Geometry and Computer Science

Information Geometry: Background and Applications in Machine Learning

Giovanni Pistone
www.giannidiorestino.it
presenting joint work with
Luigi Malagò
http://www.luigimalago.it/
http://rist.ro/en.html

Pescara (IT), February 8–10, 2017

Introduction

Information Geometry IG is the (differential) geometry of statistical models. IG comes in (at least) 4 flavours:

- If the set of parameters is an open subset of a real finite dimensional vector space, the Fisher-Rao manifold is obtained from the interpretation of the Fisher information matrix as a Riemannian metric.
- If the model is an exponential family (finite or infinite dimensional), then an affine geometry, called exponential manifold is derived thom the linear structure of of the log-likelihood.
- Under regularity conditions, the Wasserstein distance on probability measures is actually a distance of a Wasserstein Riemannian manifold.
- When the state space is finite, one can consider algebraic varieties of probability measures instead of manifolds: it is the so-called Algebraic Statistics approach.

Summary

- PART I is a general overview of an approach to IG based on the notion of statistical bundle
- PART II introduces the Fisher-Rao manifold and the exponential manifold of the centered Gaussian model.
- PART III is a very short discussion of the Fisher-Rao geometry of neural nets.
- PART IV introduces the second order geometry of the full Gaussian model.

PART I

1. Setup: statistical model, exponential family
2. Setup: random variables
3. Fisher-Rao computation
4. Amari's gradient
5. Statistical bundle
6. Why the statistical bundle?
7. Regular curve
8. Statistical gradient
9. Computing grad
10. Differential equations
11. Polarization measure
12. Polarization gradient flow

Setup: statistical model, exponential family

- On a sample space $(\Omega, \mathcal{F})$, with reference probability measure ν, and a parameter' set $\Theta \in \mathbb{R}^d$, we have a statistical model

$$\Omega \times \Theta \ni (x, \theta) \mapsto p(x;\theta) \quad \mathbb{P}_\theta (A) = \int_A p(x;\theta)\ \nu(dx)$$

- For each fixed $x \in \Omega$ the mapping $\theta \mapsto p(x;\theta)$ is the likelihood of x. We routinely assume $p(x;\theta) > 0$, $x \in \Omega$, $\theta \in \Theta$, and define the log-likelihood to be $\ell(x;\theta) = \log p(x;\theta)$.
- The simplest model shows a linear form of the log-likelihood

$$\ell(x;\theta) = \sum_{j=1}^{d} \theta_j T_j(x) - \theta_0$$

The T_j's are the sufficient statistics, and $\theta_0 = \psi(\boldsymbol{\theta})$ is the cumulant generating function. Such a model is called exponential family.

- B. Efron and T. Hastie. *Computer age statistical inference*, volume 5 of *Institute of Mathematical Statistics (IMS) Monographs*. Cambridge University Press, New York, 2016. Algorithms, evidence, and data science

Setup: random variables

- A random variable is a measurable function on $(\Omega, \mathcal{F})$. The space $L^0(\mathbb{P}_\theta)$ of (classes of) random variables does not depend on θ. The space of $L^\infty(\mathbb{P}_\theta)$ of (classes of) bounded random variables does not depend on θ. However, the space $L^\alpha(\mathbb{P}_\theta)$, for any $\alpha \in [0, \infty[$ of $\mathbb{P}_\theta$ of (classes of) integrable random variables does depend on θ!
- For special classes of statistical models and special α's it is possible to assume the equality of spaces of α-integrable random variables.
- In general, it is better to think to the decomposition $L^\alpha(\mathbb{P}_\theta) = \mathbb{R} \oplus L_0^\alpha(\mathbb{P}_\theta)$, $X = \mathbb{E}_{\mathbb{P}_\theta}[X] + (X - \mathbb{E}_{\mathbb{P}_\theta}[X])$ and to extend the statistical model to a bundle $\{(\mathbb{P}_\theta, U) | U \in L_0^\alpha(\mathbb{P}_\theta)\}$.
- Many authors have observed that each fiber of such a bundle is the proper expression of the tangent space of the statistical models seen as a manifold e.g., Phil Dawid (1975).

Fisher-Rao computation

$$
\begin{aligned}
\frac{d}{d\theta}\mathbb{E}_{\mathbb{P}_\theta}[X] &= \frac{d}{d\theta}\sum_{x\in\Omega} X(x)p(x;\theta) \\
&= \sum_{x\in\Omega} X(x)\frac{d}{d\theta}p(x;\theta) \\
&= \sum_{x\in\Omega} X(x)\frac{d}{d\theta}\log\left(p(x;\theta)\right)p(x;\theta) \qquad \text{(check } X=1\text{)} \\
&= \sum_{x\in\Omega} \left(X(x)-\mathbb{E}_{\mathbb{P}_\theta}[X]\right)\frac{d}{d\theta}\log\left(p(x;\theta)\right)p(x;\theta) \\
&= \mathbb{E}_{\mathbb{P}_\theta}\left[\left(X-\mathbb{E}_{\mathbb{P}_\theta}[X]\right)\frac{d}{d\theta}\log\left(p(\theta)\right)\right] \\
&= \left\langle \left(X-\mathbb{E}_{p(\theta)}[X]\right), \frac{d}{d\theta}\log\left(p(\theta)\right)\right\rangle_{p(\theta)}
\end{aligned}
$$

- $Dp_\theta = \frac{d}{d\theta}\log p_\theta$ is the score|velocity of the curve $\theta \mapsto p_\theta$
- C. Radhakrishna Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.*, 37:81–91, 1945

Amari's gradient

- Let $f(p) = f(p(x) \colon x \in \Omega)$ be a smooth function on the open simplex of densities $\Delta^\circ(\Omega)$.

$$\begin{aligned}
\frac{d}{d\theta} f(p_\theta) &= \sum_{x \in \Omega} \frac{\partial}{\partial p(x)} f(p(x;\theta) \colon x \in \Omega) \frac{d}{d\theta} p(x;\theta) \\
&= \sum_{x \in \Omega} \frac{\partial}{\partial p(x)} f(p(x;\theta) \colon x \in \Omega) \frac{\frac{d}{d\theta} p(x;\theta)}{p(x;\theta)} \; p(x;\theta) \\
&= \left\langle \nabla f(p(\theta)), \frac{d}{d\theta} \log p_\theta \right\rangle_{p(\theta)} \\
&= \left\langle \nabla f(p(\theta)) - \mathbb{E}_{\mathbb{P}_\theta} \left[\nabla f(p_\theta) \right], D p_\theta \right\rangle_{p(\theta)}
\end{aligned}$$

