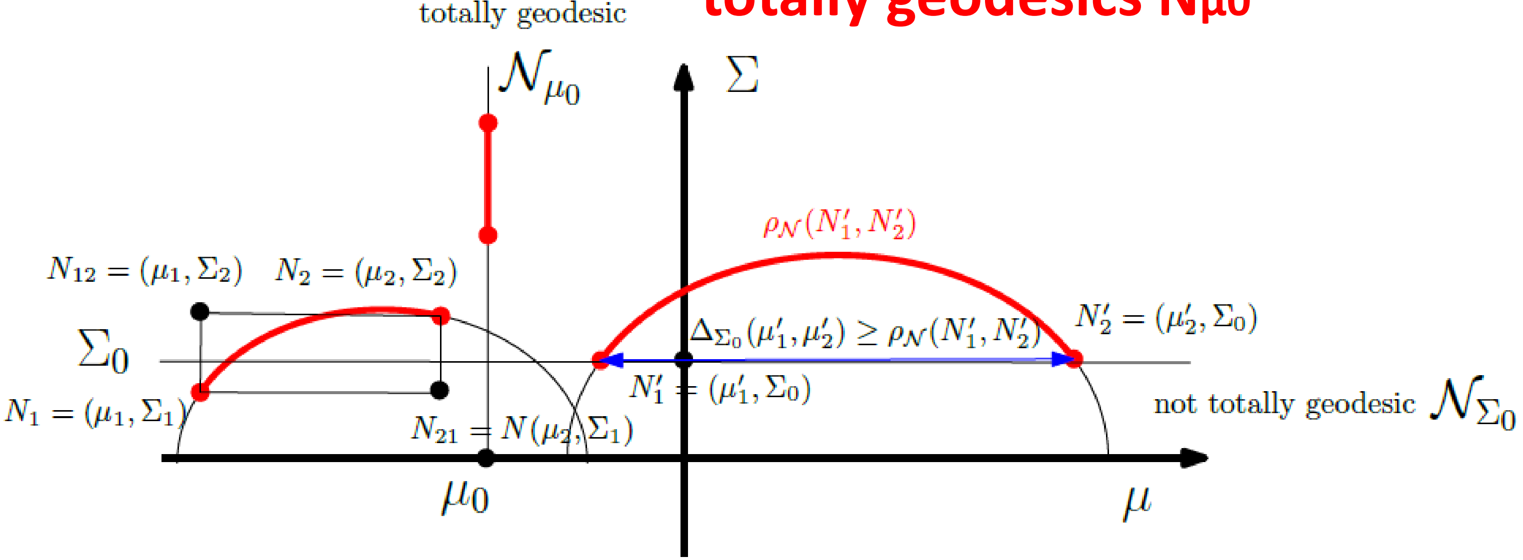
Rao's distance:

$$\rho(P_1, P_2) = \sqrt{\sum_i \log^2 \lambda_i(P_1^{-1}P_2)}$$

$$\rho(P_1, P_2) = \|\mathrm{Log}(P_1^{-\frac{1}{2}}P_2P_1^{-\frac{1}{2}})\|_F$$

Submanifolds of constant covariance matrices

totally geodesics $N_{\mu 0}$



not totally geodesics $N_{\Sigma 0}$

Proposition . *The Fisher–Rao distance $\rho_{\mathcal{N}}((\mu_1, \Sigma), (\mu_2, \Sigma))$ between two MVNs with same covariance matrix is*

$$\begin{aligned}
\rho_{\mathcal{N}}((\mu_1, \Sigma), (\mu_2, \Sigma)) &= \rho_{\mathcal{N}}((0,1), (\Delta_\Sigma(\mu_1, \mu_2), 1)), \\
&= \sqrt{2} \log \left(\frac{\sqrt{8 + \Delta_\Sigma^2(\mu_1, \mu_2)} + \Delta_\Sigma(\mu_1, \mu_2)}{\sqrt{8 + \Delta_\Sigma^2(\mu_1, \mu_2)} - \Delta_\Sigma(\mu_1, \mu_2)} \right), \\
&= \sqrt{2} \operatorname{arccosh} \left(1 + \frac{1}{4} \Delta_\Sigma^2(\mu_1, \mu_2) \right),
\end{aligned}$$

where $\Delta_\Sigma(\mu_1, \mu_2) = \sqrt{(\mu_2 - \mu_1)^\top \Sigma^{-1} (\mu_2 - \mu_1)}$ *is the Mahalanobis distance.*

Fisher-Rao geodesics from multivariate normals with boundary values conditions in closed form

Fisher-Rao geodesic $N_t = N(\mu(t), \Sigma(t)) = \gamma_{\mathrm{FR}}^{\mathcal{N}}(N_0, N_1; t)$:

[Kobayashi 2023]

- For $i \in \{0,1\}$, let $G_i = M_i\, D_i\, M_i^\top$, where

$$M_i = \begin{bmatrix} \Sigma_i^{-1} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \Sigma_i \end{bmatrix}, D_i = \begin{bmatrix} I_d & 0 & 0 \\ \mu_i^\top & 1 & 0 \\ 0 & -\mu_i & I_d \end{bmatrix}$$

