UNIVERSITÀ DEGLI STUDI DI GENOVA

One-day workshop on Algebraic Statistics

Geometries of the Gaussian Model

Giovanni Pistone

www.giannidiorestino.it

DE CASTRO STATISTICS Collegio Carlo Alberto

Jan 23, 2017

Abstract

In Information Geometry, it is possible to define a number of different geometrical structures on the full Gaussian model: the Fisher-Rao Riemannian Manifold (S.T. Skovgaard 1981), the Wasserstein Riemannian Manifold (A. Takatsu 2011), the Exponential and Mixture Affine manifolds (G. Pistone & C. Sempi 1995). We discuss the features of these geometries, including the second order properties (e.g. Hessians), with special emphasis of the Wasserstein case. This turns out to be a special case of a more general set-up introduced in 2001 by R. Otto.

This talk is based on joint work in progress with Luigi Malagò (Rist, Cluj-Napoca, Romania) and Luigi Montrucchio (Collegio Carlo Alberto, Moncalieri, Italy).

- L. T. Skovgaard. A Riemannian geometry of the multivariate normal model. *Scand. J. Statist.*, 11(4):211–223, 1984
- A. Takatsu. Wasserstein geometry of Gaussian measures. *Osaka J. Math.*, 48(4):1005–1026, 2011
- 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
- F. Otto. The geometry of dissipative evolution equations: the porous medium equation. *Comm. Partial Differential Equations*, 26(1-2):101–174, 2001

Summary

1. Gaussian model
2. Fisher-Rao manifold
3. Exponential manifold
4. Wasserstein manifold
5. Gradient (short!)
6. Covariant derivative (very short!)

Gaussian model

- A random variable Y with values in $\mathbb{R}^d$ has distribution $\mathrm{N}(\boldsymbol{\mu}, \Sigma)$ if $Z = (Z_1, \dots, Z_d)$ is IID $\mathrm{N}(0,1)$ and $X = \boldsymbol{\mu} + AZ$ with $A \in \mathrm{M}(d)$ and $AA^* = \Sigma \in \mathrm{Sym}^+(d)$. 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 $X \sim \mathrm{N}(0, \Sigma_X)$, then $Y = TX \sim \mathrm{N}(0, T\Sigma_X T^*)$, $T \in \mathrm{M}(d)$.
- If $X \sim \mathrm{N}(0, \Sigma_X)$ and $Y \sim \mathrm{N}(0, \Sigma_Y)$, then $X = TY$ 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 $\Sigma \in \mathrm{Sym}^{++}(d) = \mathrm{Sym}^+(d) \cap \mathrm{Gl}(d)$ then $\mathrm{N}(0, \Sigma)$ has density

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

Fisher-Rao manifold I

- The Gaussian model $\mathrm{N}(0, \Sigma)$, $\Sigma \in \mathrm{Sym}^{++}(d)$ is parameterised 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 $(A, B) \mapsto \langle A, B \rangle_2 = \frac{1}{2}\operatorname{Tr}(AB)$ and $\mathrm{Sym}^{++}(d)$ is an open cone. The log-likelihood in the concentration C is

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

- 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 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 $d(C \mapsto C^{-1})[W] = -C^{-1}WC^{-1}$

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

- 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 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 extressed 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 manifold IV

- In the Fisher-Rao Riemannian manifold $(\text{Sym}^{++}(d), F)$ 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 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 paramerised 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-likelyhoods defined by

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

- 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 $\mathrm{Sym}\,(d)$ 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 viceversa.

- For each $A \in \mathrm{Sym}^{++}(d)$ 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 characterised 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

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

W-manifold: Gini's dissimilarity

- Given $\Sigma_1, \Sigma_2 \in \mathrm{Sym}^{++}(d)$, define

$$\Gamma(\Sigma_1, \Sigma_2) = \left\{ \Sigma \in \mathrm{Sym}^{++}(2d) \middle| \Sigma = \begin{bmatrix} \Sigma_1 & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_2 \end{bmatrix} \right\}$$

- Given $(X, Y) \sim \mathrm{N}_{2d}(0, \Sigma)$,

$$\Sigma \in \Gamma(\Sigma_1, \Sigma_2) \quad \Leftrightarrow \quad X \sim \mathrm{N}(0, \Sigma_1) \wedge Y \sim \mathrm{N}(0, \Sigma_2)$$

- We look for the index of dissimilarity defined by

$$W(\Sigma_1, \Sigma_2) = \inf_{\Sigma \in \Gamma(\Sigma_1, \Sigma_2)} \mathbb{E}_\Sigma \left[\|X - Y\|^2 \right]$$

