Diffeological Fisher metric

Hông Vân Lê
Institute of Mathematics, CAS, Praha

Les Houches July 28, 2020

OUTLINE

1) Intrinsic Riemannian metric on statistical models.

2) Diffeological statistical models.

3) Diffeological Fisher metric.

4) Diffeological Crámer-Rao inequality.

1. Intrinsic Riemannian metric on statistical models

• A statistical model is a subset $P_{\mathcal{X}}$ of the set $\mathcal{P}(\mathcal{X})$ of all probability measures on $\mathcal{X}$.

• Geometry of $P_{\mathcal{X}}$ is induced from $(\mathcal{S}(\mathcal{X}), ||, ||_{TV})$.

- $(V, \|\cdot\|)$ - a Banach space, $\mathcal{X} \stackrel{i}{\hookrightarrow} V$ and $x_0 \in \mathcal{X}$. Then $v \in V$ is called a tangent vector of $\mathcal{X}$ at x_0, if there is a C^1-map $c : \mathbf{R} \to \mathcal{X}$.
- The tangent (double) cone

$$C_x\mathcal{X} := \{v \in V |\, v \text{ is tangent to } \mathcal{X} \text{ at } x\}.$$

- The tangent space $T_x\mathcal{X} := \mathrm{Lin}(C_x\mathcal{X})$.
- The tangent cone fibration

$C\mathcal{X} := \cup_x \in \mathcal{X} T_x\mathcal{X}$

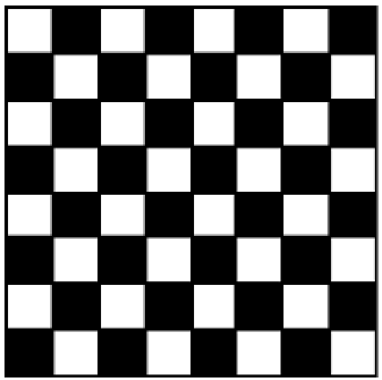
- The tangent fibration $T\mathcal{X} := \cup_{x\in\mathcal{X}} T_x\mathcal{X} \subset V \times V$ is endowed with the induced topology.

Example. $P_{\mathcal{X}} := \{p_\eta \mu_0 \in \mathcal{P}(\mathcal{X})\}$, $\mu_0 \in \mathcal{P}(\mathcal{X})$,

$$p_\eta := g^1\eta_1 + g^2\eta_1 + g^3(1 - \eta_1 - \eta_2)$$

and $g^i \geq 0$ such that $\mathbb{E}_{\mu_0}(g^i) = 1$ and $\eta = (\eta_1, \eta_2) \in D_b \subset \mathbf{R}^2$ is a parameter, which will be specified as follows.

Let us divide the square D in smaller squares and color them in black and white like a chessboard. Let D_b be the closure of the subset of D colored in black.



• Any $v \in C_\xi P_{\mathcal{X}}$ is dominated by ξ. Hence the logarithmic representation of v

$$\log v := dv/d\xi \in L^1(\mathcal{X}, \xi).$$

• The logarithmic representation of $C_\xi P_{\mathcal{X}}$ $\log(C_\xi P_{\mathcal{X}}) := \{\log v | v \in C_\xi P_{\mathcal{X}}\} \subset L^1(\mathcal{X}, \xi)$.

• $P_{\mathcal{X}}$ will be called almost 2-integrable, if

$$\log(C_\xi P_{\mathcal{X}}) \subset L^2(\mathcal{X}, \xi) \; \forall \xi \in P_{\mathcal{X}}.$$

In this case the Fisher metric $\mathfrak{g}$ on $\mathcal{P}_{\mathcal{X}}$ ais defined as follows.

For $v, w \in C_\xi P_{\mathcal{X}}$

$$\mathfrak{g}_\xi(v, w) := \int_{\mathcal{X}} \log v \cdot \log w \, d\xi.$$

Since $T_\xi P_{\mathcal{X}}$ is the linear hull of $C_\xi P_{\mathcal{X}}$, this formula extends uniquely to a positive quadratic form on $T_\xi P_{\mathcal{X}}$, which is called the Fisher metric.