- The natural|statistical gradient is

$$\operatorname{grad} f(p) = \nabla f(p) - \mathbb{E}_p \left[\nabla f(p) \right]$$

S.-I. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, feb 1998

Statistical bundle

1. $$B_p = \left\{ U\colon \Omega \to \mathbb{R} \middle| \mathbb{E}_p\left[U\right] = \sum_{x \in \Omega} U(x)\ p(x) = 0 \right\}, \quad p \in \Delta^\circ(\Omega)$$

2. $$\langle U, V \rangle_p = \mathbb{E}_p\left[UV\right] = \sum_{x \in \Omega} U(x)V(x)\ p(x) \quad \text{metric}$$

3. $$S\Delta^\circ(\Omega) = \left\{(p, U) | p \in \Delta^\circ(\Omega), U \in B_p\right\}\ .$$

4. A vector field|estimating function F of the statistical bundle is a section of the bundle i.e.,

   $$F\colon \Delta^\circ(\Omega) \ni p \mapsto (p, F(p)) \in T\Delta^\circ(\Omega)$$

- G. Pistone. Nonparametric information geometry. In F. Nielsen and F. Barbaresco, editors, *Geometric science of information*, volume 8085 of *Lecture Notes in Comput. Sci.*, pages 5–36. Springer, Heidelberg, 2013. First International Conference, GSI 2013 Paris, France, August 28-30, 2013 Proceedings.

Why the statistical bundle?

- The notion of statistical bundle appears as a natural set up for IG, where the notions of score and statistical gradient do not require any parameterization nor chart to be defined.
- The setup based on the full simplex $\Delta(\Omega)$ is of interest in applications to data analysis. Methods based on the simplex lead naturally to the treatment of the infinite sample space case in cases where no natural parametric model is available.
- There are special affine atlases such that the tangent space identifies with the statistical bundle.
- The construction extends to the affine space generated by the simplex, see the paper [1].
- In the statistical bundle there is a natural treatment of differential equations e.g., gradient flow.

[1] L. Schwachhöfer, N. Ay, J. Jost, and H. V. Lê. Parametrized measure models. *Bernoulli*, 2017. Forthcoming paper

Regular curve

Theorem

1. *Let $I \ni t \mapsto p(t)$ be a C^1 curve in $\Delta^\circ(\Omega)$.*

$$\frac{d}{dt}\mathbb{E}_{p(t)}\left[f\right] = \left\langle f - \mathbb{E}_{p(t)}\left[f\right], Dp(t)\right\rangle_{p(t)}, \quad Dp(t) = \frac{d}{dt}\log\left(p(t)\right)$$

2. *Let $I \ni t \mapsto \eta(t)$ be a C^1 curve in $A_1(\Omega)$ such that $\eta(t) \in \Delta(\Omega)$ for all t. For all $x \in \Omega$, $\eta(x;t) = 0$ implies $\frac{d}{dt}\eta(x;t) = 0$.*

$$\frac{d}{dt}\mathbb{E}_{\eta(t)}\left[f\right] = \left\langle f - \mathbb{E}_{\eta(t)}\left[f\right], D\eta(t)\right\rangle_{\eta(t)}$$

$$D\eta(x;t) = \frac{d}{dt}\log\left|\eta(x;t)\right| \quad \textit{if } \eta(x;t) \neq 0, \textit{ otherwise 0.}$$

3. *Let $I \ni t \mapsto \eta(t)$ be a C^1 curve in $A_1(\Omega)$ and assume that $\eta(x;t) = 0$ implies $\frac{d}{dt}\eta(x;t) = 0$. Hence, for each $f\colon \Delta(\Omega) \to \mathbb{R}$,*

$$\frac{d}{dt}\mathbb{E}_{\eta(t)}\left[f\right] = \left\langle f - \mathbb{E}_{\eta(t)}\left[f\right], D\eta(t)\right\rangle_{\eta(t)}$$

Statistical gradient

Definition

1. Given a function $f \colon \Delta^\circ(\Omega) \to \mathbb{R}$, its statistical gradient is a vector field $\Delta^\circ(\Omega) \ni p \mapsto (p, \operatorname{grad} F(p)) \in S\Delta^\circ(\Omega)$ such that for each regular curve $I \ni t \mapsto p(t)$ it holds

$$\frac{d}{dt} f(p(t)) = \left\langle \operatorname{grad} f(p(t)), Dp(t) \right\rangle_{p(t)}, \quad t \in I \ .$$

2. Given a function $f \colon A_1(\Omega) \to \mathbb{R}$, its statistical gradient is a vector field $A_1(\Omega) \ni \eta \mapsto (\eta, \operatorname{grad} f(\eta)) \in TA_1(\Omega)$ such that for each curve $t \mapsto \eta(t)$ with a score Dp, it holds

$$\frac{d}{dt} f(\eta(t)) = \left\langle \operatorname{grad} f(\eta(t)), D\eta(t) \right\rangle_{\eta(t)}$$

Computing grad

1. If f is a C^1 function on an open subset of $\mathbb{R}^\Omega$ containing $\Delta^\circ(\Omega)$, by writing $\nabla f(p)\colon \Omega \ni x \mapsto \frac{\partial}{\partial p(x)} f(p)$, we have the following relation between the statistical gradient and the ordinary gradient:
$$\operatorname{grad} f(p) = \nabla f(p) - \mathbb{E}_p\left[\nabla f(p)\right] \ .$$
2. If f is a C^1 function on an open subset of $\mathbb{R}^\Omega$ containing $A_1(\Omega)$, we have:
$$\operatorname{grad} f(\eta) = \nabla f(\eta) - \mathbb{E}_\eta\left[\nabla f(\eta)\right] \ .$$

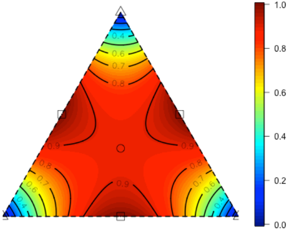
Differential equations

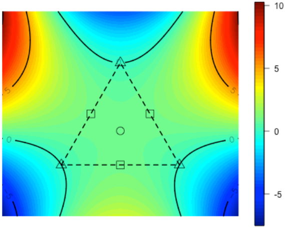
Definition (Flow)

1. Given a vector field $F\colon \Delta^\circ(\Omega)$ or $F\colon A_1(\Omega)$, the trajectories along the vector field are the solution of the (statistical) differential equation
$$\frac{D}{dt}p(t) = F(p(t)) \ .$$
2. A flow of the vector field F is a mapping $S\colon \Delta^\circ(\Omega) \times \mathbb{R}_{>0} \ni (p,t) \mapsto S(p,t) \in \Delta^\circ(\Omega)$, respectively $S\colon A_1(\Omega) \times \mathbb{R}_{>0} \ni (p,t) \mapsto S(p,t) \in A_1(\Omega)$, such that $S(p,0) = p$ and $t \mapsto S(p,t)$ is a trajectory along F.
3. Given $f\colon \Delta^\circ(\Omega) \to \mathbb{R}$, or $f\colon A_1(\Omega) \to \mathbb{R}$, with statistical gradient $p \mapsto (p, \operatorname{grad} f(p)) \in S\Delta^\circ(\Omega)$, respectively $\eta \mapsto (\eta, \operatorname{grad} f(p)) \in SA_1(\Omega)$, a solution of the statistical gradient flow equation, starting at $p_0 \in \Delta^\circ(\Omega)$, respectively $\eta_0 \in A_1(\Omega)$, at time t_0, is a trajectory of the field $-\operatorname{grad} f$ starting at p_0, respectively η_0.