- Consider the Riemannian geodesic in $\mathrm{Sym}_+(2d+1, \mathbb{R})$ with respect to the trace metric: $G(t) = G_0^{\frac{1}{2}} \left(G_0^{-\frac{1}{2}} G_1 G_0^{-\frac{1}{2}}\right)^t G_0^{\frac{1}{2}}$
- Retrieve $N(t) = \gamma_{\mathrm{FR}}^{\mathcal{N}}(N_0, N_1; t) = N(\mu(t), \Sigma(t))$ from $G(t)$:

$\Sigma(t) = [G(t)]_{1:d,1:d}^{-1}, \mu(t) = \Sigma(t)\,[G(t)]_{1:d,d+1}$ where $[G]_{1:d,1:d}$ denotes the block matrix with rows and columns ranging from 1 to d extracted from $(2d+1) \times (2d+1)$ matrix G, and $[G]_{1:d,d+1}$ is similarly the column vector of $\mathbb{R}^d$ extracted from G

Ingredient: **Riemannian submersion** and MVN geodesics from horizontal geodesics

- Get **closed-form geodesics with boundary values**
- However, **no closed-form Rao distance** because of the integration of length element problem

$$\mathrm{Length}(c) = \int_0^1 \sqrt{\langle \dot{c}(t), \dot{c}(t) \rangle_{c(t)}}\mathrm{d}t = \int_0^1 \mathrm{d}s_{\mathcal{N}}(t)\mathrm{d}t = \int_0^1 \|\dot{c}(t)\|_{c(t)}\mathrm{d}t.$$

Fisher-Rao MVN distance: An upper bound

- Geodesics with boundary conditions form **1d totally geodesic submanifolds**
- Cut the geodesics in many small parts using T+1 geodesic points

$$\tilde{\rho}^c_{\mathcal{N}}(N_1, N_2) := \frac{1}{T}\sum_{i=1}^{T-1}\sqrt{D_J\left[c\left(\frac{i}{T}\right), c\left(\frac{i+1}{T}\right)\right]}.$$

$$D_J[p_{(\mu_1,\Sigma_1)} : p_{(\mu_2,\Sigma_2)}] = \mathrm{tr}\left(\frac{\Sigma_2^{-1}\Sigma_1 + \Sigma_1^{-1}\Sigma_2}{2} - I\right) + \Delta\mu^\top \frac{\Sigma_1^{-1} + \Sigma_2^{-1}}{2}\Delta\mu.$$

- **Upper bound** for nearby points Rao distance by the square root of Jeffreys divergence (or any other f-divergence)

$$I_f[p:q] \approx \frac{f''(1)}{2}\mathrm{d}s^2_{\mathrm{Fisher}},$$

Infinitesimal Fisher-Rao distance: $\mathrm{d}s \approx \sqrt{\frac{2\, I_f[p:q]}{f''(1)}}.$

Property (Fisher–Rao upper bound). *The Fisher–Rao distance between normal distributions is upper bounded by the square root of the Jeffreys divergence:* $\rho_{\mathcal{N}}(N_1, N_2) \leq \sqrt{D_J(N_1, N_2)}$.

[N. 2023]

Diffeomorphic embeddings of MVN(d) onto SPD(d+1)

[Calvo & Oller 1990]

The **diffeomorphisms** $\{f_\beta\}$ foliates the SPD cone P(d+1)

$$f_\beta(N) = f_\beta(\mu, \Sigma) = \left[\begin{array}{cc} \Sigma + \beta\mu\mu^\top & \beta\mu \\ \beta\mu^\top & \beta \end{array} \right] \in \mathcal{P}(d+1)$$

Using half trace metric in P(d+1), we get the following **metrics on MVN(d)**:

$$\begin{aligned} \mathrm{d}s^2_{\mathrm{CO}} &= \frac{1}{2}\mathrm{tr}\left(\left(f^{-1}(\mu,\Sigma)\mathrm{d}f(\mu,\Sigma)\right)^2\right), \\ &= \frac{1}{2}\left(\frac{\mathrm{d}\beta}{\beta}\right)^2 + \beta \mathrm{d}\mu^\top \Sigma^{-1}\mathrm{d}\mu + \frac{1}{2}\mathrm{tr}\left(\left(\Sigma^{-1}\mathrm{d}\Sigma\right)^2\right). \end{aligned}$$

β=1

When **β=1** (constant), we thus get a **Fisher isometric embedding** of MVN(d) into SPD(d+1):

$$\mathrm{d}s^2_{\mathrm{Fisher}} = \mathrm{d}\mu^\top\Sigma^{-1}\mathrm{d}\mu + \frac{1}{2}\mathrm{tr}\left(\left(\Sigma^{-1}\mathrm{d}\Sigma\right)^2\right)$$

Fisher-Rao MVN distance: A lower bound

- Embed isometrically the Gaussian manifold N(d) into a **submanifold of codimension 1 into the SPD cone of dimension d+1** (non-totally geodesic):

$$f(N) = f(\mu, \Sigma) = \begin{bmatrix} \Sigma + \mu\mu^\top & \mu \\ \mu^\top & 1 \end{bmatrix}$$

[Calvo & Oller 1990]

- Use SPD geodesic in the (d+1)-dimensional cone: $\Sigma_t = \Sigma_0^{\frac{1}{2}} (\Sigma_0^{-\frac{1}{2}} \Sigma_1 \Sigma_0^{-\frac{1}{2}})^t \Sigma_0^{\frac{1}{2}}$