2. Diffeological statistical models

• A parameterized statistical model is a parameter set Θ together with a mapping $\mathbf{p} : \Theta \to \mathcal{P}(\mathcal{X})$.

• In "Information Geometry" (AJLS2017) a parameterized statistical model - $(M, \mathcal{X}, \mathbf{p})$,
M - a Banach manifold,
$i \circ \mathbf{p} : M \xrightarrow{\mathbf{p}} \mathcal{P}(\mathcal{X}) \xrightarrow{i} \mathcal{S}(\mathcal{X})_{TV}$ is a C^1-map.

Example. Let $\mathcal{X} = [0,1]$, μ_0- Lebesgue,
$\mathcal{P}_\mathcal{X} = \{f \cdot \mu_0 | \; f \in C^\infty_{>0}(\mathcal{X}), \& \int_\mathcal{X} f d\mu_0 = 1\}$.
Then there does not exist $(M, \mathcal{X}, \mathbf{p})$ s.t.
$\mathcal{P}_\mathcal{X} = \mathbf{p}(M)$, M -a Banach manifold.

Assume the opposite. $\Longrightarrow \forall\, m \in M$:
$d\mathbf{p}(T_m M) = \{f \in C^\infty(\mathcal{X}) | \; \int_\mathcal{X} d\mu_0 = 0\}$.

But this is not the case, since the space $C^\infty([0,1])$ cannot be the image of a linear bounded map from a Banach space M to $L_1([0,1])$.

• A C^k-diffeology $\mathcal{D}$ of $\mathcal{X} \neq \emptyset$ is a subset of $\mathcal{X}^U$, $U \subset \mathbf{R}^n$ is open, $n \in \mathbf{N}$, that satisfies the following.

D1. Covering. $\mathcal{D}$ contains all the constant mappings $\mathbf{x} : r \mapsto x$, $\forall\, n,\ r \in \mathbf{R}^n \,\&\, x \in \mathcal{X}$.

D2. Locality. Let $P \in \mathcal{X}^U$. If $\forall\ r \in U$ there exists an open neighborhood V of r s.t. $P_{|V} \in \mathcal{D}$ then $P \in \mathcal{D}$.

D3. Smooth compatibility. For every $P \in \mathcal{D}$, for every real domain V, for every $F \in C^k(V, U)$, we have $P \circ F \in \mathcal{D}$.

• A C^k-diffeological space is a nonempty set equipped with a C^k-diffeology $\mathcal{D}$. Elements $P \in \mathcal{D}$ are called C^k-maps from U to $\mathcal{X}$.

• $(\mathcal{P}_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}})$ is called a C^k-diffeological statistical model, if for any $P \in \mathcal{D}_{\mathcal{X}}$, $i \circ P : U \to \mathcal{S}(\mathcal{X})$ is a C^k-map.

• The tangent cone $C_\xi(\mathcal{P}_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}}) \subset C_\xi \mathcal{P}_{\mathcal{X}}$ consists of tangent vectors of C^k-curves in $\mathcal{D}_{\mathcal{X}}$.

• The tangent space $T_\xi(\mathcal{P}_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}})$ is the linear hull of $C_\xi(\mathcal{P}_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}})$.

• Let V be a locally convex vector space. A map $\varphi : \mathcal{P}_{\mathcal{X}} \to V$ is called Gateaux-differentiable on $(\mathcal{P}_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}})$ if for any C^k-curve c in $\mathcal{D}_{\mathcal{X}}$ the composition $\varphi \circ c : \mathbf{R} \to V$ is differentiable.

Example. Let $(M, \mathcal{X}, \mathbf{p})$ be a parameterized statistical model. Then $(\mathbf{p}(M), \mathcal{D}_{\mathcal{X}})$ is a C^1-diffeological statistical model where $\mathcal{D}_{\mathcal{X}}$ consists of all C^1-maps $q : \mathbf{R}^n \supset U \to \mathbf{p}(M)$ such that there exists a C^1-map $q^M : U \to M$ and $q = \mathbf{p} \circ q^M$.