Polarization measure

$$\mathrm{POL}\colon \Delta_n \ni p \mapsto 1 - 4\sum_{x=0}^{n}\left(\frac{1}{2} - p(x)\right)^2 p(x) = 4\sum_{x=0}^{n} p(x)^2(1-p(x)) \ .$$

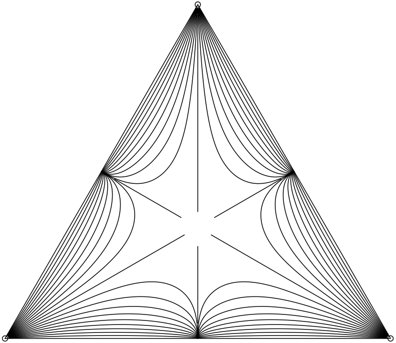


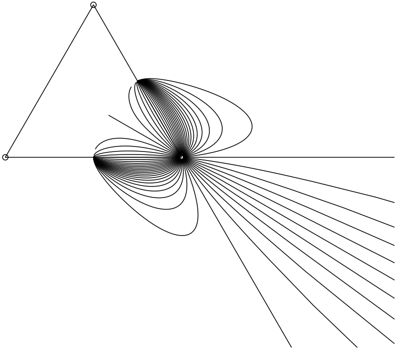


- M. Reynal-Querol. Ethnicity, political systems and civil war. *Journal of Conflict Resolution*, 46(1):29–54, February 2002
- G. Pistone and M. Rogantin. The gradient flow of the polarization measure. with an appendix. arXiv:1502.06718, 2015

Polarization gradient flow

$$\dot{p}(x;t) = p(x;t)\left(8p(x;t) - 12p(x;t)^2 - 8\sum_{y\in\Omega} p(y;t)^2 + 12\sum_{y\in\Omega} p(y;t)^3\right)$$





- L. Malagò and G. Pistone. Natural gradient flow in the mixture geometry of a discrete exponential family. *Entropy*, 17(6):4215–4254, 2015

PART II

1. Gaussian model
2. Fisher-Rao Riemannian manifold
3. Exponential manifold

Gaussian model

- A random variable Y with values in $\mathbb{R}^d$ has distribution $\mathrm{N}\left(\boldsymbol{\mu}, \Sigma\right)$ if $Z = (Z_1, \dots, Z_d)$ is IID $\mathrm{N}\left(0, 1\right)$ and $X = \boldsymbol{\mu} + AZ$ with $A \in \mathrm{M}(d)$ and $AA^* = \Sigma \in \mathrm{Sym}^+\left(d\right)$. Notice the state-space definition.
- We can take for example $A = \Sigma^{1/2}$ or any $A = \Sigma^{1/2}R^*$ with $R^*R = I$.
- If $X \sim \mathrm{N}\left(0, \Sigma_X\right)$, then $Y = TX \sim \mathrm{N}\left(0, T\Sigma_X T^*\right)$, $T \in \mathrm{M}(d)$.
- If $X \sim \mathrm{N}\left(0, \Sigma_X\right)$ and $Y \sim \mathrm{N}\left(0, \Sigma_Y\right)$, then $Y \sim TX$ with

$$T = \Sigma_Y^{1/2} \left(\Sigma_Y^{1/2} \Sigma_X \Sigma_Y^{1/2} \right)^{-1/2} \Sigma_Y^{1/2}$$

- If $\Sigma \in \mathrm{Sym}^{++}\left(d\right) = \mathrm{Sym}^+\left(d\right) \cap \mathrm{Gl}(d)$ then $\mathrm{N}\left(0, \Sigma\right)$ has density

$$p(\boldsymbol{x}; \Sigma) = (2\pi)^{-d/2} \det\left(\Sigma\right)^{-1/2} \exp\left(-\frac{1}{2}\boldsymbol{x}^*\Sigma^{-1}\boldsymbol{x}\right)$$

Fisher-Rao Riemannian manifold I

- The Gaussian model $\mathrm{N}(0,\Sigma)$, $\Sigma \in \mathrm{Sym}^{++}(d)$ is parameterized either by the covariance $\Sigma \in \mathrm{Sym}^{++}(d)$ or by the concentration $C = \Sigma^{-1} \in \mathrm{Sym}^{++}(d)$.
- The vector space of symmetric matrices $\mathrm{Sym}(d)$ has the scalar product $(A,B) \mapsto \langle A,B\rangle_2 = \frac{1}{2}\mathrm{Tr}(AB)$ and $\mathrm{Sym}^{++}(d)$ is an open cone. The log-likelihood in the concentration C is

$$\begin{aligned}\ell(\boldsymbol{x};C) &= \log\left((2\pi)^{-d/2}\det(C)^{1/2}\exp\left(-\frac{1}{2}\boldsymbol{x}^* C\boldsymbol{x}\right)\right)\\ &= -\frac{d}{2}\log(2\pi) + \frac{1}{2}\log\det C - \frac{1}{2}\mathrm{Tr}(C\boldsymbol{x}\boldsymbol{x}^*)\\ &= -\frac{d}{2}\log(2\pi) + \frac{1}{2}\log\det C - \langle C, \boldsymbol{x}\boldsymbol{x}^*\rangle_2\end{aligned}$$

- Fisher's score in the direction $V \in \mathrm{Sym}(d)$ is the directional derivative $d(C \mapsto \ell(\boldsymbol{x};C))[V] = \frac{d}{dt}\ell(\boldsymbol{x};C+tV)\big|_{t=0}$

- J. R. Magnus and H. Neudecker. *Matrix differential calculus with applications in statistics and econometrics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester, 1999. Revised reprint of the 1988 original, §8.3