- Notice that

$$\mathbb{E}_\Sigma \left[\|X - Y\|^2 \right] = \mathrm{Tr}(\Sigma_1) + \mathrm{Tr}(\Sigma_2) - 2\,\mathrm{Tr}(\Sigma_{12})$$

W-manifold: An equivalent problem

- If $\Sigma_1, \Sigma_2 \in \mathrm{Sym}^{++}(d)$, then

$$\begin{bmatrix} \Sigma_1 & K \\ K^* & \Sigma_2 \end{bmatrix} \in \mathrm{Sym}^{+}(2d) \iff \Sigma_1 - K^* \Sigma_2^{-1} K \in \mathrm{Sym}^{+}(d)$$

- We can consider the problem

$$\gamma = \min_K -2\,\mathrm{Tr}(K)$$
$$\Sigma_1 - K^* \Sigma_2^{-1} K \in \mathrm{Sym}^{+}(d)$$

- A feasible K is such that the Shur complement is zero:

$$\Sigma_1 - K^* \Sigma_2^{-1} K$$

The unique symmetric solution is

$$K = \Sigma_1^{1/2} (\Sigma_1^{1/2} \Sigma_2^{-1} \Sigma_1^{1/2})^{-1/2} \Sigma_1^{1/2}$$

W-manifold: Linear programming I

- Write $\boldsymbol{E} = \text{Sym}\,(2d)$ and $\boldsymbol{F} = \text{Sym}\,(d) \times \text{Sym}\,(d)$; $P_1 = \begin{bmatrix} I_d & 0_d \end{bmatrix}$, and $P_2 = \begin{bmatrix} 0_d & I_d \end{bmatrix}$ and define the marginalization operator as

$$A\colon E \ni \Sigma \mapsto (P_1 \Sigma P_1^*, P_2 \Sigma P_2^*) \in F$$

- We have

$$\mathbb{E}_\Sigma\left[\langle X, Y\rangle\right] = \mathbb{E}_\Sigma\left[\sum_{i=1}^{d} X_i Y_i\right] = \sum_{i=1}^{d} (\Sigma_{12})_{ii} = \text{Tr}\,(\Sigma_{12}) =$$

$$\text{Tr}\,(P_1 \Sigma P_2^*) = \text{Tr}\left(\frac{1}{2}(P_2^* P_1 + P_1^* P_2)\Sigma\right) = \langle \Sigma, P_2^* P_1 + P_1^* P_2 \rangle_{\boldsymbol{E}}$$

- The problem becomes the canonical probelm

$$\gamma = \inf_{\Sigma \in \boldsymbol{E}} \langle \Sigma, -(P_2^* P_1 + P_1^* P_2)\rangle_{\boldsymbol{E}}$$
$$A(\Sigma) = (\Sigma_1, \Sigma_2)$$
$$\Sigma \geq_{\text{Sym}^+(2d)} 0$$

- The canonical problem is feasible: take $\Sigma = \text{diag}\,(\Sigma_1, \Sigma_2)$.

W-manifold: Linear programming II

- The adjoint $A^*\colon \boldsymbol{F} \to \boldsymbol{E}$ is defined by

$$\langle A^*(F_1, F_2), C\rangle_{\boldsymbol{E}} = \langle (F_1, F_2), A(C)\rangle_{\boldsymbol{F}}$$

- We have

$$\begin{aligned}\langle (F_1, F_2), A(C)\rangle_{\boldsymbol{F}} &= \frac{1}{2}\operatorname{Tr}(F_1 P_1 C P_1^*) + \frac{1}{2}\operatorname{Tr}(F_2 P_2 C P_2^*)\\ &= \frac{1}{2}\operatorname{Tr}((P_1^* F_1 P_1 + P_2^* F_2 P_2) C)\\ &= \langle P_1^* F_1 P_1 + P_2^* F_2 P_2, C\rangle_{\boldsymbol{E}}\end{aligned}$$

hence

$$A^*(F_1, F_2) = P_1^* F_1 P_1 + P_2^* F_2 P_2 = \operatorname{diag}(F_1, F_2)$$

- The dual problem is

$$\begin{aligned}\beta = sup_{(F_1,F_2)\in \boldsymbol{F}} &\langle (\Sigma_1, \Sigma_2), (F_1, F_2)\rangle_{\boldsymbol{F}}\\ A^*(F_1, F_2) &\leq_{\mathrm{Sym}^+(2d)} -(P_2^* P_1 + P_1^* P_2)\end{aligned}$$