Example. Any statistical model $\mathcal{P}_{\mathcal{X}}$ can be endowed with a structure of a C^k-diffeological statistical model for any $k \in \mathbf{N}^+ \cup \infty$, where its diffeology $\mathcal{D}_{\mathcal{X}}^{(k)}$ consists of all mappings $P : U \to \mathcal{P}_{\mathcal{X}}$ such that the composition $i \circ P : U \to \mathcal{S}(\mathcal{X})$ is of the class C^k, where U is any open domain in $\mathbf{R}^n$ for $n \in \mathbf{N}$.

3. Diffeological Fisher metric.

• $(\mathcal{P}_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}})$ is called almost 2-integrable, if $\log(C_\xi(\mathcal{P}_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}})) \subset L^2(\mathcal{X}, \xi)$ for all $\xi \in P_{\mathcal{X}}$.

• An almost 2-integrable $(\mathcal{P}_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}})$ will be called 2-integrable, if for any $\mathbf{p} \in \mathcal{D}_{\mathcal{X}}$, the function $v \mapsto |d\mathbf{p}(v)|_{\mathfrak{g}}$ is continuous on TU. The Fisher metric on an 2-integrable $(\mathcal{P}_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}})$ is called the diffeological Fisher metric.

Example. $(M, \mathcal{X}, \mathbf{p})$ is 2-integrable, iff $(\mathbf{p}(M), \mathbf{p}_*(\mathcal{D}_M))$ is a 2-integrable C^1-diffeological statistical model.

Example. λ - a σ-finite measure on $\mathcal{X}$. Friedrich (1991) set $\mathcal{P}(\lambda) := \{\mu \in \mathcal{P}(\mathcal{X}) | \, \mu \ll \lambda\}$ with the following diffeology $\mathcal{D}(\lambda)$. A curve $c : \mathbf{R} \to \mathcal{P}(\lambda)$ is a C^1-curve, iff

$$\log \dot{c}(t) \in L^2(\mathcal{X}, c(t)).$$

Then $(\mathcal{P}(\lambda), \mathcal{D}(\lambda))$ is an almost 2-integrable C^1-diffeological statistical model.

The diffeological Fisher metric serves as a information quantity wrt Markov kernels, regarded as probabilistic morphisms.

• (1962) Lawvere proposed a category $\{\mathcal{X}, T : \mathcal{X} \sim \mathcal{Y} | T$ is a Markov kernel, $\Longleftrightarrow \overline{T} : \mathcal{X} \to (\mathcal{P}(\mathcal{Y}), \Sigma_w)$ is measurable$\}$. Here Σ_w the smallest σ-algebra on $\mathcal{P}(\mathcal{Y})$ such that $I_f : \mu \mapsto \int_{\mathcal{X}} f d\mu$ for $f \in \mathcal{F}_s(\mathcal{X})$ and $\mu \in \mathcal{P}(\mathcal{X})$ is measurable for all $f \in \mathcal{F}_s(\mathcal{X})$, i.e. f is simple.

• T is called a probabilistic morphism.

- $T : \mathcal{X} \rightsquigarrow \mathcal{Y}$, induces a linear map $T_* = S_*(T) : \mathcal{S}(\mathcal{X}) \to \mathcal{S}(\mathcal{Y})$

$$T_*(\mu)(B) := \int_{\mathcal{X}} \overline{T}(x)(B) d\mu(x) \qquad (1)$$

for any $\mu \in \mathcal{S}(\mathcal{X})$ and $B \in \Sigma_{\mathcal{Y}}$.

- $T_*(\mathcal{P}(\mathcal{X})) \subset \mathcal{P}(\mathcal{Y})$.

- Given $T : \mathcal{X} \rightsquigarrow \mathcal{Y}$ and a C^k-diffeological statistical model $(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}})$, then $(T_*(P_{\mathcal{X}}), T_*(\mathcal{D}_{\mathcal{X}}))$ is a C^k-statistical model.

- A mapping $\mathbf{p} : U \to T_*(P_{\mathcal{X}})$ belongs to $T_*(\mathcal{D}_{\mathcal{X}})$ iff $\forall\ r \in U$ $\exists$ an open neighborhood $V \subset U$ of r s.t. either $\mathbf{p}_{|V} = const$, or there exists a mapping $\mathbf{q} \in \mathcal{D}_{\mathcal{X}}$ such that $\mathbf{p}_{|V} = T_* \circ \mathbf{q}$.

