An Introduction to Information Geometry

Giulia Bertagnolli

5/10/22

Table of contents

Preface

This are the lecture notes for the short course (8 hours) *The Geometry of Statistical Models* (Trento - March, 2023). Feedback, as well as reports on typos and errors, are welcome.

1 Introduction

The name C. R. Rao, Professor Emeritus of Statistics at Penn State University, is ubiquitous in statistics and it was him in 1945, who firstly understood the geometrical meaning of Fisher's information (Rao 1945). Some results by Efron (Efron 1975), in 1975, inspired Shun-ichi Amari, who discovered the family of affine $\alpha-$connections. Chentsov independently obtained the same results in 1972 (his work became known to the community only in 1982, with the English version of his work (Chentsov 1982)). Other recurring names in the field, just to name a few, are the ones of A. P. David, Lauritzen—who formalised the concept of a *statistical manifold* for finite sample spaces—Nagaoka, co-author, with Amari, of the very first book on information geometry (Amari and Nagaoka 2000), Giovanni Pistone for his works on non-parametric IG, see e.g. (Pistone 2013), and Ay-Jost-Lê-Schwachhöfer, for their recent book (Ay et al. 2017). Let us start with a very brief and informal introduction of the contents of this short course.

Intuitively, we will start from a sample space Ω and define a differentiable structure on the set $\mathcal{P}(\Omega)$ of probability measures on the sample space. Curves on this (probability/statistical) manifold are 1-dimensional parametric, statistical models. If I is an open interval of $\mathbb{R}$ and the mapping

$$I \ni \theta \mapsto p(\cdot;\theta)\nu$$

is smooth, then we can compute the velocity, acceleration, etc. of the curve and, consequently, we can describe the geometry of the statistical model. (I, Ω, p, ν) is called a (regular) 1-dimensional statistical model.

Observations

i. We have only introduced the sample space Ω, but we will need also a $\sigma-$algebra, i.e. $(\Omega, \mathcal{E})$ ans a $\sigma-$finite measure ν on this space. Then, as we will see, $p(\cdot;\theta)$ is a *density* w.r.t. the reference/dominating measure ν (i.e. we are in an absolute-continuous framework).
ii. When writing $p(x;\theta)$, $x \in \Omega$ is a sample, but we may also consider a random variable $X : \Omega \to \mathbb{R}$, and x represents an *observable* on Ω.
iii. When Ω is infinite, $\mathcal{P}(\Omega)$ is infinite-dimensional.
iv. Given a statistic $\kappa : \Omega \to \Omega'$, what happens to the geometric structure on $\mathcal{P}(\Omega)$? It turns out that the Fisher metric is invariant under sufficient statistics.

Let us look at some examples of manifolds of interest in IG: the set of positive-definite matrices of dimension $n \times n$ is a sub-manifold of dimension $\frac{n(n+1)}{2}$ of the n^2-dimensional manifold of all real matrices of that dimension; the set of neural networks, identified by the connection weights $\mathbf{W}$...

In the remaining of this section, we provide a brief recap of the main definitions of differential geometry, which are *useful* for understanding IG. *Usefull*, but not mandatory, as we will see in Chapter 3.

1.1 Differential Geometry Recap

The core objects of IG are manifolds, more specifically *differentiable* manifolds. So, we need a brief recap of some concepts and tools of differential geometry. For more details see (Sernesi 1994; Lang 2012; Petersen 2006) or the lecture notes of your favourite "Geometric analysis" course (also Moretti 2020).

A manifold is a set M endowed with a manifold structure, which is defined as a collection of *local charts*, an *atlas*.
A *local chart* is a pair (U, φ) where $U \subset M$ and $\varphi : U \to \varphi(U) \subset \mathbb{R}^n$[1] is a bijection and $\varphi(U)$ is open in $\mathbb{R}^n$.
Two charts (U, φ), (V, ψ) are said to be $\mathcal{C}^k-$*compatible* if either $U \cap V = \emptyset$, or the map $\psi \circ \varphi^{-1} : \varphi(U \cap V) \to \psi(U \cap V) \subset \mathbb{R}^n$ is a bijection, and both this and its inverse $\varphi \circ \psi^{-1} : \psi(U \cap V) \to \varphi(U \cap V) \subset \mathbb{R}^n$ are of class $\mathcal{C}^k$, i.e. $\psi \circ \varphi^{-1}$ is a *diffeomorphism* of class $\mathcal{C}^k$ between open sets of $\mathbb{R}^n$. An *atlas* of class $\mathcal{C}^k$ is a collection of charts $\{(U_\alpha, \varphi_\alpha)\}_{\alpha \in I}$, where $\cup_{\alpha \in I} U_\alpha = M$ and the transition maps are pair-wise $\mathcal{C}^k-$compatible. Finally, we say that the atlas $\{(U_\alpha, \varphi_\alpha)\}_{\alpha \in I}$ defines a structure of $\mathcal{C}^k-$manifold on M and $\dim M = n$. If the charts are $\mathcal{C}^\infty-$compatible we talk about smooth charts, atlas, and manifold. On the other hand, if $k = 0$ we call the manifold a *topological* manifold.