- SPD path is of length necessarily smaller than the MVN geodesic in submanifold f(N). Thus get a **lower bound** on Rao distance:

$$\boxed{\rho_{\mathcal{N}}(N_1, N_2) \geq \rho_{\mathrm{CO}}(\underbrace{f(\mu_1, \Sigma_1)}_{\bar{P}_1}, \underbrace{f(\mu_2, \Sigma_2)}_{\bar{P}_2}) = \sqrt{\frac{1}{2} \sum_{i=1}^{d+1} \log^2 \lambda_i(\bar{P}_1^{-1} \bar{P}_2)}.}$$

- Cut MVN geodesics into and apply lower bound piecewisely : **Fine lower bound**

Fisher-Rao MVN geodesic:
Numerical midpoint geodesic with quadratic convergence

Computing SPD geodesics points require all eigenvalues/eigenvectors:

$$\Sigma_t = \Sigma_0^{\frac{1}{2}} \left(\Sigma_0^{-\frac{1}{2}} \Sigma_1 \Sigma_0^{-\frac{1}{2}}\right)^t \Sigma_0^{\frac{1}{2}}$$

For **t=1/2**, we can compute $\Sigma_{1/2}$ with **quadratic convergence** (thus bypassing eigendecomposition) as follows:

Matrix AHM mean:

A_{t+1}=ArithmeticMean(A_t,B_t)

B_{t+1}=HarmonicMean(A_t,B_t)

initialized with $A_0=\Sigma_0$ and $B_0=\Sigma_1$

$$\mathrm{ArithmeticMean}(A,B) = \frac{1}{2}(A+B)$$

$$\mathrm{HarmonicMean}(A,B) = 2(A^{-1}+B^{-1})^{-1}$$

Converge to the matrix geometric mean

$$A^{1/2}(A^{-1/2}BA^{-1/2})^{1/2}A^{1/2}$$

[Nakamura 2012]

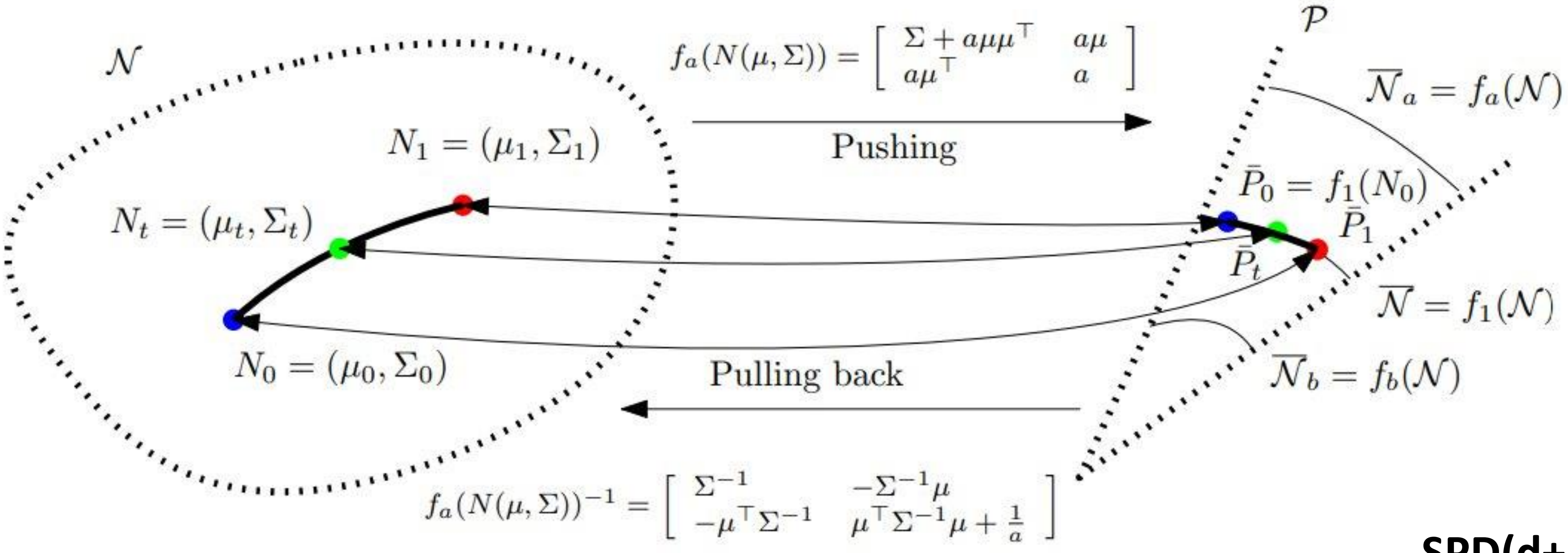
New fast distances between multivariate normals

$$\rho_{\text{Hilbert}}(N_0, N_1) := \rho_{\text{Hilbert}}(f(N_0), f(N_1))$$

$$\rho_{\text{Hilbert}}(P_0, P_1) = \log\left(\frac{\lambda_{\max}(P_0^{-\frac{1}{2}} P_1 P_0^{-\frac{1}{2}})}{\lambda_{\min}(P_0^{-\frac{1}{2}} P_1 P_0^{-\frac{1}{2}})}\right)$$

$$= \log\left(\frac{\lambda_{\max}(P_0^{-1} P_1)}{\lambda_{\min}(P_0^{-1} P_1)}\right)$$

Gaussian(d) manifold



SPD(d+1) cone

New fast distances between multivariate normals

- Use Calvo & Oller isometric cone embedding f(μ,Σ) $f(N) = f(\mu, \Sigma) = \begin{bmatrix} \Sigma + \mu\mu^\top & \mu \\ \mu^\top & 1 \end{bmatrix}$