Theorem 1. Given $T : \mathcal{X} \rightsquigarrow \mathcal{Y}$ and an almost 2-integrable C^k-d.s.m. $(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}})$, then

(1) $(T_*(P_{\mathcal{X}}), T_*(\mathcal{D}_{\mathcal{X}}))$ is an almost 2-integrable C^k-d.s.m.

(2) For any $\mu \in P_{\mathcal{X}}$, $v \in T_\mu(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}})$

$$\mathfrak{g}_\mu(v, v) \geq \mathfrak{g}_{T_*\mu}(T_*v, T_*v)$$

with the equality if T is sufficient w.r.t. $P_{\mathcal{X}}$.

• $T : \mathcal{X} \rightsquigarrow \mathcal{Y}$ is called sufficient for $P_{\mathcal{X}}$ if there exists $\underline{\mathbf{p}} : \mathcal{Y} \rightsquigarrow \mathcal{X}$ s.t. $\forall\ \mu \in P_{\mathcal{X}}$ and $h \in L(\mathcal{X})$ (bounded measurable functions on $\mathcal{X}$)

$$T_*(h\mu) = \underline{\mathbf{p}}^*(h) T_*(\mu)$$

$$\Longleftrightarrow \underline{\mathbf{p}}^*(h) = \frac{dT_*(h\mu)}{dT_*(\mu)} \in L^1(\mathcal{Y}, T_*(\mu)).$$

In this case we call $\mathbf{p} : \mathcal{Y} \to \mathcal{P}(\mathcal{X})$ defining $\underline{\mathbf{p}} : \mathcal{Y} \rightsquigarrow \mathcal{X}$ the conditional mapping for T.

Example. Let λ be a σ-finite measure on $\mathcal{X}$. In (Friedrich1991) Friedrich considered the group $\mathcal{G}(\mathcal{X}, \Sigma_{\mathcal{X}}, \lambda)$ of all measurable 1-1 mappings $\Phi : \mathcal{X} \to \mathcal{X}$ such that $\Phi_*(\lambda) \ll \lambda$. Clearly $\Phi_*(P(\lambda)) \subset P(\lambda)$. It is not hard to see that Φ is a sufficient statistic and hence sufficient probabilistic morphism w.r.t. $P(\lambda)$. Hence Theorem 1 implies the following

Corollary 1. (Friedrich1991) The group $\mathcal{G}(\mathcal{X}, \Sigma_{\mathcal{X}}, \lambda)$ acts isometrically on $P(\lambda)$.

4. Diffeological Crámer-Rao inequality

- An estimator is a map $\hat{\sigma} : \mathcal{X} \to P_{\mathcal{X}}$.

- It is simpler to estimate only a "coordinate" $\varphi(\xi)$, where $\xi \in \mathcal{P}_{\mathcal{X}}$ and $\varphi \in Map(P_{\mathcal{X}}, V)$.

- A φ-estimator $\hat{\sigma}_\varphi$ is a composition $\varphi \circ \hat{\sigma} : \mathcal{X} \stackrel{\hat{\sigma}}{\to} P_{\mathcal{X}} \stackrel{\varphi}{\to} V$.

Example. Assume that $k : \mathcal{X} \times \mathcal{X} \to \mathbf{R}$ is a symmetric and positive definite kernel function and V be the associated RKHS. For any $x \in \mathcal{X}$ we denote by k_x the function on $\mathcal{X}$ defined by $k_x(y) := k(x, y)$ for any $y \in \mathcal{X}$. Then k_x is an element of V. Let $P_{\mathcal{X}} = \mathcal{P}(\mathcal{X})$. Then we define the kernel mean embedding $\varphi : \mathcal{P}(\mathcal{X}) \to V$ as follows (MFSS2017)

$$\varphi(\xi) := \int_{\mathcal{X}} k_x d\xi(x),$$

where the integral should be understood as a Bochner integral.