Fisher-Rao Riemannian manifold II

- As $d\left(C \mapsto \frac{1}{2}\log\det C\right)[V] = \frac{1}{2}\operatorname{Tr}\left(C^{-1}V\right) = \left\langle C^{-1}, V\right\rangle_2$, the Fisher's score is

$$S(\boldsymbol{x}; C)[V] = d(C \mapsto \ell(\boldsymbol{x}; C))[V] = \\ \left\langle C^{-1}, V\right\rangle_2 - \langle V, \boldsymbol{x}\boldsymbol{x}^*\rangle_2 = \left\langle C^{-1} - \boldsymbol{x}\boldsymbol{x}^*, V\right\rangle_2$$

- Notice that $\mathbb{E}_\Sigma\left[C^{-1} - XX^*\right] = C^{-1} - \Sigma = 0$
- The covariance of the Fisher's score in the directions V and W is equal to minus (the expected value of) the second derivative. As $d(C \mapsto C^{-1})[W] = -C^{-1}WC^{-1}$

$$\operatorname{Cov}_{C^{-1}}\left(S(\boldsymbol{x}; C)[V], S(\boldsymbol{x}; C)[W]\right) = -d^2\ell(\boldsymbol{x}; C)[V, W] = \\ \left\langle C^{-1}WC^{-1}, V\right\rangle_2 = \frac{1}{2}\operatorname{Tr}\left(C^{-1}WC^{-1}V\right)$$

- T. W. Anderson. *An introduction to multivariate statistical analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, third edition, 2003

Fisher-Rao Riemannian manifold III

- If we make the same computation with respect to the parameter Σ, because of the special properties of $C \mapsto \Sigma$, we get the same result:

$$\mathrm{Cov}_{\Sigma}\left(S(\boldsymbol{x};\Sigma)[V], S(\boldsymbol{x};\Sigma)[W]\right) = \frac{1}{2}\operatorname{Tr}\left(\Sigma^{-1}W\Sigma^{-1}V\right)$$

- As $\mathrm{Sym}^{++}(d)$ is an open subset of the Hilbert space $\mathrm{Sym}(d)$, then $\mathrm{Sym}^{++}(d)$ is (trivially) a manifold. The velocity $t \mapsto D\Sigma(t)$ of a curve $t \mapsto \Sigma(t)$ is expressed as the ordinary derivative $t \mapsto \dot{\Sigma}(t)$.
- The tangent space of $\mathrm{Sym}^{++}(d)$ is $\mathrm{Sym}(d)$. In fact, a smooth curve $t \mapsto \Sigma(t) \in \mathrm{Sym}^{++}(d)$ has velocity $\dot{\Sigma}(t) \in \mathrm{Sym}(d)$, and, given any $\Sigma \in \mathrm{Sym}^{++}(d)$ and $V \in \mathrm{Sym}(d)$, the curve $\Sigma(t) = \Sigma^{1/2}\exp\left(t\Sigma^{-1/2}V\Sigma^{-1/2}\right)\Sigma^{1/2}$ has $\Sigma(0) = \Sigma$ and $\dot{\Sigma}(0) = V$.
- Each tangent space $T_{\Sigma}\,\mathrm{Sym}^{++}(d) = \mathrm{Sym}(d)$ has a scalar product

$$F_{\Sigma}(U, V) = \frac{1}{2}\operatorname{Tr}\left(\Sigma^{-1}W\Sigma^{-1}V\right), \quad V, W \in T_{\Sigma}\,\mathrm{Sym}^{++}(d)$$

- The metric (family of scalar products) $F = \left\{F_{\Sigma} \middle| \Sigma \in \mathrm{Sym}^{++}(d)\right\}$ defines the Fisher-Rao Riemannian manifold

Fisher-Rao Riemannian manifold IV

- In the Fisher-Rao Riemannian manifold $(\mathrm{Sym}^{++}(d), F)$ the length of the curve $[0,1] \ni t \mapsto \Sigma(t)$ is

$$\int_0^1 dt \sqrt{F_{\Sigma(t)}(\dot{\Sigma}(t), \dot{\Sigma}(t))}$$

- The Fisher-Rao distance between Σ_1 and Σ_2 is the minimal length of a curve connecting the two points. The value of the distance is

$$F(\Sigma_1, \Sigma_2) = \sqrt{\frac{1}{2} \operatorname{Tr}\left(\log\left(\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}\right) \log\left(\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}\right)\right)}$$

- The geodesics from Σ_1 to Σ_2 is

$$\gamma \colon t \mapsto \Sigma_1^{1/2}\left(\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}\right)^t \Sigma_1^{1/2}$$

- R. Bhatia. *Positive definite matrices*. Princeton Series in Applied Mathematics. Princeton University Press, Princeton, NJ, 2007, §6.1

Fisher-Rao Riemannian manifold V

- The velocity of the geodesics is

$$\dot{\gamma}\colon t \mapsto \Sigma_1^{1/2} \left(\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}\right)^t \log\left(\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}\right)\Sigma_1^{1/2}$$

  From that, one checks that the norm of the velocity is constant and equal to the distance.

- The velocity at $t = 0$ is

$$\dot{\gamma}(0) = \Sigma_1^{1/2} \log\left(\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}\right)\Sigma_1^{1/2}$$

  and the equation can be solved for the final point $\Sigma_2 = \gamma(1)$,

$$\Sigma_2 = \Sigma_1^{1/2} \exp\left(\Sigma_1^{-1/2}\dot{\gamma}(0)\Sigma_1^{-1/2}\right)\Sigma_1^{1/2}$$

  so that the geodesics is expressed in terms of the initial point Σ and the initial velocity V by the Riemannian exponential

$$\operatorname{Exp}_{\Sigma}(tV) = \Sigma^{1/2} \exp\left(\Sigma^{-1/2}(tV)\Sigma^{-1/2}\right)\Sigma^{1/2}$$

Exponential manifold I

- An affine manifold is defined by an atlas of charts such that all change-of-charts mappings are affine mappings. Exponential families are affine manifolds if one takes as charts the centered log-likelihood.
- We study the full Gaussian model parameterized by the concentration matrix $C = \Sigma^{-1} \in \mathrm{Sym}^{++}(d)$ as an affine manifold.
- The charts in the exponential atlas $\left\{ s_A \middle| A \in \mathrm{Sym}^{++}(d) \right\}$ are the centered log-likelihoods defined by

$$\begin{aligned} s_A(C) &= (\ell_C - \ell_A) - \mathbb{E}_A\left[\ell_C - \ell_A\right] \\ &= \langle A - C, XX^* \rangle_2 - \left\langle A - C, A^{-1} \right\rangle_2 \end{aligned}$$