- In the cone, use **Hilbert projective metric distance** and **LERP pregeodesics**

$$\rho_{\mathrm{Hilbert}}(P_0, P_1) = \log\left(\frac{\lambda_{\max}(P_0^{-\frac{1}{2}} P_1 P_0^{-\frac{1}{2}})}{\lambda_{\min}(P_0^{-\frac{1}{2}} P_1 P_0^{-\frac{1}{2}})}\right)$$
$$= \log\left(\frac{\lambda_{\max}(P_0^{-1} P_1)}{\lambda_{\min}(P_0^{-1} P_1)}\right)$$

Projective metric on SPD

$\rho_{\mathrm{Hilbert}}(P_0, P_1) = 0$ if and only if $P_0 = \lambda P_1$

But proper metric on f(N)

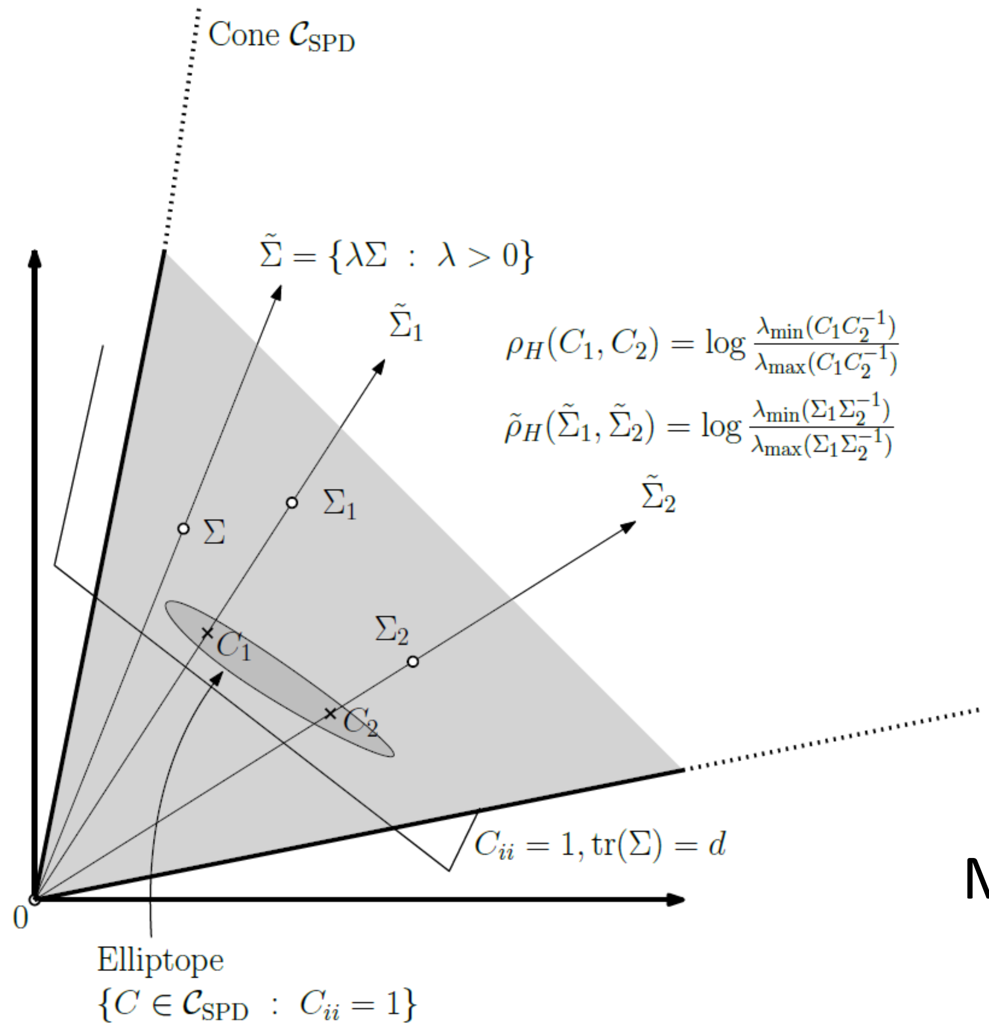
$$\gamma_{\mathrm{Hilbert}}(P_0, P_1; t) := \left(\frac{\beta\alpha^t - \alpha\beta^t}{\beta - \alpha}\right) P_0 + \left(\frac{\beta^t - \alpha^t}{\beta - \alpha}\right) P_1$$

$\alpha = \lambda_{\min}(P_1^{-1} P_0)$ and $\beta = \lambda_{\max}(P_1^{-1} P_0)$

- **Pullback** the geodesics and distance into the Gaussian manifold

$$\rho_{\mathrm{Hilbert}}(N_0, N_1) := \rho_{\mathrm{Hilbert}}(f(N_0), f(N_1))$$

Hilbert projective metric distance in the SPD cone



$$\rho_H(C_1, C_2) = \log \frac{\lambda_{\min}(C_1 C_2^{-1})}{\lambda_{\max}(C_1 C_2^{-1})}$$

$$\tilde{\rho}_H(\tilde{\Sigma}_1, \tilde{\Sigma}_2) = \log \frac{\lambda_{\min}(\Sigma_1 \Sigma_2^{-1})}{\lambda_{\max}(\Sigma_1 \Sigma_2^{-1})}$$

$$\rho_H(\lambda_1 p_1, \lambda_2 p_2) = \rho_H(p_1, p_2), \quad \forall \lambda_1, \lambda_2 > 0$$