Remarks

i. One can give M a topology in a unique way such that each each U_α is open and the φ_α are topological isomorphisms (or *homeomorphism*, i.e. bijectinve and bi-continuous).
ii. Given two atlases of class $\mathcal{C}^k$, they are *equivalent* if their union is still an atlas of class $\mathcal{C}^k$ and it is the equivalent class of atlases of class $\mathcal{C}^k$ that defines a $\mathcal{C}^k-$manifold on M.
iii. We assume here that everyone has some familiarity with the fundamentals of differential geometry, so we do not make examples. For a more thorough introduction on differential geometry, see(Lang 2012; Sernesi 1994).

[1] Here, φ could go, in general, to a topological linear space, i.e. a linear space with a topology making the operations of sum and scalar multiplication, continuous(Lang 2012) (e.g. a Banach space). In this case, the transition map $\psi \circ \varphi^{-1}$ would be an $\mathcal{C}^k-$isomorphism of topological spaces. Here you might ask what is the differentiability for a map between topological spaces, for which a good reference(Lang 2012).

Given a chart at $p \in M$, i.e. $U \ni p$ and a $\varphi : U \in \mathbb{R}^n$, this is determined by its n component functions $\{\xi^i : U \to \mathbb{R}\}_{i=1}^n$, such that $\varphi(p) = (\xi^1(p), \dots, \xi^n(p))$. These are called the n local coordinates on U defined by the chart φ. Given two local charts at $p \in U \cap V \subset M$, (U, φ), (V, ψ), with coordinate systems $[\xi^i], [\rho^i]$ respectively, the compositions $\psi \circ \varphi^{-1}$ and $\varphi \circ \psi^{-1}$ are the change of coordinates maps.

Let us look at an example, which will play an important role in understanding *affine connections.*

Example: Affine manifold

A real affine space of dimension n $\mathbb{A}^n$ is a triplet $(\mathbb{A}^n, V, \vec{\ })$, where $\mathbb{A}^n$ is the set of points, V is an $n-$dimensional vector space over $\mathbb{R}$–called the space of translations– and$\vec{\ }$is a map from $\mathbb{A}^n \times \mathbb{A}^n$ to V satisfying the following properties:

(i) for each fixed $p \in \mathbb{A}^n$ and vector $v \in V$ there exists a unique $q \in \mathbb{A}^n$ such that $\overrightarrow{pq} = v$

(ii) $\overrightarrow{pq} + \overrightarrow{qr} = \overrightarrow{pr}$.

Each affine space is a connected and path-connected topological manifold with a natural $\mathcal{C}^\infty$ differential structure. For each point $O \in \mathbb{A}^n$ (the origin) and vector basis $\{e_i\}_{i=1}^n \subset V$ we can consider the map $f : \mathbb{A}^n \to \mathbb{R}^n$ which takes a point $p \in \mathbb{A}^n$ into the n coordinates of $\overrightarrow{Op}$ w.r.t. the basis $\{e_i\}_{i=1}^n \subset V$, which is a bijection. Furthermore the Euclidean topology on $\mathbb{R}^n$ induces a topology on $\mathbb{A}^n$, which does not depend on the choice of the origin and basis. f defines a global chart on $\mathbb{A}^n$–called the Cartesian coordinate system with origin $O \in \mathbb{A}^n$ and axes $\{e_i\}_{i=1}^n \subset V$–and each mapping f defines a smooth atlas on the affine space. Given two of these maps f, g which are determined by different origins and bases in V, $g \circ f^{-1} : \mathbb{R}^n \to \mathbb{R}^n$ and $f \circ g^{-1} : \mathbb{R}^n \to \mathbb{R}^n$ are linear and non-homogeneous coordinate transformations and are hence smooth.

Let us now introduce the concept of differentiability of functions on a manifold.

A continuous map $f : M \to N$ between two differentiable manifolds of dimension n and m resp. is *smooth* (or also *differentiable*, or a *morphism*) at $p \in M$ if $\psi \circ f \circ \varphi^{-1} : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable for all charts (U, φ), (V, ψ) such that $p \in U$ and $f(p) \in V$. We indicate by $D(M|N)$ the class of smooth functions between M and N, or just by $D(M)$, when $N = \mathbb{R}$.