W-manifold: Value of the dissimilarity

- The dual problem is

$$\beta = sup_{(F_1,F_2)\in \boldsymbol{F}} \left(\mathrm{Tr}\left(\Sigma_1 F_1\right) + \mathrm{Tr}\left(\Sigma_2 F_2\right)\right)$$
$$\begin{bmatrix} (-F_1) & I \\ I & (-F_2) \end{bmatrix} \in \mathrm{Sym}^+\left(2d\right)$$

- It holds $\gamma = \beta$
- The optimal value is

$$W(\Sigma_1, \Sigma_2)^2 = \mathrm{Tr}\left(\Sigma_1\right) + \mathrm{Tr}\left(\Sigma_2\right) - 2\,\mathrm{Tr}\left((\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\right)$$

- D. C. Dowson and B. V. Landau. The Fréchet distance between multivariate normal distributions. *J. Multivariate Anal.*, 12(3):450–455, 1982
- C. R. Givens and R. M. Shortt. A class of Wasserstein metrics for probability distributions. *Michigan Math. J.*, 31(2):231–240, 1984

Wasserstein Riemannian manifold I

- In $\mathrm{M}(d) = \mathbb{R}^{d \times d}$ with scalar product

$$(A, B) \mapsto \langle A, B \rangle_2 = \frac{1}{2}\operatorname{Tr}(AB^*) = \frac{1}{2}\operatorname{Tr}(B^*A) = \frac{1}{2}\operatorname{Tr}(BA^*) = \frac{1}{2}\operatorname{Tr}(A^*B)$$

  the symmetric matrices $\operatorname{Sym}(d)$ i.e., $A^* = A$, form a vector subspace whose orthogonal complement is the space of antisymmetric matrices i.e., $A^* = -A$.

- We recall that for $A, B \in \mathrm{M}(d)$ we have the vectorized form

$$\langle A, B \rangle_2 = \frac{1}{2}\mathbf{vec}(A)^* \mathbf{vec}(B)$$
$$\mathbf{vec}(AB) = (I \otimes A)\mathbf{vec}(B) = (B \otimes I)\mathbf{vec}(A)$$

- 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, §2.4

Wasserstein Riemannian manifold II

- The set $\operatorname{Sym}^{++}(d) \subset \operatorname{Sym}(d)$ of positive definite matrices is an open convex cone.
- As a sub-manifold of Sym d its tangent bundle is

$$T \operatorname{Sym}^{++}(d) = \left\{(V, Z) \middle| V \in \operatorname{Sym}^{++}(d), Z \in \operatorname{Sym}(d)\right\} .$$

- The immersion $\operatorname{Sym}^{++}(d) \to \operatorname{Sym}(d)$ induces on each tangent space $T_C \operatorname{Sym}^{++}(d)$ the (trivial) metric.

$$(C, H, K) \mapsto \langle H, K \rangle_C = \langle H, K \rangle_2 , \quad H, K \in \operatorname{Sym}(d)$$

- We are going to use a different contruction as in the example

$$f : \mathbb{R}^2 \setminus (0, 0) \ni (x, y) \mapsto x^2 + y^2 \in]0, +\infty[$$

Wasserstein Riemannian manifold III

- Let $f\colon H \to \mathcal{N}$ be a smooth surjection of from Hilbert space H onto a manifold $\mathcal{N}$. Assume that for each $A \in H$ the tangent mapping at A, $df(A)\colon H \to T_{f(A)}\mathcal{N}$, is surjective.
- In such a case, for each $C \in \mathcal{N}$, the fiber $f^{-1}(C)$ is a submanifold.
- Given a point $A \in f^{-1}(C)$, a vector $U \in H$ is *vertical* if it is tangent to the manifold $f^{-1}(A)$. Each such a tangent vector U is the velocity at $t = 0$ of some smooth curve $t \mapsto \gamma(t)$ with $\gamma(0) = A$ and $\dot{\gamma}(0) = U$. Precisely, from $f(\gamma(t)) = C$ for all t we derive the characterisation of vertical vectors. We have $d_A f(A) = 0$ i.e., the tangent space at A is $T_A f^{-1}(f(A)) = \mathrm{Ker}(df(A))$. Consider the orthogonal space to the tangent space $T_A f^{-1}(f(A))$. Such a space is called the space of *horizontal* vectors.

$$\mathcal{H}_A = \mathrm{Ker}(df(A))^{\perp} = \mathrm{Im}\left(df(A)^*\right)\ .$$