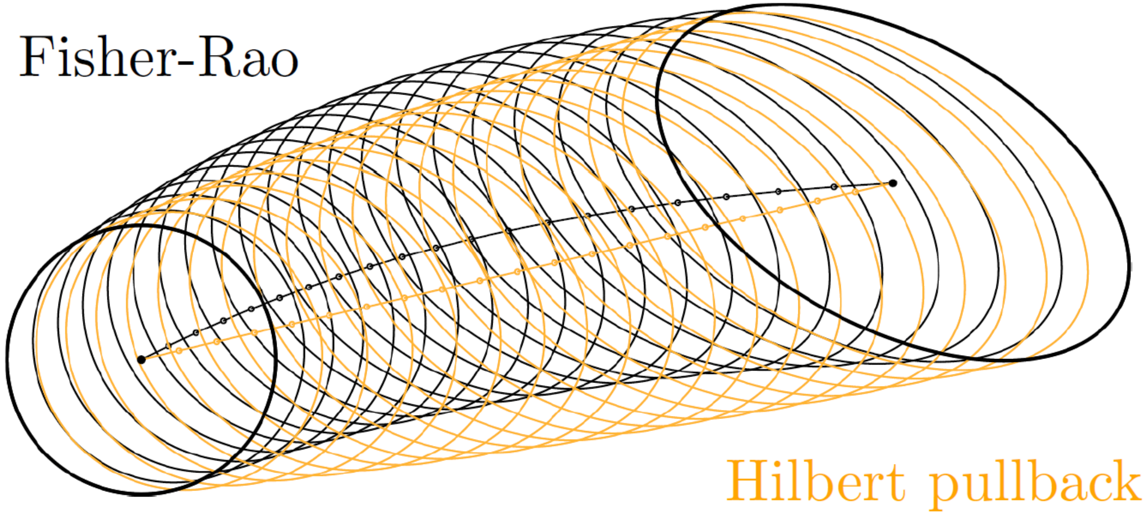
Metric distance in the elliptope of correlation matrices

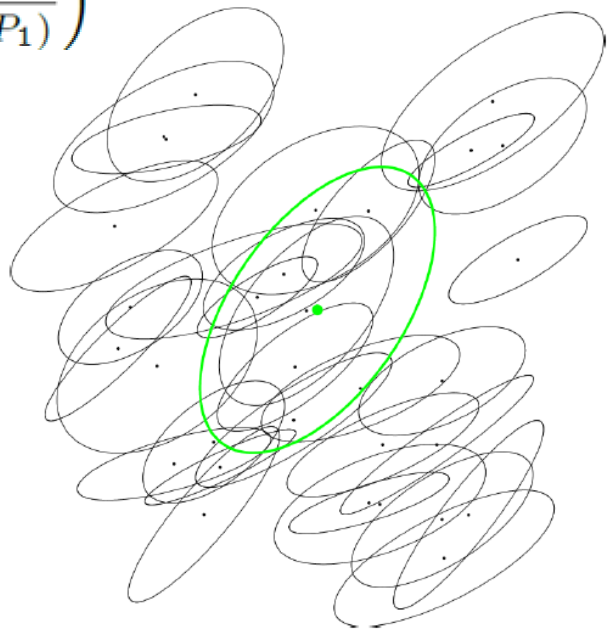
N. and Sun. "Clustering in Hilbert's projective geometry: The case studies of the probability simplex and the elliptope of correlation matrices." *Geometric structures of information* (2019): 297-331.

Pullback Hilbert distance/geodesics between MVNs

Only require to calculate **extreme eigenvalues** (eg., power method iteration)

$$\rho_{\text{Hilbert}}(P_0, P_1) = \log\left(\frac{\lambda_{\max}(P_0^{-1}P_1)}{\lambda_{\min}(P_0^{-1}P_1)}\right)$$





$$\rho_{\text{Hilbert}}(N_0, N_1) := \rho_{\text{Hilbert}}(f(N_0), f(N_1))$$

Applications: Approximation of the smallest enclosing ball (SEB) of a set of multivariate normals (quantization/clustering of Gaussian mixtures)

Summary: A (1+ε)-approximation of Rao's distance between multivariate normal distributions

Algorithm 2. $\tilde\rho_{\mathrm{FR}}(N_0,N_1)=\mathrm{ApproximateRaoMVN}(N_0,N_1,\epsilon)$:

- $l=\rho_{\mathrm{CO}}(N_0,N_1)$; /* Calvo & Oller lower bound (Proposition 2.1) */
- $u=\sqrt{D_J(N_0,N_1)}$; /* Jeffreys divergence D_J (Proposition 1) */
- if $\left(\frac{u}{l}>1+\epsilon\right)$
  - $N=\mathrm{GeodesicMidpoint}(N_0,N_1)$; /* see Algorithm 1 for $t=\frac{1}{2}$. */
  - return $\mathrm{ApproximateRaoMVN}(N_0,N,\epsilon)+\mathrm{ApproximateRaoMVN}(N,N_1,\epsilon)$;

  else return u;

Algorithm 1. Fisher-Rao geodesic $N_t=N(\mu(t),\Sigma(t))=\gamma^{\mathcal{N}}_{\mathrm{FR}}(N_0,N_1;t)$:

- For $i\in\{0,1\}$, let $G_i=M_i\,D_i\,M_i^\top$, where

$$M_i = \begin{bmatrix}\Sigma_i^{-1} & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & \Sigma_i\end{bmatrix}, \quad (8)$$

$$D_i = \begin{bmatrix} I_d & 0 & 0\\ \mu_i^\top & 1 & 0\\ 0 & -\mu_i & I_d\end{bmatrix}, \quad (9)$$

  where I_d denotes the identity matrix of shape $d\times d$. That is, matrices G_0 and $G_1\in\mathrm{Sym}_+(2d+1,\mathbb{R})$ can be expressed by *block Cholesky factorizations*.