- S. Amari and H. Nagaoka. *Methods of information geometry*. American Mathematical Society, Providence, RI, 2000. Translated from the 1993 Japanese original by Daishi Harada, Ch. 2–3
- G. Pistone and C. Sempi. An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. *Ann. Statist.*, 23(5):1543–1561, October 1995
- G. Pistone. Nonparametric information geometry. In F. Nielsen and F. Barbaresco, editors, *Geometric science of information*, volume 8085 of *Lecture Notes in Comput. Sci.*, pages 5–36. Springer, Heidelberg, 2013. First International Conference, GSI 2013 Paris, France, August 28-30, 2013 Proceedings

Exponential manifold II

- We use the scalar product defined on $\operatorname{Sym}(d)$ by $\langle A, B\rangle_2 = \frac{1}{2}\operatorname{Tr}(AB)$, and write $X \otimes X = XX^*$. The chart at A is

$$s_A(C)) = \left\langle A - C, X \otimes X - A^{-1} \right\rangle_2$$

- The image of each s_A is a set of second order polynomials of the type

$$\frac{1}{2}\sum_{i,j=1}^{d}(a_{ij} - c_{ij})(x_i x_j - a^{ij}), \quad A^{-1} = [a^{ij}]_{i,j=1}^{d} \ ,$$

  that is, a second order symmetric polynomial of order 2, without first order terms, with zero expected value at $\mathrm{N}\left(0, A^{-1}\right)$. And vice-versa.

- For each $A \in \operatorname{Sym}^{++}(d)$ the vector space of such polynomials is the model space for the affine manifold in the chart s_A. Such a space is an expression of the tangent space at A if the velocity $DC(0)$ of the curve $t \mapsto C(t)$, $C(0) = A$, is computed as

$$DC(0) = \left.\frac{d}{dt} s_{C(0)}(C(t))\right|_{t=0} = \left\langle \dot{C}(0), C^{-1}(0) - X \otimes X \right\rangle_2$$

Exponential manifold III

- Define the score space at A to be the vector space generated by the image of s_A, namely

$$S_A \operatorname{Sym}^{++}(d) = \left\{ \left\langle V, \boldsymbol{x} \otimes \boldsymbol{x} - A^{-1} \right\rangle_2 \middle| V \in \operatorname{Sym}(d) \right\}$$

- The image of the chart s_A in this vector space is characterized by a $V = A - C,\ C \in \operatorname{Sym}^{++}(d)$.
- Each score space is a fiber of the score bundle $S \operatorname{Sym}^{++}(d)$.
- On each fiber $S_A \operatorname{Sym}^{++}(d)$ we have the scalar product induced by $L^2(\mathrm{N}\left(0, A^{-1}\right)$, namely the Fisher information operator,

$$\begin{aligned} \mathbb{E}_{A^{-1}}\left[V(X)W(X)\right] &= \mathbb{E}_{A^{-1}}\left[\left\langle V, X \otimes X - A^{-1} \right\rangle_2 \left\langle W, X \otimes X - A^{-1} \right\rangle_2\right] \\ &= F_A(V, W) \end{aligned}$$

- The change-of-chart $s_B \circ s_A^{-1} \colon S_A \operatorname{Sym}^{++}(d) \to S_B \operatorname{Sym}^{++}(d)$ is affine with linear part

$$^{e}\mathbb{U}_A^B \colon \left\langle V, X \otimes X - A^{-1} \right\rangle_2 \mapsto \left\langle V, X \otimes X - B^{-1} \right\rangle_2$$

Exponential manifold IV

- Note that the exponential transport ${}^e\mathbb{U}_A^B$ is the identity on the parameter V and it coincides with the centering of a random variable.
- The mixture transport is the dual ${}^m\mathbb{U}_B^A = ({}^e\mathbb{U}_A^B)^*$, hence for each $W \in \mathrm{Sym}\,(d)$,

$$F_B({}^e\mathbb{U}_A^B V, W) = F_A(V, {}^m\mathbb{U}_B^A W)$$

- We have

$$\begin{aligned}
{}^m\mathbb{U}_B^A \left\langle W, X \otimes X - B^{-1} \right\rangle_2 = \\
\left\langle AB^{-1}WB^{-1}A, X \otimes X - A^{-1} \right\rangle_2 = \\
\left\langle B^{-1}WB^{-1}, (AX) \otimes (AX) - A^{-1} \right\rangle_2
\end{aligned}$$

PART III

1. Conditional independence
2. Regression
3. Gaussian regression: the joint density
4. Gaussian regression: the geometry
5. Gaussian regression: comments

Conditional independence

- Given 3 random variables X, Y, Z, we say that X and Y are independent, given Z, if for all bounded $f(X)$ and $\psi(Y)$ we have

  $$\mathbb{E}\left[\phi(X)\psi(Y)|Z\right] = \mathbb{E}\left[\phi(X)|Z\right]\mathbb{E}\left[\psi(Y)|Z\right] \qquad \text{[Product Rule]}$$

  which in turn is equivalent to, for alla bounded $\phi(X)$

  $$\mathbb{E}\left[\psi(Y)|X, Z\right] = \mathbb{E}\left[\psi(Y)|Z\right] \qquad \text{[Sufficiency]}$$

- If moreover the joint distribution of X, Y has a density given Z of the form $p(x, y|z)$ with respect to a product measure on $(\operatorname{supp} X) \times (\operatorname{supp} Y)$, then conditional independence is equivalent to

  $$p(x, y|z) = p_1(x|z)p_2(y|z) \qquad \text{[Factorization]}$$

  and to

  $$p(y|x, z) = p(y|z) \qquad \text{[Sufficiency]}$$

Regression

- Consider now generic random variables X, Y and assume $Z = f(X; \boldsymbol{w})$, where $\boldsymbol{w} \in \mathbb{R}^N$ is a parameter. The σ-algebra generated by X and $f(X; \boldsymbol{w})$ is equal to the σ-algebra generated by $f(X; \boldsymbol{w})$, hence sufficiency holds,

$$\mathbb{E}\left[\psi(Y) | X, f(X; \boldsymbol{w})\right] = \mathbb{E}\left[\psi(Y) | f(X; \boldsymbol{w})\right]$$

  and

$$\mathbb{E}\left[\phi(X)\psi(Y) | f(X; \boldsymbol{w})\right] = \mathbb{E}\left[\phi(X) | f(X; \boldsymbol{w})\right] \mathbb{E}\left[\psi(Y) | f(X; \boldsymbol{w})\right]$$