- The notion of *submersion* is discussed in M. P. do Carmo. *Riemannian geometry*. Mathematics: Theory & Applications. Birkhäuser Boston Inc., Boston, MA, 1992. Translated from the second Portuguese edition by Francis Flaherty, Ch. 8, Ex. 8–10, or S. Lang. *Differential and Riemannian manifolds*, volume 160 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, third edition, 1995, §II.2

Wasserstein Riemannian manifold IV

- The general linear group $\mathrm{Gl}(d)$ is an open subset of $\mathrm{M}(d)$, which is an Hilbert space of dimension $d \times d$ with the scalar product

$$\langle X, Y \rangle_2 = \frac{1}{2} \operatorname{Tr}(Y^* X)$$

- The mapping

$$\sigma \colon \ \mathrm{Gl}(d) \subset \mathrm{M}(d) \to \mathrm{Sym}^{++}(d) \subset \mathrm{M}(d)$$

defined by

$$A \mapsto \sigma(A) = AA^*$$

has an obvious meaning for Gaussian distributions: it is the computation of the covariance matrix $\Sigma = AA^*$ of the random vector $X = AZ$ when $Z \sim \mathrm{N}(0, I)$.

- As the mapping σ is not 1-to-1 we cannot use A as a parameter. The choice $A = \Sigma^{1/2}$ does not lead to the metric we want.

Wasserstein Riemannian manifold V

- We say that the submersion f is Riemannian if for all A the linear map $df(A) \colon \mathcal{H}_A \to T_C\mathcal{N}$ is an isometry i.e.,

$$U, V \in \mathcal{H}_A \Rightarrow \langle d_U f(A), d_V f(A) \rangle_{f(A)} = \langle U, V \rangle_H \ .$$

- As a linear isometry is 1-to-1 we can write

$$X, Y \in T_C\mathcal{N} \Rightarrow \langle X, Y \rangle_C = \left\langle df(f^{-1}(C))\big|_{\mathcal{H}_{f^{-1}(C)}}^{-1} X, df(f^{-1}(C))\big|_{\mathcal{H}_{f^{-1}(C)}}^{-1} Y \right\rangle_H \ .$$

- A Riemannian submersion preserves the length of curves. Let $[0,1] \ni t \mapsto \gamma(t)$ be a smooth curve in H and consider its image $[0,1] \ni t \mapsto f(\gamma(t))$. The velocity of the image is $t \mapsto df(\gamma(t))[\dot\gamma(t)]$ and its length is

$$\int_0^1 dt \ \langle df(\gamma(t))[\dot\gamma(t)], df(\gamma(t))[\dot\gamma(t)] \rangle_{f(\gamma(t))}^{1/2} = \int_0^1 dt \ \|\dot\gamma(t)\|_H$$

Wasserstein Riemannian manifold VI

- The us derive the Wasserstein Riemannian metric by letting the mapping $A \mapsto AA^*$, A invertible matrix, to be a Riemannian submersion.
- For each $A \in \mathrm{Gl}(d)$ the matrix $\Sigma = AA^*$ belongs in $\mathrm{Sym}^{++}(d)$ and, viceversa, each element of $\mathrm{Sym}^{++}(d)$ has such a presentation.
- The mapping $\sigma \colon \mathrm{Gl}(d) \to \mathrm{Sym}^{++}(d)$ given by $\sigma(A) = AA^*$ has derivative at A in the direction $X \in \mathrm{M}(d)$ given by

$$d_X \sigma(A) = XA^* + AX^*$$

- In vectorized form, we can write

$$d_X \sigma(A) = XA^* + AX^* = \mathbf{bind}\left((A \otimes I)\,\mathbf{vec}\,(X) + (I \otimes A)\,\mathbf{vec}\,(X^*)\right)$$

where **bind** is the inverse of **vec**.

Wasserstein Riemannian manifold VII

- Let us discuss the problem of defining a metric such as σ is a Riemannian submersion. The mapping $\sigma\colon \mathrm{Gl}(d)$ is onto $\mathrm{Sym}^{++}(d)$, which is an open subset of $\mathrm{Sym}\,(d)$, precisely the interior of the cone $\mathrm{Sym}^{+}(d)$. The vector space $\mathrm{Sym}\,(d)$ is a sub-vector space of $\mathrm{M}(d)$ with dimension $\frac{1}{2}d(d+1)$ that inherits the Hilbert structure of the super-space.
- Consider the matrix A as a point in the fiber manifold $\sigma^{-1}(AA^*)$. The derivative of σ at A in the direction $X \in \mathrm{M}(d)$ is the symmetric matrix:
$$d\sigma(A)[X] = XA^* + AX^* \in \mathrm{Sym}\,(m) \ .$$
- The linear mapping $X \mapsto XA^* + AX^*$ is surjective, because for each $W \in \mathrm{Sym}\,(d)$ we can define $X = \frac{1}{2}W(A^*)^{-1}$ to satisfy the equation $XA^* + AX^* = W$ is true. hence, the fiber $\sigma^{-1}(AA^*)$ is a submanifold of $\mathrm{Gl}(d)$.
- Let us compute the splitting od $\mathrm{M}(d)$ into the kernel of $d\sigma(A)$ and the horizontal vectors,
$$\mathrm{M}(d) = \mathrm{Ker}(d\sigma(A)) \oplus \mathcal{H}_A$$