- Consider the Riemannian geodesic in $\mathrm{Sym}_+(2d+1,\mathbb{R})$ with respect to the trace metric:

$$G(t)=G_0^{\frac{1}{2}}\left(G_0^{-\frac{1}{2}}G_1G_0^{-\frac{1}{2}}\right)^t G_0^{\frac{1}{2}}.$$

  In order to compute the matrix power G^p for $p\in\mathbb{R}$, we first calculate the Singular Value Decomposition (SVD) of G: $G=O\,L\,O^\top$ (where O is an orthogonal matrix and $L=\mathrm{diag}(\lambda_1,\ldots,\lambda_{2d+1})$ a diagonal matrix) and then get the matrix power as $G^p=O\,L^p\,O^\top$ with $L^p=\mathrm{diag}(\lambda_1^p,\ldots,\lambda_{2d+1}^p)$.

- Retrieve $N(t)=\gamma^{\mathcal{N}}_{\mathrm{FR}}(N_0,N_1;t)=N(\mu(t),\Sigma(t))$ from $G(t)$:

$$\Sigma(t) = [G(t)]^{-1}_{1:d,1:d}, \quad (10)$$

$$\mu(t) = \Sigma(t)\,[G(t)]_{1:d,d+1}, \quad (11)$$

  where $[G]_{1:d,1:d}$ denotes the block matrix with rows and columns ranging from 1 to d extracted from $(2d+1)\times(2d+1)$ matrix G, and $[G]_{1:d,d+1}$ is similarly the column vector of $\mathbb{R}^d$ extracted from G.

$$\overline{\mathcal{N}}=f(N):=\begin{bmatrix}\Sigma+\mu\mu^\top & \mu\\ \mu^\top & 1\end{bmatrix}\in\mathcal{P}(d+1)$$

$$\rho_{\mathrm{CO}}(N_0,N_1)=\frac{1}{\sqrt{2}}\sum_{i=1}^{d+1}\log^2\lambda_i(\overline{\mathcal{N}}_0^{-\frac{1}{2}}\overline{\mathcal{N}}_1\overline{\mathcal{N}}_0^{-\frac{1}{2}})$$

$$D_J(N_1,N_2)=\mathrm{tr}\left(\frac{\Sigma_2^{-1}\Sigma_1+\Sigma_1^{-1}\Sigma_2}{2}-I\right)+(\mu_2-\mu_1)^\top\frac{\Sigma_1^{-1}+\Sigma_2^{-1}}{2}(\mu_2-\mu_1)$$

A recursive algorithm

Summary and concluding remarks

- Geodesics with initial values or boundary values are known in **closed-form**
- **Rao distance's lower bound** using isometric embedding into SPD(d+1).
  Thus get **arbitrarily fine lower bounds** using piecewise MVN Rao geodesics
- **Arbitrarily fine upper bound** using square root of Jeffreys divergence
  on piecewise MVN Rao geodesics

- **Pullback** of SPD cone distance via Calvo & Oller **isometric embedding**:
  Fast distance & geodesic requiring only **extremal eigenvalues**

- Gaussian/MVN manifold is not NPC/Hadamard/CAT(0) because there
  are some positive sectional curvatures. SPD cone is NPC.
- Siegel considered a complex matrix metric which yields a NPC space

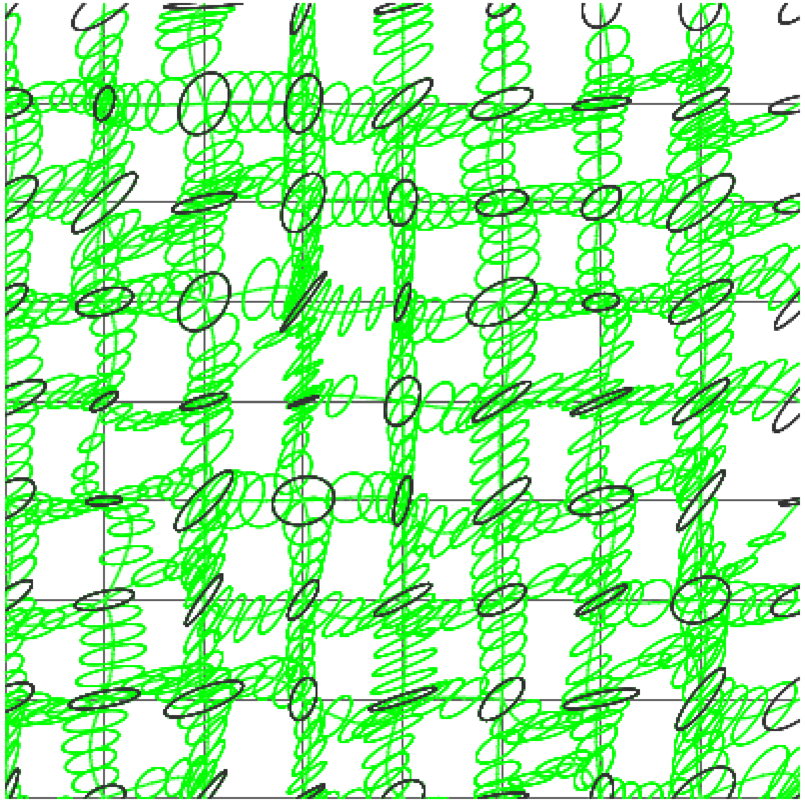
[Cabanes N, 2021]