Remark In classical statistics one considers only parameter estimations for parameterized statistical models. In this case, an estimators a map from $\mathcal{X}$ to the parameter set Θ of a statistical model $\mathbf{p}(\Theta) \subset \mathcal{P}(\mathcal{X})$. Usually one assumes that the parametrization $\mathbf{p} : \Theta \to \mathbf{p}(\Theta)$ is 1-1, hence in this case, a parameter estimation is equivalent to a nonparametric estimation in the sense of our Definition.

The notion of a φ-estimation occurs in classical statistics under different name e.g. substitution estimator, estimand, etc.

- V' - the topological dual of V.

- $\varphi^l := l \circ \varphi$, $l \in V'$, $\varphi \in Map(P_{\mathcal{X}}, V) = \mathcal{P}_{\mathcal{X}}^V$.

$L^2_\varphi(\mathcal{X}, P_{\mathcal{X}}) := \{\hat{\sigma} \in P_{\mathcal{X}}^{\mathcal{X}} | \varphi^l \circ \hat{\sigma} \in L^2_\xi(\mathcal{X}), \xi \in P_{\mathcal{X}}, l \in V'\}$.

- The φ-mean value $\varphi_{\hat{\sigma}} \in P_{\mathcal{X}}^{V''}$ of $\hat{\sigma}$ is

$\varphi_{\hat{\sigma}}(\xi)(l) := \mathbb{E}_\xi(\varphi^l \circ \hat{\sigma})$ for $\xi \in P_{\mathcal{X}}$ and $l \in V'$.

- $V \subset V''$.

- $b^\varphi_{\hat{\sigma}} := \varphi_{\hat{\sigma}} - \varphi \in Map(\mathcal{P}_{\mathcal{X}}, V'')$ is the bias of the φ-estimator $\hat{\sigma}_\varphi$.

- Mean square error quadratic function on V'

$$MSE_{\xi}^{\varphi}[\widehat{\sigma}](l,h) = \mathbb{E}_{\xi}[(\varphi^{l}\circ\widehat{\sigma}(x)-\varphi^{l}(\xi))\cdot(\varphi^{h}\circ\widehat{\sigma}(x)-\varphi^{h}(\xi))].$$

- Variance quadratic function $V_{\xi}^{\varphi}[\widehat{\sigma}](l,h)$

$$= \mathbb{E}_{\xi}[\varphi^{l}\circ\widehat{\sigma}(x)-E_{\xi}(\varphi^{l}\circ\widehat{\sigma}(x))\cdot\varphi^{h}\circ\widehat{\sigma}(x)-E_{\xi}(\varphi^{h}\circ\widehat{\sigma}(x))].$$

- $MSE_{\xi}^{\varphi}[\widehat{\sigma}](l,h) = V_{\xi}^{\varphi}[\widehat{\sigma}](l,h)+\langle b_{\widehat{\sigma}}^{\varphi}(\xi),l\rangle\cdot\langle b_{\widehat{\sigma}}^{\varphi}(\xi),h\rangle.$

Remark Assume that V is a real Hilbert space with a scalar product $\langle \cdot, \cdot \rangle$ and the associated norm $\| \cdot \|$. Then the scalar product defines a canonical isomorphism $V = V'$, $v(w) := \langle v, w \rangle$ for all $v, w \in V$. For $\hat{\sigma} \in L^2_\varphi(\mathcal{X}, P_\mathcal{X})$ the mean square error $MSE^\varphi_\xi(\hat{\sigma})$ of the φ-estimator $\varphi \circ \hat{\sigma}$ is defined by

$$MSE^\varphi_\xi(\hat{\sigma}) := \mathbb{E}_\xi(\|\varphi \circ \hat{\sigma} - \varphi(\xi)\|^2). \qquad (2)$$

The RHS of (2) is well-defined, since $\hat{\sigma} \in L^2_\varphi(\mathcal{X}, P_\mathcal{X})$ and therefore

$$\langle \varphi \circ \widehat{\sigma}(x), \varphi \circ \widehat{\sigma}(x) \rangle \in L^1(\mathcal{X}, \xi),$$

$$\langle \varphi \circ \widehat{\sigma}(x), \varphi(\xi) \rangle \in L^2(\mathcal{X}, \xi).$$