Wasserstein Riemannian manifold VIII

- As the vector space tangent to $\sigma^{-1}(AA^*)$ at A is the kernel of the derivative at A:

$$T_A\sigma^{-1}(AA^*) = \mathrm{Ker}(d(A \mapsto \sigma(A))[X]) = \\ \{X \in \mathrm{M}(d) | XA^* + AX^* = 0\} = \{X \in \mathrm{M}(d) | (AX^*)^* = -AX^*\} \ ,$$

  it consists of all matrices A such that AX^* is anti-symmetric.

- A matrix W is horizontal at A if, and only if for each vertical $X \in T_A\sigma^{-1}(AA^*)$ we have

$$0 = \langle W, X \rangle_2 = \frac{1}{2}\mathrm{Tr}\,(X^*W) = \frac{1}{2}\mathrm{Tr}\left(AX^*WA^{-1}\right) = \\ \frac{1}{2}\mathrm{Tr}\left((XA^*)^*(WA^{-1})\right) = \left\langle WA^{-1}, XA^* \right\rangle_2 \ ,$$

  or, equivalently, for each X such that XA^* is anti-symmetric.

- In conclusion, the vector space of horizontal vectors is

$$\mathcal{H}_A = (T_A\sigma^{-1}(AA^*))^{\perp} = \left\{ W \in \mathrm{M}(d) \middle| WA^{-1} \in \mathrm{Sym}\,(d) \right\} \ .$$

Wasserstein Riemannian manifold IX

- Let $X \in \mathrm{M}(d)$ and consider the decomposition of $X = X_V + X_H$ with X_V vertical at A and X_H horizontal at A. Then $d\sigma(A)[X] = d\sigma(A)[X_H]$ and the restriction of the derivative $d\sigma(A)$ to the vector space $\mathcal{H}_A$ of horizontal vectors at A is 1-to-1 onto the tangent space of $\mathrm{Sym}^{++}(d)$ at AA^*, that is $\mathrm{Sym}\,(d)$.
- In such a restriction we have for each $W \in \mathcal{H}_A$

$$U = d\sigma(A)[W] = WA^* + AW^* = WA^{-1}AA^* + A(WA^{-1}A)^*$$
$$= (WA^{-1})AA^* + AA^*(WA^{-1})^* = (WA^{-1})AA^* + AA^*(WA^{-1}) \ ,$$

so that the inverse mapping of the restriction is given by

$$W = \left(d\sigma(A)|_{\mathcal{H}_A}\right)^{-1}(U) = L(U; AA^*)A \ ,$$

where $L = L(U; C)$ is the solution of the Liapunov equation

$$V = LC + CL, \quad V, L \in \mathrm{Sym}\,(d)\,, C \in \mathrm{Sym}^{++}(d) \ .$$

Wasserstein Riemannian manifold X

- The integral form solution is

$$L(V;C) = \int_0^\infty dt\ \mathrm{e}^{-tC} V \mathrm{e}^{-tC}\ .$$

- In vectorized form the Liapunov equation is

$$\mathbf{vec}\,(V) = (C \otimes I + I \otimes C)\,\mathbf{vec}\,(L)\ ,$$

hence the solution is

$$L(V;C) = \mathbf{bind}\left((C \otimes I + I \otimes C)^{-1}\,\mathbf{vec}\,(V)\right)\ .$$

- A solution based on the spectral decomposition $C = U\Lambda U^*$, $\Lambda = \mathrm{diag}\,(\lambda_j \colon j = 1, \dots, d)$ and $U * U = I$. The solution in the U basis is

$$(U^* L U) = \left[\frac{1}{\lambda_i + \lambda_j}\right]_{i,j=1}^{d} \circ (U^* V U)$$

- R. Bhatia. *Positive definite matrices*. Princeton Series in Applied Mathematics. Princeton University Press, Princeton, NJ, 2007, Ex. 1.2.10.