References

- James: The variance information manifold and the functions on it, Multivariate Analysis–III. Academic Press (1973)
- Skovgaard: A Riemannian geometry of the multivariate normal model, Scandinavian journal of statistics (1984)
- Eriksen: Geodesics connected with the Fisher metric on the multivariate normal manifold, Proceedings of the GST Workshop, Lancaster (1987)
- Calvo and Oller: A distance between multivariate normal distributions based in an embedding into the Siegel group, Journal of multivariate analysis (1990)
- Calvo and Oller: "An explicit solution of information geodesic equations for the multivariate normal model", Statistics & Risk Modeling (1991)
- Kobayashi: Geodesics of multivariate normal distributions and a Toda lattice type Lax pair, arXiv:2304.12575 (2023)
- Nielsen: **A Simple Approximation Method for the Fisher–Rao Distance between Multivariate Normal Distributions**, Entropy (2023)
- Nielsen: **Fisher-Rao and pullback Hilbert cone distances on the multivariate Gaussian manifold with applications to simplification and quantization of mixtures**, ICML TAG-ML (2023)

- Siegel: Symplectic geometry (1964)
- Cabanes and Nielsen: Classification in the Siegel Space for Vectorial Autoregressive Data, GSI (2021)

Thank you!



<u>Open problem</u>:

Closed-form formula for MVN Rao distance?

SPD Riemannian geometry wrt trace metric

- Levi-Civita metric connection

$$\nabla^G_{X_P} Y_P = DY[P][X_P] - \frac{1}{2}\left(X_P P^{-1} Y_P + Y_P P^{-1} X_P\right)$$

Fréchet derivative

$$\gamma_G(P, Q; \alpha) = G_\alpha(P, Q)$$

$$G_\alpha(P, Q) = P^{\frac{1}{2}} \left(P^{-\frac{1}{2}}\, Q\, P^{-\frac{1}{2}}\right)^\alpha P^{\frac{1}{2}}$$

Geodesic arclength parameterization:

$$\rho_{\mathcal{N}}\left(\gamma^{\mathrm{FR}}_{\mathcal{N}}(p_{\lambda_1}, p_{\lambda_2}; s), \gamma^{\mathrm{FR}}_{\mathcal{N}}(p_{\lambda_1}, p_{\lambda_2}; t)\right) = |s - t|\, \rho_{\mathcal{N}}(p_{\lambda_1}, p_{\lambda_2}), \quad \forall s, t \in [0, 1].$$

Matrix Karcher centers as matrix means

- Arithmetic weighted mean matrix $A_\alpha(P,Q) = (1-\alpha)P + \alpha Q$

yields a ∇^{A}-geodesic with respect to metric $g_P^A(X,Y) = \mathrm{tr}(X^\top Y)$(Euclidean)

- Harmonic weighted mean matrix $H_\alpha(P,Q) = ((1-\alpha)P^{-1} + \alpha Q^{-1})^{-1}$

yields a geodesic ∇^H with respect to metric $g_P^H(X,Y) = \mathrm{tr}(P^{-2}XP^{-2}Y)$ (isometric to g, Euclidean)

- Geometric weighted mean matrix $G_\alpha(P,Q) = P^{\frac{1}{2}} \left(P^{-\frac{1}{2}}\, Q\, P^{-\frac{1}{2}}\right)^\alpha P^{\frac{1}{2}}$

yields a geodesic wrt metric $g_P^G(X,Y) = \mathrm{tr}(P^{-1}XP^{-1}Y)$ (Non-positively curved)

- (SPD, g^G, ∇^A, ∇^H) is a dually flat space, is ∇^G Levi-Civita connection

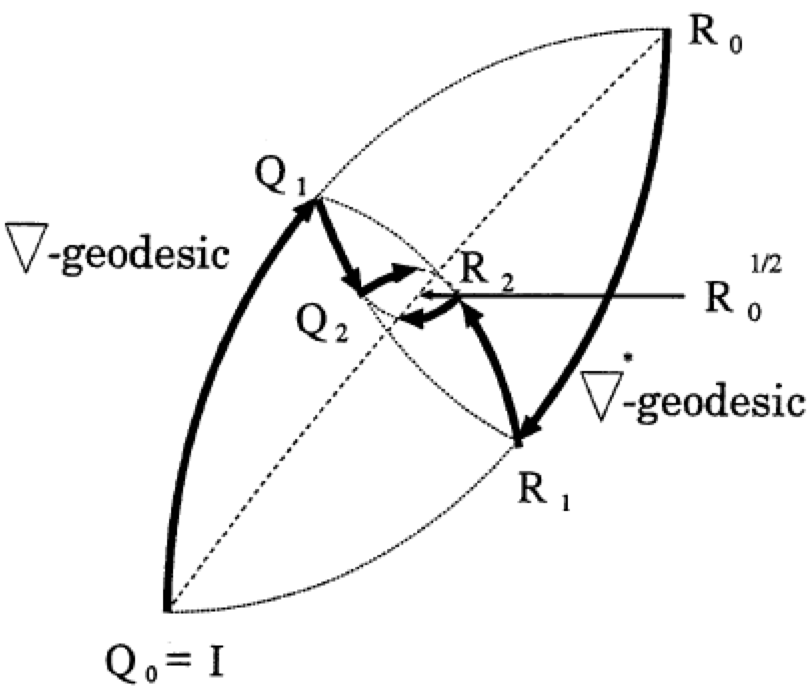


Fig. 2. The matrix AHM algorithm.