1.1.1 Tangent spaces and differentials

Let us begin with derivations and differentiations in $\mathbb{R}^n$.

With this identification of vectors with (directional) derivatives in mind, let us define the tangent spaces of a manifold M.

Definition 1.1 (Derivations)**.** Given a smooth manifold M, a *derivation* in $p \in M$ is a $\mathbb{R}-$linear map $D_p : D(M) \to \mathbb{R}$ such that for all $f, g \in D(M)$

$$D_p fg = f(p) D_p g + g(p) D_p f.$$

With the following linear structure

$$\left(aD_p + bD'_p\right) f := aD_p f + bD'_p f \quad \forall a, b \in \mathbb{R},\ \forall f \in D(M)$$

the set $\mathcal{D}_p(M)$ of of all derivations at p becomes an $\mathbb{R}-$vector space.

We first observe that the space of derivations is not empty. Given a chart (U, φ) at p with coordinates $[\xi^i]$ the operators

$$\left.\frac{\partial}{\partial \xi^i}\right|_p : f \mapsto \left.\frac{\partial f \circ \varphi^{-1}}{\partial \xi^i}\right|_{\varphi(p)}$$

are derivations. The subspace of $\mathcal{D}_p(M)$ spanned by $\frac{\partial}{\partial \xi^i}$ has the same dimension as M and does not depend on the choice of the chart at p. Let $[\rho^i]$ be another local coordinate system at p defined by the chart (ψ, V), then we have

$$\left.\frac{\partial}{\partial \rho^k}\right|_p = \left.\frac{\partial \xi^r}{\partial \rho^k}\right|_{\psi(p)} \left.\frac{\partial}{\partial \xi^r}\right|_p \tag{1.1}$$

where the terms $\left.\frac{\partial \xi^r}{\partial \rho^k}\right|_{\psi(p)}$ are the coefficients of the Jacobian J of the change of coordinates transformation, which is non singular. By definition, indeed, we have that $\left.\frac{\partial \xi^r}{\partial \xi^s}\right|_p = \delta^r_s$ and we can compose the maps as follows $\varphi \circ \psi^{-1} \circ \psi \circ \varphi^{-1}$ which is the identity on $\varphi(U \cap V)$, so that $\delta^r_s = \left.\frac{\partial \xi^r}{\partial \xi^s}\right|_p = \left.\frac{\partial \xi^r}{\partial \rho^k}\right|_{\psi(p)} \left.\frac{\partial \rho^k}{\partial \xi^s}\right|_{\varphi(p)}$, i.e. the matrix J is invertible, hence non singular. Therefore the spaces spanned by $\left.\frac{\partial}{\partial \xi^i}\right|_p$ and $\left.\frac{\partial}{\partial \rho^k}\right|_p$ coincide. It remains to prove that the dimension of the span of the n derivations is n, i.e. that the n derivations are linearly independent. But we refer to any book on differential geometry for this.

Definition 1.2 (Tangent space)**.** The tangent space of M at $p \in M$ is indicated by $T_p M$ and is the subspace of $\mathcal{D}_p(M)$ spanned by the n derivations $\frac{\partial}{\partial \xi^i}$. It has dimension n and does not depend on the choice of the chart at p.

The space of all derivations on M at p coincides with T_pM.

Let us go back to the affine manifold and consider its tangent space at $p \in \mathbb{A}^n$, $T_p\mathbb{A}^n$. It turns out that there is a natural *isomorphism* between $T_p\mathbb{A}^n$ and V.

Definition 1.3.

Tangent and cotangent bundles

$$TU := \left\{(p, v) \mid p \in U, \quad v \in T_pM\right\}, \quad T^*U := \left\{(p, \omega) \mid p \in U, \omega \in T_p^*M\right\}$$

Definition 1.4 (Differential of a mapping or push forward)**.** Let M, N be two smooth manifolds and $f : M \to N$ a smooth function. The differential of f at $p \in M$ or push forward of f at p is the linear mapping

$$\begin{aligned} df_p : T_pM &\to T_{f(p)}N \\ X_p &\mapsto dfX_p \end{aligned} \tag{1.2}$$

defined by $dfX_p(g) := X_p(g \circ f)$ for all vectors $X_p \in T_pM$ and all smooth functions $g \in D(N)$.