- R. Pascanu and Y. Bengio. Revisiting natural gradient for deep networks. v7, 2014
- S.-i. Amari. *Information geometry and its applications*, volume 194 of *Applied Mathematical Sciences*. Springer, [Tokyo], 2016
- B. Efron and T. Hastie. *Computer age statistical inference*, volume 5 of *Institute of Mathematical Statistics (IMS) Monographs*. Cambridge University Press, New York, 2016. Algorithms, evidence, and data science

Gaussian regression: the joint density

- For example, assume Y has real values and $Y = f(X; \boldsymbol{w}) + N$, $N \sim \mathrm{N}(0,1)$, X and Y independent, $\boldsymbol{w} \in \mathbb{R}^N$. Then the distribution of Y given $X = x$ is $\mathrm{N}(f(x; \boldsymbol{w}), 1)$, which depend on $f(x; \boldsymbol{w})$. The joint distribution of X and Y is given by

$$\mathbb{E}\left[\phi(X)\psi(Y)\right] = \mathbb{E}\left[\phi(X)\mathbb{E}\left[\psi(Y)|X\right]\right] = \mathbb{E}\left[\phi(X)\int \frac{1}{\sqrt{2\pi}}\psi(y)\mathrm{e}^{-\frac{1}{2}(y-f(x;\boldsymbol{w}))^2}\right]$$

- The joint density (if any) of X and Y is

$$p(x, y; \boldsymbol{w}) = q(x)r(y|f(x; \boldsymbol{w})) = q(x)\left(\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}(y-f(x;\boldsymbol{w}))^2}\right)$$

- The log-density is

$$\ell(x, y; \boldsymbol{w}) = \log\left(q(x)\right) - \frac{1}{2}\log\left(2\pi\right) - \frac{1}{2}(y - f(x; \boldsymbol{w}))^2$$

Gaussian regression: the geometry

- Consider the statistical model $\left\{p(x,y;\boldsymbol{w}) \middle| \boldsymbol{w} \in \mathbb{R}^N\right\}$
- The vector of scores is

$$\nabla(\boldsymbol{w} \mapsto \ell(x,y;\boldsymbol{w})) = (y - f(x;\boldsymbol{w}))\nabla(\boldsymbol{w} \mapsto f(x;\boldsymbol{w}))$$

- The tangent space at $\boldsymbol{w}$ is the space of random variables

$$T_{\boldsymbol{w}} = \operatorname{Span}\left((X - f(X;\boldsymbol{w}))\frac{\partial}{\partial w_j} f(X;\boldsymbol{w}) \middle| j = 1,\dots,N\right)$$

- The Fisher matrix is

$$I(\boldsymbol{w}) = \mathbb{E}_{\boldsymbol{w}}\left[(Y - f(X;\boldsymbol{w}))^2 \nabla f(X;\boldsymbol{w}) \nabla f(X;\boldsymbol{w})^*\right] = \\ \mathbb{E}_{\boldsymbol{w}}\left[\mathbb{E}_{\boldsymbol{w}}\left[(Y - f(X;\boldsymbol{w}))^2 \middle| f(X;\boldsymbol{w})\right] \nabla f(X;\boldsymbol{w}) \nabla f(X;\boldsymbol{w})^*\right] = \\ \mathbb{E}\left[\nabla f(X;\boldsymbol{w}) \nabla f(X;\boldsymbol{w})^*\right]$$

Gaussian regression: comments

- Consider the case of the perceptron with input $\boldsymbol{x} = (x_1, \dots, x_N)$, parameters $\boldsymbol{w} = (w_0, w_1, \dots, w_N) = (w_0, \boldsymbol{w}^1)$, activation function $S(u)$, and

$$f(\boldsymbol{x}; \boldsymbol{w}) = S(\boldsymbol{w}^1 \cdot x - w_0) \quad \nabla f(x; \boldsymbol{w}) = S'(\boldsymbol{w}^1 \cdot x - w_0)(-1, \boldsymbol{x})$$

- The Fisher information is

$$I(\boldsymbol{w}) = \mathbb{E}\left[S'(\boldsymbol{w}^1 \cdot X - w_0)^2 (-1, \boldsymbol{X}) \otimes (-1, \boldsymbol{X})\right] =$$
$$\mathbb{E}\left[S'(\boldsymbol{w}^1 \cdot X - w_0)^2 \begin{bmatrix} 1 & X^* \\ X & XX^* \end{bmatrix}\right]$$

PART IV: Full Gaussian model

1. Riemannian metric
2. Riemannian gradient
3. Levi-Civita covariant derivative
4. Acceleration
5. Geodesics

Riemannian metric

- We parameterize the full Gaussian model $\mathcal{N} = \{\mathrm{N}(\boldsymbol{\mu}, \Sigma)\}$ with $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\Sigma \in \mathrm{Sym}^{++}(d)$. The tangent space at $(\boldsymbol{\mu}, \Sigma)$, is $T_{\boldsymbol{\mu},\Sigma}\mathcal{N} = \mathbb{R}^d \times \mathrm{Sym}(d)$.
- For each couple $(\boldsymbol{u}, U), (\boldsymbol{v}, V) \in T_{\boldsymbol{\mu},\Sigma}\mathcal{N}$ the scalar product of the metric at $(\boldsymbol{\mu}, \Sigma)$ splits:

$$\left\langle (\boldsymbol{u}, U), (\boldsymbol{v}, V) \right\rangle_{\boldsymbol{\mu},\Sigma} = \left\langle \boldsymbol{u}, \boldsymbol{v} \right\rangle_{\boldsymbol{\mu},\Sigma} + \left\langle U, V \right\rangle_{\boldsymbol{\mu},\Sigma}$$

  with

$$\begin{aligned} \left\langle \boldsymbol{u}, \boldsymbol{v} \right\rangle_{\boldsymbol{\mu},\Sigma} &= \boldsymbol{u}^* \Sigma^{-1} \boldsymbol{v} = \mathrm{Tr}\left(\Sigma^{-1} \boldsymbol{v} \boldsymbol{u}^*\right) \\ \left\langle U, V \right\rangle_{\boldsymbol{\mu},\Sigma} &= \frac{1}{2} \mathrm{Tr}\left(U \Sigma^{-1} V \Sigma^{-1}\right) \end{aligned}$$

- L. T. Skovgaard. A Riemannian geometry of the multivariate normal model. *Scand. J. Statist.*, 11(4):211–223, 1984