$$Q_{n+1} = \tfrac{1}{2}(Q_n + R_n),$$
$$R_{n+1} = 2(Q_n^{-1} + R_n^{-1})^{-1}, \quad n = 0, 1, 2, \ldots .$$

Theorem 9. *The sequences* $\{Q_n\}_{n=0,1,2,\ldots}$ *and* $\{R_n\}_{n=0,1,2,\ldots}$ *with* $Q_0 = I$ *tend to the common limit* $\mathrm{G} = R_0^{1/2}$ *in a quadratic order.*

Theorem 10. *The AHM algorithm on the space* PD(m) *of positive-definite symmetric matrices generates sequences* $\{Q_n\}_{n=0,1,2,\ldots}$ *and* $\{R_n\}_{n=0,1,2,\ldots}$ *which converge quadratically to the midpoint*

$$\mathrm{G} = Q_0^{1/2}(Q_0^{-1/2} R_0 Q_0^{-1/2})^{1/2} Q_0^{1/2} \tag{31}$$

of the Riemannian geodesics from Q_0 *to* R_0.

Siegel upper/disk space: Non-Positive Curvature (NPC)

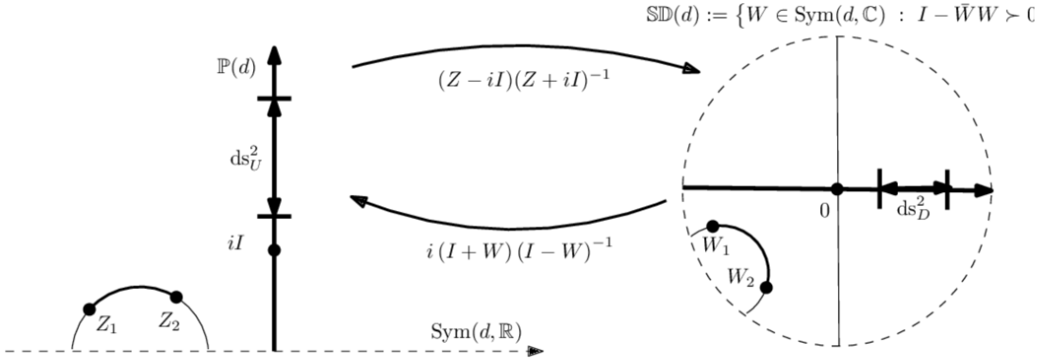
Siegel disk: $SD_N = \{M \in \mathbb{C}^{N\times N}, I - MM^H > 0\}$ $\quad SD_N = \{M \in \mathbb{C}^{N\times N}, \|M\| < 1\}$

$$\|M\| = \sup_{X \in \mathbb{C}^{N\times N}, \|X\|=1} \left(\|MX\|\right)$$

Siegel metric/line element: $ds^2 = \mathrm{trace}\left(\left(I - \Omega\Omega^H\right)^{-1} d\Omega \left(I - \Omega^H \Omega\right)^{-1} d\Omega^H\right)$

Siegel disk distance: $C = (\Psi - \Omega)\left(I - \Omega^H \Psi\right)^{-1}\left(\Psi^H - \Omega^H\right)\left(I - \Omega\Psi^H\right)^{-1}$

$$d^2_{SD_N}(\Omega, \Psi) = \frac{1}{4}\mathrm{trace}\left(\log^2\left(\frac{I + C^{1/2}}{I - C^{1/2}}\right)\right)$$
$$= \mathrm{trace}\left(\mathrm{arctanh}^2\left(C^{1/2}\right)\right)$$



Siegel geodesic: $\zeta(t) : t \mapsto \exp_\Omega(tV)$ $\quad \exp_0(V) = \tanh(Y)\, Y^{-1} V \quad \text{where } Y = \left(VV^H\right)^{1/2}$

Theorem . *The sectional curvature at zero of the plan* σ *defined by* E_1 *and* E_2 :

$$-4 \leqslant K(\sigma) \leqslant 0 \quad \forall \sigma$$

Summary: A (1+ε)-approximation of Rao's distance between multivariate normal distributions

```
ApproxRaoDistMVN(N0,N1,ε>0):

LB=CalvoOllerLowerBound(N0,N1);
UB=SqrtJeffreysUpperBound(N0,N1);
if (UB/LB>1+ε)
        {/* N is midpoint geodesic */
          N=GeodesicMidpoint(N0,N1);
          return ApproxRaoDistMVN(N0,N,ε)+ApproxRaoDistMVN(N,N1,ε);}
            else
          return UB;
```

Instead of exact midpoint, may use the matrix arithmetic-harmonic mean (quadratic convergence)