1.1.2 Vector and tensor fields

A vector field is a mapping $X : p \mapsto X_p \in T_pM$, which associates to each point p in the manifold M a tangent vector. We indicate by $\mathfrak{X}(M)$ the set of all vector fields on M. Observe that this set is not empty, for instance, the $n-$mappings defined by $\frac{\partial}{\partial \xi^i} : p \to \left.\frac{\partial}{\partial \xi^i}\right|_p$ are vector fields, formed by the natural basis given by the coordinate system $\left[\xi^i\right]$. Each vector field X may be written as $X_p = X_p^i \partial_i|_p$, where $\partial_i := \frac{\partial}{\partial \xi^i}$ and X_p^i, for $i = 1, \dots, n$, are the scalar components of X w.r.t. the coordinate system $\left[\xi^i\right]$.

- change of basis

If the components of the vector field are C^∞ w.r.t. some coordinate system, then they are smooth w.r.t. any coordinate system, and X is then called a *smooth vector field*. With the following structure

$$X + Y : p \mapsto X_p + Y_p \qquad cX : p \mapsto cX_p$$

the set $\mathfrak{X}(M)$ becomes a linear space. More generally, a mapping $t : M \to \mathcal{A}_{\mathbb{R}}(T_pM)$ which associates to a point $M \ni p$ a tensor t_p in the tensor algebra generated by T_pM, T_p^*M, and $\mathbb{R}$, is said to by a tensor field.

- multilinear maps
- tensor products

Assigning a smooth tensor field T on M is equivalent to assign a set of smooth functions which map

$$(\xi^1, \ldots, \xi^n) \mapsto T^{i_1 \cdots i^m}{}_{j_1 \cdots j_k}(\xi^1, \ldots, \xi^n)$$

in every local coordinate patch of M such that they satisfy the rules of transformation of the components of a tensor, i.e.

$$T^{i_1 \cdots i_m}{}_{j_1 \cdots j_k},(\xi^1, \ldots, \xi^n) = \left.\frac{\partial \xi^{i_1}}{\partial \rho^{k_1}}\right|_p \cdots \left.\frac{\partial \xi^{i_m}}{\partial \rho^{k_m}}\right|_p \left.\frac{\partial \rho^{l_1}}{\partial \xi^{j_1}}\right|_p \cdots \left.\frac{\partial \rho^{l_m}}{\partial \xi^{j_m}}\right|_p T'^{k_1 \cdots k_m}{}_{l_1 \cdots l_k}(\rho^1, \cdots, \rho^n)$$

Remark. Each vector field $X \in \mathfrak{X}(M)$ defines a derivation at each point $p \in M$: take any differentiable $f \in D(M)$ then $X_p(f) := X^i(p) \frac{\partial f}{\partial \xi^i}\Big|_p$. In general, every smooth vector field X defines a linear mapping from $D(M)$ to $D(M)$ by $f \mapsto X(f)$, where $X(f)(p) =: X_p(f)$ fore every $p \in M$.

The differential of $f \in D(M)$ at p is the 1-form defined, in local coordinates, by

$$df_p = \left.\frac{\partial f}{\partial \xi^i}\right|_p d\xi^i|_p.$$

Varying $p \mapsto df_p$ we have defined a smooth vector field df, called the differential of f (note the absence of “at p”).

A particularly important tensor of covariant degree 2, i.e. a tensor in $[T_pM]^0_2$ is the Riemannian metric tensor, which we are introduce in the following section.

1.1.3 Riemannian manifolds

Assume that, for each $p \in M$, an inner product $\langle\ ,\ \rangle_p$ is defined on T_pM. The mapping $g : p \mapsto \langle\ ,\ \rangle_p \in [T_pM]^0_2$, or, equivalently, assume we have a smooth covariant tensor field on M of degree 2, determining a symmetric, positive definite quadratic form $g(p) : T_pM \times T_pM \to \mathbb{R}$. g is called **Riemannian metric** on M and (M, g) is then called **Riemannian manifold**. Observe that this metric is, in general, not unique and it is not naturally determined by the structure of M as a manifold.

We assume the existence of a Riemannian metric on M, but the following can be proved:

Theorem 1.1. *If M is a connected, smooth manifold, it is possible to define a Riemannian metric g on M.*

Proof. See [ADD CITATION HERE].

□

Given a coordinate system $[\xi^i]$ at p, using our usual notation $\partial_i := \frac{\partial}{\partial \xi^i}$ (more precisely, we should write $\partial_i|_p$ but it should be obvious from the context), we can see that the components g_{ij}, for $i, j = 1, \dots, n$, of g at p, are determined by

$$g_{ij}(p) = \langle \partial_i, \partial_j \rangle_p, \tag{1.3}$$

so that :

- the tensor at p is written as $g(p) = g_{ij}(p) d^i|_p \otimes d^j|_p$, where $\{d^i|_p\} = \{d\xi^i|_p\}$, for $i = 1, \dots, n$ is the dual basis of $\{\partial_i\}$ in the cotangent space $T^*_p M$;
- the scalar product between two tangent vecctor at p is $\langle v, w \rangle_p = g_{ij}(p) v^i w^j$, for any two vectors $v = v^i \partial_i|_p, w = w^i \partial_i|_p \in T_p M$;
- and the norm of any $v^i \partial_i|_p = v \in T_p M$ is given by $\|v\|_p^2 = g_{ij}(p) v^i v^j$.