Wasserstein Riemannian manifold XI

- The mapping

$$\sigma\colon \mathcal{H}_A \ni B \mapsto BB^* \in \mathrm{Sym}^{++}(d)\,, \quad AA^* = \Sigma \in \mathrm{Sym}^{++}(d)$$

is actually globally invertible. For each $C \in \mathrm{Sym}^{++}(d)$, the solution of

$$C = BB^* = (BA^{-1}A)(BA^{-1}A)^* = (BA^{-1})\Sigma(BA^{-1})^*, (BA^{-1}) \in \mathrm{Sym}(d)\,,$$

is the solution of a Riccati equation,

$$B = C^{1/2}(C^{1/2}\Sigma C^{1/2})^{-1/2})C^{1/2}A\;.$$

- Let us push-forward the scalar product on $\mathcal{H}_A$ to $T_{AA^*}\mathrm{Sym}^{++}(d)$ as

$$W_{AA^*}(U,V) = \left\langle \left(d\sigma(A)|_{\mathcal{H}_A}\right)^{-1}(U), \left(d\sigma(A)|_{\mathcal{H}_A}\right)^{-1}(V)\right\rangle_2 =$$
$$\langle L(U;AA^*)A, L(V;AA^*)A\rangle_2 = \frac{1}{2}\mathrm{Tr}\,(A^*L(V;AA^*)L(U;AA^*)A) =$$
$$\frac{1}{2}\mathrm{Tr}\,(L(V;AA^*)AA^*L(U;AA^*))\;\;,$$

which depends on AA^* only.

Wasserstein Riemannian manifold XII

- We can take $A = \Sigma^{1/2}$ and see that, in such a case, $W \in \mathcal{H}_A$, that is $WA^{-1} = W\Sigma^{1/2} \in \mathrm{Sym}\,(d)$ and, so that

$$U_i = L(U_i; \Sigma)\Sigma + \Sigma L(U_i; \Sigma), \quad i = 1, 2\ , \tag{1}$$

$$W_\Sigma(U_1, U_2) = \frac{1}{2}\operatorname{Tr}\left(L(U_2; \Sigma)\Sigma L(U_1; \Sigma)\right)\ . \tag{2}$$

- Consider the mapping $U \mapsto L(U; \Sigma)\Sigma^{1/2}$. It maps the scalar product w_Σ to $\langle \cdot, \cdot \rangle_2$:

$$w_\Sigma(U_1, U_2) = \left\langle L(U_1; \Sigma)\Sigma^{1/2}, L(U_2; \Sigma)\Sigma^{1/2} \right\rangle_2$$

Wasserstein Riemannian manifold XIII

- Let us construct now a Wasserstein geodesics connecting two matrices $\Sigma_0, \Sigma_1 \in \mathrm{Sym}^{++}(d)$. Define the symmetric matrix
$$T = \Sigma_1^{1/2}(\Sigma_1^{1/2}\Sigma_0\Sigma_1^{1/2})^{-1/2}\Sigma_1^{1/2} \ .$$
The matrix T is the unique solution in $\mathrm{Sym}^{+}(d)$ of the Riccati equation $T\Sigma_0 T = \Sigma_1$.
- We define a curve in $\mathrm{Sym}^{++}(d)$ connecting Σ_0 and Σ_1 as follows. First we define
$$A_0 = \Sigma_0^{1/2}, A_1 = (T - I)\Sigma_0^{1/2},$$
so that $A_0, A_1 \in \mathcal{H}_{\Sigma_0^{1/2}}$ because $A_0(\Sigma_0^{1/2})^{-1} = I \in \mathrm{Sym}(d)$ and $A_1(\Sigma_0^{1/2})^{-1} = T - I \in \mathrm{Sym}(d)$. It follows that the the strait line from A_0 to A_1 belongs to the vector space of horizontal vectors at $\Sigma_0^{1/2}$,
$$[0,1] \ni t \mapsto A(t) = A_0 + tA_1 \in \mathcal{H}_{\Sigma_0^{1/2}}, \quad t \in \mathbb{R} \ .$$
and it is a geodesics in $\mathrm{M}(d)$.
- As a consequence, $t \mapsto \Sigma(t) = A(t)A^*(t)$ is a geodesics in the Wasserstein metric connecting Σ_0 to Σ_1.

Wasserstein Riemannian manifold XIV

- In conclusion, the curve

$$t \mapsto \Sigma(t) = A(t)A(t)^* = (I+t(T-I))\Sigma_0(I+t(T-I)) \in \mathrm{Sym}^{++}(d)$$

connects $\Sigma_0 = \Sigma(0)$ to $\Sigma_1 = \Sigma(1)$ and has minimal length.