Riemannian gradient

Given a smooth function $\mathcal{N} \ni (\boldsymbol{\mu}, \Sigma) \mapsto f(\boldsymbol{\mu}, \Sigma) \in \mathbb{R}$ and a smooth curve $t \mapsto (\boldsymbol{\mu}(t), \Sigma(t)) \in \mathcal{N}$,

$$
\begin{aligned}
&\frac{d}{dt} f(\boldsymbol{\mu}(t), \Sigma(t)) \\
&= \dot{\boldsymbol{\mu}}(t)^* \nabla_1 f(\boldsymbol{\mu}(t), \Sigma(t)) + \operatorname{Tr}\left(\nabla_2 f(\boldsymbol{\mu}(t), \Sigma(t)) \dot{\Sigma}(t)\right) \\
&= \dot{\boldsymbol{\mu}}(t)^* \Sigma(t)^{-1} (\Sigma(t) \nabla_1 f(\boldsymbol{\mu}(t), \Sigma(t))) + \\
&\quad \frac{1}{2} \operatorname{Tr}\left(\Sigma(t)^{-1} (2\Sigma(t) \nabla_2 f(\boldsymbol{\mu}(t), \Sigma(t)) \Sigma(t)) \Sigma(t)^{-1} \dot{\Sigma}(t)\right) \\
&= \left\langle (\Sigma(t) \nabla_1 f(\boldsymbol{\mu}(t), \Sigma(t)), 2\Sigma(t) \nabla_2 f(\boldsymbol{\mu}(t), \Sigma(t))), \frac{d}{dt} (\boldsymbol{\mu}(t), \Sigma(t)) \right\rangle_{\boldsymbol{\mu}(t), \Sigma(t)}
\end{aligned}
$$

- The *Riemannian gradient* is

$$
\operatorname{grad} f(\boldsymbol{\mu}, \Sigma) = (\Sigma \nabla_1 f(\boldsymbol{\mu}, \Sigma), 2\Sigma \nabla_2 f(\boldsymbol{\mu}, \Sigma) \Sigma)
$$

- For example, $f(\boldsymbol{\mu}, \Sigma) = \mathbb{E}_{\boldsymbol{\mu}, \Sigma}[f(X)] = \mathbb{E}_{\mathbf{0}, I}\left[f(\Sigma^{-1/2}(X - \boldsymbol{\mu}))\right]$.

Levi-Civita covariant derivative I

- Given a smooth curve $\gamma \colon t \mapsto (\boldsymbol{\mu}(t), \Sigma(t)) \in \mathcal{N}$ and smooth vector fields on the curve $t \mapsto X(t) = (\boldsymbol{u}(t), U(t))$ and $t \mapsto Y(t) = (\boldsymbol{v}(t), V(t))$, we have

$$\frac{d}{dt} \langle X(t), Y(t) \rangle_{\gamma(t)} = \frac{d}{dt} \left(\langle \boldsymbol{u}(t), \boldsymbol{v}(t) \rangle_{\gamma(t)} + \langle U(t), V(t) \rangle_{\gamma(t)} \right) = \frac{d}{dt} \boldsymbol{v}(t)^* \Sigma^{-1}(t) \boldsymbol{u}(t) + \frac{1}{2} \frac{d}{dt} \operatorname{Tr} \left(U(t) \Sigma^{-1}(t) V(t) \Sigma^{-1}(t) \right)$$

- The first term is

$$\frac{d}{dt} \boldsymbol{v}(t)^* \Sigma^{-1}(t) \boldsymbol{u}(t) = \dot{\boldsymbol{v}}(t)^* \Sigma^{-1}(t) \boldsymbol{u}(t) + \boldsymbol{v}(t)^* \Sigma^{-1}(t) \dot{\boldsymbol{u}}(t) - \boldsymbol{v}(t)^* \Sigma^{-1}(t) \dot{\Sigma}(t) \Sigma^{-1}(t) \boldsymbol{u}(t) = \langle \boldsymbol{u}(t), \dot{\boldsymbol{v}}(t) \rangle_{\boldsymbol{\mu}(t), \Sigma(t)} + \langle \dot{\boldsymbol{u}}(t), \boldsymbol{v}(t) \rangle_{\boldsymbol{\mu}(t), \Sigma(t)} + \left\langle \boldsymbol{u}(t), -\frac{1}{2} \dot{\Sigma}(t) \Sigma^{-1}(t) \boldsymbol{v}(t) \right\rangle_{\boldsymbol{\mu}(t), \Sigma(t)} + \left\langle -\frac{1}{2} \dot{\Sigma}(t) \Sigma^{-1}(t) \boldsymbol{u}(t), \boldsymbol{v}(t) \right\rangle_{\boldsymbol{\mu}(t), \Sigma(t)}$$

Levi-Civita covariant derivative II

- We define the first component of the covariant derivative to be

$$\frac{D}{dt}\boldsymbol{w}(t) = \dot{\boldsymbol{w}}(t) - \frac{1}{2}\dot{\Sigma}(t)\Sigma^{-1}(t)\boldsymbol{w}(t)$$

  because

$$\frac{d}{dt}\left\langle \boldsymbol{u}(t), \boldsymbol{v}(t)\right\rangle_{\boldsymbol{\mu}(t),\Sigma(t)} = \left\langle \boldsymbol{u}(t), \frac{D}{dt}\boldsymbol{v}(t)\right\rangle_{\boldsymbol{\mu}(t),\Sigma(t)} + \left\langle \frac{D}{dt}\boldsymbol{u}(t), \boldsymbol{v}(t)\right\rangle_{\boldsymbol{\mu}(t),\Sigma(t)}$$

- If $\boldsymbol{w}(t) = \dot{\boldsymbol{\mu}}(t)$, then the first component of the acceleration of the curve is

$$\frac{D}{dt}\frac{d}{dt}\boldsymbol{\mu}(t) = \ddot{\boldsymbol{\mu}}(t) - \frac{1}{2}\dot{\Sigma}(t)\Sigma^{-1}(t)\dot{\boldsymbol{\mu}}(t)$$

Levi-Civita covariant derivative III

- The derivative of the second term in the splitting is

$$
\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\operatorname{Tr}\left(U(t)\Sigma^{-1}(t)V(t)\Sigma^{-1}(t)\right) = \\
&\qquad \frac{1}{2}\operatorname{Tr}\left(\frac{d}{dt}\left(U(t)\Sigma^{-1}(t)\right)V(t)\Sigma^{-1}(t)\right) + \\
&\qquad \frac{1}{2}\operatorname{Tr}\left(U(t)\Sigma^{-1}(t)\frac{d}{dt}\left(V(t)\Sigma^{-1}(t)\right)\right) = \\
&\frac{1}{2}\operatorname{Tr}\left(\left(\dot{U}(t)\Sigma^{-1}(t) - U(t)\Sigma^{-1}(t)\dot{\Sigma}(t)\Sigma^{-1}(t)\right)V(t)\Sigma^{-1}(t)\right) + \\
&\frac{1}{2}\operatorname{Tr}\left(U(t)\Sigma^{-1}(t)\left(\dot{V}(t)\Sigma^{-1}(t) - V(t)\Sigma^{-1}(t)\dot{\Sigma}(t)\Sigma^{-1}(t)\right)\right) = \\
&\quad \frac{1}{2}\operatorname{Tr}\left(\left(\dot{U}(t) - U(t)\Sigma^{-1}(t)\dot{\Sigma}(t)\right)\Sigma^{-1}(t)V(t)\Sigma^{-1}(t)\right) + \\
&\qquad \frac{1}{2}\operatorname{Tr}\left(U(t)\Sigma^{-1}(t)\left(\dot{V}(t) - V(t)\Sigma^{-1}(t)\dot{\Sigma}(t)\right)\Sigma^{-1}(t)\right)
\end{aligned}
$$