Furthermore, we can define the **length of a (piecewise) smooth curve** $\gamma : I \ni t \mapsto \gamma(t) \in M$, where $I \subset \mathbb{R}$ is a bounded interval, as

$$L_g(\gamma) = \int_I \sqrt{|g\left(\gamma'(t), \gamma'(t)\right)|} dt.$$

Remark. $L_g(\gamma)$ is re-parametrisation invariant.

Given the length of a curve, we can define a distance function in (M, g) so that (M, d_g) is a metric space, in the following way:

$$d_g(p, q) := \inf \left\{ L_{\mathbf{g}}(\gamma) \mid \gamma : [a, b] \to M, \gamma \text{ piecewise smooth}, \gamma(a) = p, \gamma(b) = q \right\}. \tag{1.4}$$

A curve γ achieving the minimum in (Equation 1.4) is called *geodesic*.

Now, we can ask: how does a change of basis modify the metric tensor? Suppose we are given another coordinate system $[\rho^i]$ at p and let us define $\tilde{\partial}_k = \frac{\partial}{\partial \rho^k}$, then, simply recalling (Equation 1.1), we have:

$$\left\langle \tilde{\partial}_k, \tilde{\partial}_\ell \right\rangle = \tilde{g}_{k\ell} = g_{ij} \left(\frac{\partial \xi^i}{\partial \rho^k} \right) \left(\frac{\partial \xi^j}{\partial \rho^\ell} \right)$$

and

$$g_{ij} = \tilde{g}_{k\ell} \left(\frac{\partial \rho^k}{\partial \xi^i} \right) \left(\frac{\partial \rho^\ell}{\partial \xi^j} \right),$$

(observe that there is a dependence on p everywhere in the previous formulas, but we will often “forget” to write it explicitly).

The coefficients $g_{ij}(p)$ form a square matrix $G(p)$, which is symmetric and positive definite, so, its inverse $G(p)^{-1}$ exists. Let $g^{ij}(p)$ be its ij—th element, then

$$g_{ij}g^{jk} = \delta_i^k = \begin{cases} 1 & (k = i) \\ 0 & (k \neq i) \end{cases}$$

from which we can also obtain the change-of-coordinates relations (as exercise).

On a Riemannian manifold we can also define the *gradient* of a differentiable f, denoted here by $\operatorname{grad} f$, as the vector field satisfying

$$g(v, \operatorname{grad} f) = df(v) \tag{1.5}$$

for all $v \in TM$.

1.1.4 Affine connections and covariant derivatives

In this section our goal is to compare tangent spaces $T_p(M)$ and $T_q(M)$, and the respective vectors, when $p \neq q \in M$ or, in general, to compare vector field $X, Y \in \mathfrak{X}(M)$ by giving a meaning to the derivative $\nabla_X Y$ of a vector field X w.r.t. the vector field Y.

Let us start with our an affine manifold $\mathbb{A}^n$. [PUT EXAMPLE HERE]

Definition 1.5.

Affine connection and covariant derivative

Let M be a differentiable manifold. An affine connection or covariant derivative operator ∇, is a map

$$\nabla : \mathfrak{X}(M) \times \mathfrak{X}(M) \ni (X, Y) \mapsto \nabla_X Y \in \mathfrak{X}(M)$$

which satisfies the following properties for every $p \in M$

i. $\left(\nabla_{fY+gZ}X\right)_p = f(p)\left(\nabla_Y X\right)_p + g(p)\left(\nabla_Z X\right)_p$ for all $f, g \in D(M)$ and vector fields $X, Y, Z \in \mathfrak{X}(M)$;
ii. $\left(\nabla_Y fX\right)_p = Y_p(f)X_p + f(p)\left(\nabla_Y X\right)_p$ for all $X, Y \in \mathfrak{X}(M)$ and $f \in D(M)$;
iii. $\left(\nabla_Y fX\right)_p = Y_p(f)X_p + f(p)\left(\nabla_Y X\right)_p$ for all scalars $a, b \in \mathbb{R}$ and $X, Y, Z \in \mathfrak{X}(M)$.

The contravariant vector field $\nabla_Y X$ is called the covariant derivative vector of X with respect to Y and the affine connection ∇.