Similarly, we define the variance of a φ-estimator $\varphi \circ \widehat{\sigma}$ at ξ as follows

$$V_\xi^\varphi(\widehat{\sigma}) := \mathbb{E}_\xi(\|\varphi \circ \widehat{\sigma} - \mathbb{E}_\xi(\varphi \circ \widehat{\sigma})\|^2).$$

If V has a countable basis of orthonormal vectors $v_1, \cdots, v_\infty$, then we have

$$MSE_\xi^\varphi(\widehat{\sigma}) = \sum_{i=1}^{\infty} MSE_\xi^\varphi[\widehat{\sigma}](v_i, v_i), \qquad (3)$$

$$V_\xi^\varphi(\widehat{\sigma}) = \sum_{i=1}^{\infty} V_\xi^\varphi[\widehat{\sigma}](v_i, v_i). \qquad (4)$$

- $(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}})$ - an almost 2-integrable C^k-d.s.m.

- $T_\xi^{\mathfrak{g}}(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}})$ - the completion of $T_\xi(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}})$ w.r.t. the diffeological Fisher metric $\mathfrak{g}$.

$$L_{\mathfrak{g}} : T_\xi^{\mathfrak{g}}(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}}) \to (T_\xi^{\mathfrak{g}}(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}}))'$$

$$L_{\mathfrak{g}}(v)(w) := \langle v, w \rangle_{\mathfrak{g}},$$

is an isomorphism.

Then we define the inverse $\mathfrak{g}^{-1}$ of the Fisher metric $\mathfrak{g}$ on $(T_\xi^{\mathfrak{g}}(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}}))'$ as follows

$$\langle L_g v, L_g w \rangle_{\mathfrak{g}^{-1}} := \langle v, w \rangle_{\mathfrak{g}}$$

• $\widehat{\sigma} \in L^2_{\varphi}(\mathcal{X}, P_{\mathcal{X}})$ is called a φ-regular estimator, if for all $l \in V'$ the function $\xi \mapsto \|\varphi^l \circ \widehat{\sigma}\|_{L^2(\mathcal{X},\xi)}$ is locally bounded, i.e., for all $\xi_0 \in P_{\mathcal{X}}$

$$\lim_{\xi \to \xi_0} \sup \|\varphi^l \circ \widehat{\sigma}\|_{L^2(\mathcal{X},\xi)} < \infty.$$

• For $\xi \in \mathcal{P}_{\mathcal{X}}$ we denote by $(\mathfrak{g}^{\varphi}_{\widehat{\sigma}})^{-1}(\xi)$ to be the following quadratic form on V':

$$(\mathfrak{g}^{\varphi}_{\widehat{\sigma}})^{-1}(\xi)(l,k) := \langle d\varphi^l_{\widehat{\sigma}}, d\varphi^k_{\widehat{\sigma}} \rangle_{\mathfrak{g}^{-1}}(\xi)$$

$$= \langle \mathrm{grad}_{\mathfrak{g}}(\varphi^l_{\widehat{\sigma}}), \mathrm{grad}_{\mathfrak{g}}(\varphi^k_{\widehat{\sigma}}) \rangle.$$

Theorem [Diffeological Cramér-Rao inequality]

Let $(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}})$ be a 2-integrable C^k-diffeological statistical model, φ a V-valued function on $P_{\mathcal{X}}$ and $\hat{\sigma} \in L^2_{\varphi}(\mathcal{X}, P_{\mathcal{X}})$ a φ-regular estimator. Then the difference $V^{\varphi}_{\xi}[\hat{\sigma}] - (\hat{\mathfrak{g}}^{\varphi}_{\hat{\sigma}})^{-1}(\xi)$ is a positive semi-definite quadratic form on V' for any $\xi \in P_{\mathcal{X}}$.

- J. Jost, H. V. Lê, , D. H. Luu and T. D. Tran, Probabilistic mappings and Bayesian nonparametrics, arXiv:1905.11448.

- H. V. Lê, Diffeological statistical models, the Fisher metric and probabilistic mappings, arXiv:1912.02090, Mathematics 2020, 8(2), 167.

Thank you for your attention!