Levi-Civita covariant derivative IV

- A similar expression is obtained from

$$\frac{1}{2}\frac{d}{dt}\operatorname{Tr}\left(\Sigma^{-1}(t)U(t)\Sigma^{-1}(t)V(t)\right)$$

  so that we can define the second component of the covariant derivative to be

$$\frac{D}{dt}W(t)=\dot{W}(t)-\frac{1}{2}\left(W(t)\Sigma^{-1}(t)\dot{\Sigma}(t)+\dot{\Sigma}(t)\Sigma^{-1}(t)W(t)\right)$$

- If $W(t)=\dot{\Sigma}(t)$, the second component of the acceleration is

$$\frac{D}{dt}\frac{d}{dt}\Sigma(t)=\ddot{\Sigma}(t)-\dot{\Sigma}(t)\Sigma^{-1}(t)\dot{\Sigma}(t)$$

Acceleration

- The acceleration of the curve $t \mapsto \gamma(t) = (\boldsymbol{\mu}(t), \Sigma(t))$ has two components,

$$\frac{D}{dt}\frac{d}{dt}\gamma(t) = \left(\frac{D}{dt}\frac{d}{dt}\boldsymbol{\mu}(t), \frac{D}{dt}\frac{d}{dt}\Sigma(t)\right)$$

given by

$$\begin{aligned}\frac{D}{dt}\frac{d}{dt}\boldsymbol{\mu}(t) &= \ddot{\boldsymbol{\mu}}(t) - \frac{1}{2}\dot{\Sigma}(t)\Sigma^{-1}(t)\dot{\boldsymbol{\mu}}(t)\\ \frac{D}{dt}\frac{d}{dt}\Sigma(t) &= \ddot{\Sigma}(t) - \dot{\Sigma}(t)\Sigma^{-1}(t)\dot{\Sigma}(t)\end{aligned}$$

Geodesics I

- Given $A, B \in \mathrm{Sym}^{++}(d)$, the curve

  $$[0,1] \ni t \mapsto \Sigma(t) = A^{1/2}(A^{-1/2}BA^{-1/2})^t A^{1/2}$$

  is known to be the geodesics for the manifold on $\mathrm{Sym}^{++}(d)$ with $\boldsymbol{\mu} = 0$.

- We have

  $$\Sigma^{-1}(t) = A^{-1/2}(A^{-1/2}BA^{-1/2})^{-t}A^{-1/2}$$

  and

  $$\dot{\Sigma}(t) = A^{1/2}\log\left(A^{-1/2}BA^{-1/2}\right)(A^{-1/2}BA^{-1/2})^t A^{1/2} = A^{1/2}(A^{-1/2}BA^{-1/2})^t \log\left(A^{-1/2}BA^{-1/2}\right)A^{1/2}$$

- R. Bhatia. *Positive definite matrices*. Princeton Series in Applied Mathematics. Princeton University Press, Princeton, NJ, 2007

Geodesics II

- We have

$$\begin{aligned}
&\dot{\Sigma}(t)\Sigma^{-1}(t)\dot{\Sigma}(t) = \\
&\quad A^{1/2}\log\left(A^{-1/2}BA^{-1/2}\right)(A^{-1/2}BA^{-1/2})^t A^{1/2}\times \\
&\quad A^{-1/2}(A^{-1/2}BA^{-1/2})^{-t}A^{-1/2}\times \\
&\quad A^{1/2}(A^{-1/2}BA^{-1/2})^t\log\left(A^{-1/2}BA^{-1/2}\right)A^{1/2} = \\
&A^{1/2}\log\left(A^{-1/2}BA^{-1/2}\right)(A^{-1/2}BA^{-1/2})^t\log\left(A^{-1/2}BA^{-1/2}\right)A^{1/2}
\end{aligned}$$

- We have

$$\begin{aligned}
&\ddot{\Sigma}(t) = \frac{d}{dt}A^{1/2}\log\left(A^{-1/2}BA^{-1/2}\right)(A^{-1/2}BA^{-1/2})^t A^{1/2} = \\
&A^{1/2}\log\left(A^{-1/2}BA^{-1/2}\right)(A^{-1/2}BA^{-1/2})^t\log\left(A^{-1/2}BA^{-1/2}\right)A^{1/2}
\end{aligned}$$

Geodesics III

- We have found that $\Sigma(t) = A^{1/2}(A^{-1/2}BA^{-1/2})^t A^{1/2}$ solves the equation $\frac{D}{dt}\frac{d}{dt}\Sigma(t) = 0$. Let us consider the equation $\frac{D}{dt}\frac{d}{dt}\boldsymbol{\mu}(t) = 0$.
- We have

$$\frac{1}{2}\dot{\Sigma}(t)\Sigma^{-1}(t)\dot{\boldsymbol{\mu}}(t) = \frac{1}{2}\left(A^{1/2}\log\left(A^{-1/2}BA^{-1/2}\right)(A^{-1/2}BA^{-1/2})^t A^{1/2}\right) \times \left(A^{-1/2}(A^{-1/2}BA^{-1/2})^{-t}A^{-1/2}\right)\dot{\boldsymbol{\mu}}(t) = \frac{1}{2}A^{1/2}\log\left(A^{-1/2}BA^{-1/2}\right)A^{-1/2}\dot{\boldsymbol{\mu}}(t)$$

- Notice that $A = \Sigma(0)$ and $A^{1/2}\log\left(A^{-1/2}BA^{-1/2}\right)A^{1/2} = \dot{\Sigma}(0)$, hence the equation becomes

$$0 = \frac{D}{dt}\frac{d}{dt}\boldsymbol{\mu}(t) = \ddot{\boldsymbol{\mu}}(t) - \frac{1}{2}\dot{\Sigma}(0)\Sigma^{-1}(0)\dot{\boldsymbol{\mu}}(t)$$