Firstly, observe that $Y_p(f)$ indicates the directional derivative of a differentiable (real-valued) function f in the direction of the vector field Y, in $p \in M$. $Y_p(f) = df_p(Y) = df(Y_p)$, where $df_p : T_pM \to \mathbb{R}$ is the differential of f at p, see Definition 1.4.

Remark. The properties listed in the definition are **pointwise**. If two vector fields X and X' have the same value at p, i.e. $X_p = X'_p$ then $\left(\nabla_X Z\right)_p = \left(\nabla_{X'} Z\right)_p$. Similarly if $Y = Y'$ in a neighbourhood of p, then $\left(\nabla_X Y\right)_p = \left(\nabla_X Y'\right)_p$.

Connection coefficients

Let us consider, as usual, a local chart (U, ϕ) at $p \in U \subset M$ with coordinate $[\xi^i]$ for $i = 1, \dots, n$ and two vector fields $X, Y \in \mathfrak{X}(M)$, which we decompose w.r.t. $\partial_i|_p$. Then

$$\begin{aligned} \left(\nabla_X Y\right)_p &= X^i(p)Y^j(p)\nabla_{\partial_i|_p}\partial_j + X^i(p)\partial_i Y^j|_p\partial_j|_p \\ \text{using } \nabla_{\partial_i|_p}\partial_j &= \left\langle \nabla_{\partial_i}\partial_j, d^k \right\rangle \partial_k|_p := \Gamma^k_{ij}(p)\partial_k|_p \\ &= X^i_p\left(\partial_i Y^j|_p + \Gamma^k_{ij}(p)Y^j_p\right)\partial_k, \end{aligned} \tag{1.6}$$

where $\partial_i Y^j|_p = \left.\frac{\partial Y^j}{\partial \xi^i}\right|_p$.

For fixed $X \in \mathfrak{X}(M)$ and $p \in M$, the linear map $Y_p \mapsto (\nabla_{Y_p} X)_p$ (and a known result which guarantees that, for $p \in M$, if $t \in \mathcal{A}_{\mathbb{R}}(T_p M)$ then there exists a differentiable tensor field Ξ in M such that $\Xi_p = t$ [ADD CITATION HERE]) defines a tensor $(\nabla X)_p \in T_p^* M \otimes T_p M$ such that the only possible contraction of Y_p and $(\nabla X)_p$ is $(\nabla_{Y_p} X)_p$. Varying $M \ni p \mapsto (\nabla X)_p$ defines a smooth (1, 1) tensor field ∇X, which in local coordinates reads

$$\partial_i X^k + \Gamma^k_{ij} X^j$$

and is called **covariant derivative tensor** of X w.r.t. the affine connection ∇.

It can be proved that assigning an affine connection on a manifold M of dimension n is completely equivalent to giving the n^3 coefficients $\Gamma^k_{ij}(p)$ in each local coordinate system, as smooth functions w.r.t. p and transform according to [ADD REFERENCE TO EQUATION HERE].

Now that we have the concept of *affine connection*, let us introduce the *parallel transport* and derivation of vectors fields along curves.

According to Remark 2. it makes sense to define the derivative in p, $\nabla_{X_p} Y$, where X_p is a vector belonging to the tangent space of M at p. [PUT EXAMPLE AND FORMULA HERE]

- covariant derivative of tensor fields

We define

$$(\nabla \eta)_{ki} = \eta_{k,i} := \frac{\partial \eta_k}{\partial \xi^i} - \Gamma^r_{ik} \eta_r$$

as the covariant derivative (tensor) of the covariant vector field η. $\nabla \eta$ is the unique tensor field of type $(0, 2)$ such that the contraction of X_p and $(\nabla \eta)_p$ is $(\nabla_{X_p} \eta)_p$.

The coefficient $T^i_{jk} := \Gamma^i_{jk} - \Gamma^i_{kj}$ define the components of a tensor field, called the **torsion tensor field of the connection**

$$T = \left(\Gamma^i_{jk} - \Gamma^i_{kj}\right) \partial_i \otimes d^j \otimes d^k.$$

The torsion tensor at p is then, a bilinear map from $\mathfrak{X}(M) \times \mathfrak{X}(M)$ to a smooth vector field (same as for ∇), defined as

$$T_p(\nabla)\left(X_p, Y_p\right) = \nabla_{X_p} Y - \nabla_{Y_p} X - [X, Y]_p$$

If the tensor field T vanishes on M for every $X, Y \in \mathfrak{X}(M)$, i.e. $[X, Y] = \nabla_X Y - \nabla_Y X$, then ∇ is said to be **torsion free**.