- Let us compute the length of the the geodesic $t \mapsto A(t)$, $t \in [0,1]$, which is equal to the Wasserstein distance of Σ_0 and Σ_1. We have

$$\left\|\dot{A}(t)\right\|_2 = \sqrt{\frac{1}{2}\operatorname{Tr}\left(\dot{A}(t)(\dot{A}(t))^*\right)} = \sqrt{\frac{1}{2}\operatorname{Tr}\left((T-I)\Sigma_0(T-I)\right)} =$$

$$\sqrt{\frac{1}{2}\left(\operatorname{Tr}(\Sigma_0) + \operatorname{Tr}(\Sigma_1) - \operatorname{Tr}(T\Sigma_0) - \operatorname{Tr}(\Sigma_0 T)\right)} =$$

$$\sqrt{\frac{1}{2}\left(\operatorname{Tr}(\Sigma_0) + \operatorname{Tr}(\Sigma_1) - \operatorname{Tr}\left((\Sigma_1^{1/2}\Sigma_0\Sigma_1^{1/2})^{1/2}\right)\right)} .$$

Wasserstein Riemannian manifold XV

- Let us compute the velocity of the geodesics $t \mapsto \Sigma(t)$:

$$\frac{d}{dt}\Sigma(t) = (T-I)\Sigma_0 + \Sigma_0(T-I) + 2t(T-I)\Sigma_0(T-I) \ ,$$

in particular

$$\dot{\Sigma}(0) = (T-I)\Sigma_0 + \Sigma_0(T-I)$$

- Recall that the linear map $\operatorname{Sym}(d) \ni A \mapsto A\Sigma_0 + \Sigma_0 A \in \operatorname{Sym}(d)$ is injective, hence surjective, and that we denote by $L(\cdot;\Sigma_0)$ the inverse map. From the first Eq. above, we have $T - I = L(\dot{\Sigma}(0);\Sigma_0)$, and hence

$$\begin{aligned}\Sigma(t) &= \Sigma_0 + t((T-I)\Sigma_0 + \Sigma_0(T-I)) + t^2(T-I)\Sigma_0(T-I) \\ &= \Sigma(0) + t\dot{\Sigma}(0) + t^2 L(\dot{\Sigma}(0);\Sigma(0))\Sigma(0)L(\dot{\Sigma}(0);\Sigma(0)) \ .\end{aligned}$$

- Given $V \in \operatorname{Sym}(d)$ and $C \in \operatorname{Sym}^{++}(d)$ we define the Riemannian exponential to be

$$\operatorname{Exp}_C(V) = C + V + L(V;C)CL(V;C) \ ,$$

so that the geodesics is $\Sigma(t) = \operatorname{Exp}_{\Sigma(0)}\left(t\dot{\Sigma}(0)\right)$.

Gradient I

- We have 3 manifold structures on $\operatorname{Sym}^{++}(d)$; Fisher-Rao Riemannian manifold, Exponential affine manifold, Wasserstein Riemannian manifold. In each case we have a definition of velocity $D\gamma(t)$ of a curve $t \mapsto \gamma(t) \in \operatorname{Sym}^{++}(d)$ and a scalar product on each of the tangent space $T_A \operatorname{Sym}^{++}(d)$, $A \in \operatorname{Sym}^{++}(d)$.
- In both the Fisher-Rao and Wasserstein manifold each tangent space is identified with the Hilbert space $\operatorname{Sym}(d)$. In the Exponential case, each tangent space is a sub-vector space od codimension 1 of an Hilbert space. Let us denote by $\mathcal{T}$ the vector space containing all tangent spaces.
- A smooth mapping $X \colon \operatorname{Sym}^{++}(d) \to \mathcal{T}$ such that $f(A) \in T_A \operatorname{Sym}^{++}(d)$, $A \in \operatorname{Sym}^{++}(d)$, is a section or vector field or estimating function.
- Given a vector field X, consider the *differential equation*

$$D\gamma(t) = X(\gamma(t)), \quad \gamma(0) = A$$

  This defines the flow of X.