Given $X, Y \in \mathfrak{X}(M)$, the term $[X, Y]_p$ is called *bracket* (or Lie bracket) and it is defined as the unique contravariant smooth vector field Z such that $Zf = (XY - YX)f = X(Y(f)) - Y(X(f))$

for each $f \in D(M)$. The bracket exists and is unique, see e.g. [Lemma 5.2; Carmo (1992)]. Or [Gallot, Hulin, Lafontaine].

Given a Riemannian manifold (M, g), there is a preferred (exactly one) affine connection ∇, which is torsion free and is completely determined by the metric, i.e. $\nabla g = 0$. This is the Levi-Civita connection. Its coefficients, called Christoffel's coefficients, are:

$$\Gamma^i_{jk} = \left\{ {i \atop jk} \right\} := \frac{1}{2} g^{is} \left(\frac{\partial g_{ks}}{\partial \xi^j} + \frac{\partial g_{sj}}{\partial \xi^k} - \frac{\partial g_{jk}}{\partial \xi^s} \right).$$

$\left\{ {i \atop jk} \right\}$ is called Christoffel's symbol.

Assume M is a smooth manifold with an affine connection ∇ and γ is a smooth curve from an open interval (a, b) to M. Then we say that the mapping $X : (a, b) \to T_{\gamma(t)}M$ defined by $t \mapsto X(t)$ is a smooth vector field along γ, if its components are smooth functions in every local chart at $\gamma(t)$ for every $t \in (a, b)$. Note that the case $X(t) = Y|_{\gamma(t)}$ for some vector field $Y \in \mathfrak{X}(M)$ is a special case of vector field along γ.

- add definitions of parallel transport in the different cases, with the connection applied to…

Flat manifolds

- affine coordinate system

Jun Zhang - tutorial

1.2 Other useful reminders

- Probability and statistics (exponential and mixture families, moments and moment generating function, score and likelihood, entropy…)
- Tensor algebra
- Calculus of variations (Euler-Lagrange equation)
- Optimal transport

2 Information geometry of statistical models

Let $\mathcal{X}$ be a set, we will now consider probability distributions on $\mathcal{X}$, i.e.

$$p : \mathcal{X} \to \mathbb{R}$$

such that $p(x) \geq 0$ for all $x \in \mathcal{X}$ and

i. $\sum_{x \in \mathcal{X}} p(x) = 1$ if $\mathcal{X}$ is a discrete set, or
ii. $\int_{\mathcal{X}} p(x)dx = 1$ (in this case p is a density function).

In general $(\mathcal{X}, \mathcal{B}, \nu)$ is a measurable space with $\sigma-$algebra (or Borel field) $\mathcal{B}$, ν a $\sigma-$finite measure. Given a probability measure P on $\mathcal{X}$ which is absolutely continuous w.r.t. ν, $p = \frac{dP}{d\nu} : \mathcal{X} \to \mathbb{R}$ is the Radon-Nikodym derivative of P. Here, we are interested in families of probability distributions on $\mathcal{X}$.

Consider a family S of probability distributions parametrised over a set of parameters $\Xi \subset \mathbb{R}^n$

$$S = \{p_\xi = p(x; \xi) : \xi = [\xi^1, \dots, \xi^n] \in \Xi\}$$

where the mapping (parametrisation) $\xi \mapsto p_\xi$ is injective. S is called an $n-$dimensional (parametric) **statistical model** on $\mathcal{X}$.

- given observations $x_1, \dots, x_n$ estimate the distribution generating the data p^* (true underlying distribution).
- p^* is unknown, but we often assume that it comes from a family of distributions, a model, and the problem becomes a parameter estimation.

Now, we want to add a differentiable structure to a statistical model S and use geometrical methods and arguments to approach usual statistical problems. Firstly, we need some **assumptions** allowing us to:

- Differentiate w.r.t. the model parameters $\to \Xi \subset \mathbb{R}$ is open and $\forall x \in \mathcal{X}$ the function $\xi \mapsto p(x; \xi)$ is smooth; we also assume that the order of integration and differentiation may be swapped.
- The support of the probability distributions does not vary with ξ, i.e. $\text{supp}(p_\xi) = \text{supp}(p) = \{x \in \mathcal{X} : p(x) > 0\}$.

We can then choose $\mathcal{X} = \mathrm{supp}(p)$ so that a statistical model S is a subset of

$$\mathcal{P}(\mathcal{X}) := \left\{ p : \mathcal{X} \to \mathbb{R} : p(x) > 0 \ \forall x \in \mathcal{X}, \int_{\mathcal{X}} p(x)\mathrm{d}x = 1 \right\}.$$

There are here some technicalities we are skipping, but we refer the reader to (Ay et al. 2017).

Given a statistical model $S = \{p_\xi : \xi \in \Xi\}$, we can take a global chart

$$\begin{aligned} \varphi : S &\to \mathbb{R}^n \\ p_\xi &\mapsto \xi \end{aligned}$$

so that ξ^i define our (global) coordinate system for S. Observe that each re-parametrisation of the model $\psi : \Xi \to \psi(\Xi) \subset \mathbb{R}^n$, where ψ is a smooth diffeomorphism, provides another equivalent (global) coordinate system for S, i.e. $\rho = \psi(\xi)$ and $S = \{p_{\psi^{-1}(\rho)} : \rho \in \psi(\Xi)\}$. We can then consider S as a differentiable manifold, called a **statistical manifold**.

2.0.1 The Fisher metric

Let S be an $n-$dimensional statistical manifold, given a point $p_\xi \in S$, or, as we will henceforth write, given a point ξ ($\in \Xi$), the Fisher information matrix of S at ξ is a $n \times n$ matrix $G(\xi) = (g_{ij}(\xi))$ defined by

$$g_{ij}(\xi) := \mathbb{E}_\xi \left[\partial_i \ell_\xi \partial_j \ell_\xi \right] = \int \partial_i \ell(x;\xi) \partial_j \ell(x;\xi) p(x;\xi) \mathrm{d}x \tag{2.1}$$

where $\partial_i = \frac{\partial}{\partial \xi^i}$, $\mathbb{E}_\xi$ denotes the expectation w.r.t. p_ξ, and $\ell_\xi(x) = \ell(x;\xi) = \log p(x;\xi)$. Assuming that the integral is finite

FIM as the covariance matrix of the score, which is symmetric and positive semi-definite (could be undefined too). When it is positive definite, it is said regular, and yields the Fisher metric on manifolds. Observe that here, due to the regularity assumption that we can exchange differentiation and integration, the expectation of the score is 0 and so we write the FIM using only the expectation. Bartlett identities??

- Probability simplex, affine space, statistical bundle.

3 Non-parametric information geometry

Geometrising a problem or a field should, in principle, provide tools which do not depend from parametrisations. Hence, it makes sense that non-parametric models should be the main object of interest in IG. Of course, dealing with infinite-dimensional spaces is not always easy (or at our reach), but we can still introduce the methods and results of a non-parametric IG in the finite-dimensional case. In this way, the finite-dimensional (parametric) theory is derived from the infinite-dimensional (non-parametric) one.

The main references here are(Pistone 2013, 2019).

- Open probability simplex
- Affine structure
- Tangent space to p: variables with zero expected value w.r.t. p.

The set of probability functions over a finite sample space X is the probability simplex. This can be seen as the set generated by $\delta-$functions, centred at each point $x \in X$. $\mathcal{P}(X)$ is a convex subset of $\mathbb{R}^X$, or, also, a convex subset of the affine space $p \in \mathbb{R}^X : \sum_{x \in X} p(x) = 1$.

4 Summary

In summary, this book has no content whatsoever.

References

Amari, Shun-ichi, and Hiroshi Nagaoka. 2000. *Methods of Information Geometry*. Vol. 191. American Mathematical Soc.

Ay, Nihat, Jürgen Jost, Hông Vân Lê, and Lorenz Schwachhöfer. 2017. *Information Geometry*. Vol. 64. Springer.

Carmo, Manfredo Perdigão do. 1992. *Riemannian Geometry*. Mathematics: Theory & Applications. Boston, Mass. [etc: Birkhäuser.

Chentsov, Nikolai Nikolaevich. 1982. "Statiscal Decision Rules and Optimal Inference." *Monog* 53.

Efron, Bradley. 1975. "Defining the Curvature of a Statistical Problem (with Applications to Second Order Efficiency)." *The Annals of Statistics* 3 (6): 1189–1242. https://www.jstor.org/stable/2958246.

Lang, Serge. 2012. *Differential and Riemannian Manifolds*. Vol. 160. Springer Science & Business Media.

Petersen, Peter. 2006. *Riemannian Geometry*. Vol. 171. Springer.

Pistone, Giovanni. 2013. "Nonparametric Information Geometry." arXiv. http://arxiv.org/abs/1306.0480.

———. 2019. "Information Geometry of the Probability Simplex: A Short Course." *arXiv: Statistics Theory*, November. https://doi.org/10.33581/1561-4085-2020-23-2-221-242.

Rao, C Radhakrishna. 1945. "Information and the Accuracy Attainable in the Estimation of Statistical Parameters." *Reson. J. Sci. Educ* 20: 78–90.

Sernesi, Edoardo. 1994. "Geometria 2 Bollati Boringhieri." Torino.