Gradient II

- Let $f\colon \operatorname{Sym}^{++}(d) \to \mathbb{R}$ be a smooth function. For each smooth curve $t \mapsto \gamma(t)$ the real runction $t \mapsto f(\Sigma(t))$ is differentiable. The natural gradient is the vector field $\operatorname{grad} f$ such that for all smooth function f and all smooth curve γ we have

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

- Given $A \in \operatorname{Sym}^{++}(d)$ and $V \in T_A \operatorname{Sym}^{++}(d)$ let γ be a smooth curve such that $\gamma(0) = A$ and $D\gamma(0) = V$. Then

$$\langle \operatorname{grad} f(A), V \rangle_A = \left. \frac{d}{dt} f(\gamma(t)) \right|_{t=0}$$

- The gradient flow of f is the flow of $\operatorname{grad} f$. Each trajectory is a solution of

$$D\gamma(t) = \operatorname{grad} f(\gamma(t))$$

Gradient III

- In both Fisher-Rao and Wasserstein the velocity is the ordinary derivative, $D\gamma(t) = \dot{\gamma}(t)$, hence

$$\frac{d}{dt} f(\gamma(t)) = df(\gamma(t))[\dot{\gamma}(t)] = \langle \nabla_2 f(\gamma(t)), \dot{\gamma}(t) \rangle_2$$

where ∇_2 is the gradient with respect to the scalar product $\langle \cdot, \cdot \rangle_2$.

- We can express the Fisher metric with the 2-metric:

$$\langle \nabla_2 f(\gamma(t)), D\gamma(t) \rangle_2 = \frac{1}{2} \operatorname{Tr} \left(\nabla_2 f(\gamma(t)) D\gamma(t) \right)$$
$$\frac{1}{2} \operatorname{Tr} \left(\gamma(t)^{-1} \gamma(t) \nabla_2 f(\gamma(t)) \gamma(t) \gamma(t)^{-1} D\gamma(t) \right) =$$
$$F_{\gamma(t)}(\gamma(t) \nabla_2 f(\gamma(t)) \gamma(t), D\gamma(t))$$

In this case

$$\operatorname{grad} f(\Sigma) = \Sigma \nabla_2 f(\Sigma) \Sigma$$

Covariant derivative I

- Given two smooth vector fields X and Y the covariant derivative is a vector field $D_X Y$ which has the properties of a derivation of Y in direction of X. The manifold structure does not define uniquely a covariant derivative.
- When $Y = \operatorname{grad} f$, then we define the Hessian as $\operatorname{Hess}_X f = D_X \operatorname{grad} F$.
- When $Y(\gamma(t)) = D\gamma(t)$, then $D_{D\gamma(t)} Y(\gamma(t))$ is the accelleration of the curve γ. By identifying curves with 0 accelleration, we can compute relevant Taylor formulæ.
- In the case of both the Fisher-Rao and the Wasserstein Riemannian structure, it is natural to use the Levi-Civita covariant derivative which has the property "derivative of the product":

$$\frac{d}{dt} \langle Y(\gamma(t)), Z(\gamma(t)) \rangle_{\gamma(t)} = \left\langle D_{\dot{\gamma}(t)} Y(\gamma(t)), Z(\gamma(t)) \right\rangle_{\gamma(t)} + \left\langle Y(\gamma(t)), D_{\dot{\gamma}(t)} Z(\gamma(t)) \right\rangle_{\gamma(t)}$$

- Levi-Civita connections are computed explicitely from derivation of the left-end side.

Covariant derivative II

- In the case of the Exponential manifold, it is more appropriate to use a different approach, using the transports ${}^{e}\mathbb{U}_{A}^{B}$, ${}^{m}\mathbb{U}_{A}^{B}$, $A, B \in \operatorname{Sym}^{++}(d)$.
- The velocity at $t+h$, $D\gamma(t+h)$ of the curve γ belongs to $S_{\gamma(t+h)} \operatorname{Sym}^{++}(d) \neq S_{\gamma(t)} \operatorname{Sym}^{++}(d)$, so we define the accelleration to be

$$\lim_{h \to 0} h^{-1} \left({}^{e}\mathbb{U}_{\gamma(t+h)}^{\gamma(t)} D\gamma(t+h) - D\gamma(t) \right)$$

or

$$\lim_{h \to 0} h^{-1} \left({}^{m}\mathbb{U}_{\gamma(t+h)}^{\gamma(t)} D\gamma(t+h) - D\gamma(t) \right)$$

- M. P. do Carmo. *Riemannian geometry*. Mathematics: Theory & Applications. Birkhäuser Boston Inc., Boston, MA, 1992. Translated from the second Portuguese edition by Francis Flaherty, Ch. 2. In this book, *covariant derivatives* are named *affine connections*.
- 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
- L. Malagò and G. Pistone. Second-order optimization over the multivariate gaussian distribution. In F. Barbaresco and F. Nielsen, editors, *Geometric Science of Information*, volume 9389 of *LNCS*, pages 349–358. Springer, 2015