An Introduction to Information Geometry

Giulia Bertagnolli

05/10/2022

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 & 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, Jost, Vân Lê, & Schwachhöfer, 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. *Useful*, but not mandatory, as we will see in Section 1.6.

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 (Lang, 2012; Petersen, 2006; Sernesi, 1994) 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 U_α is open and the φ_α are topological isomorphisms (or *homeomorphism*, i.e. bijectinve and bi-continuous). In other words, a differentiable structure on M induced a topology on it.

[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).

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 (or the maximal family, see books on differential geometry for these technicalities).
iii. (Carmo, 1992) gives an equivalent definition of differentiable manifold, which is more intuitive when we look at parametric models. In this reference a *parametrisation* of M, a set, at a point $p \in M$ is a pair (U_α, x_α), where $p \in U_\alpha \subset \mathbb{R}^n$, U_α is an open set in $\mathbb{R}^n$ and $x_\alpha :\to M$ is an injective map. Then, as before we form an atlas and this provides the differentiable structure to the set M.
iv. 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).

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 $\varphi(p) = (\xi^1(p), \ldots, \xi^n(p))$. 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 $(\mathbb{A}^n, V, \vec{\cdot})$, 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 $\vec{\cdot}$ 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.

Example: Projective manifold (Carmo, 1992)

The real projective space $P^n(\mathbb{R})$ is the set of straight lines of $\mathbb{R}^{n+1}$ passing through the origin $0 \in \mathbb{R}^{n+1}$—the set of linear subspaces of dimension 1 of $\mathbb{R}^{n+1}$ (Sernesi, 1994).

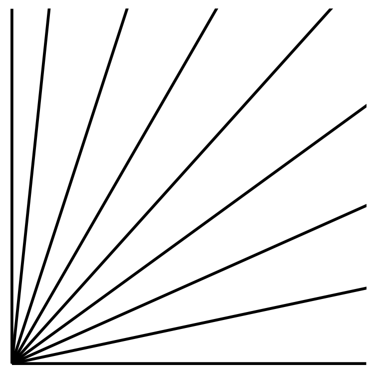

Figure 1.1: Some points in the projective plane.

For each $(x_1, \dots, x_{n+1}) \in \mathbb{R}^{n+1}$, the equivalence class $[x_1, \dots, x_{n+1}] = [\lambda(x_1, \dots, x_{n+1}) : \lambda \in \mathbb{R}]$ defines a point of the projective space. Consider the subsets (coordinate neighbourhoods) $V_i = \{[x_1, \dots, x_{n+1}] : x_i \neq 0\}$, $i = 1, \dots, n+1$, of directions not belonging to the i–th coordinate hyperplane. In V_i we have that $[x_1, \dots, x_{n+1}] = [\frac{x_1}{x_i}, \dots, 1, \dots, \frac{x_{n+1}}{x_i}]$, so we can define the mapping $\mathbf{x}_i : \mathbb{R}^n \to V_i$ by $\mathbf{x}_i\,(y_1, \dots, y_n) = [y_1, \dots, y_{i-1}, 1, y_i, \dots, y_n]$. The family $\{(R^n, \mathbf{x}_i)\}_{i=1,\dots,n+1}$ defines a differentiable structure on $P^n(\mathbb{R})$. The coordinates in each V_i are

$$\left(\frac{x_1}{x_i}, \dots, \frac{x_{i-1}}{x_i}, \frac{x_{i+1}}{x_i}, \dots, \frac{x_{n+1}}{x_i}\right)$$

and are called *inhomogeneous coordinates* corresponding to the homogeneous ones $(x_1, \dots, x_{n+1}) \in \mathbb{R}^{n+1}$. See (Carmo, 1992) for more details.

Now, $P^n(\mathbb{R})$ can also be seen as the quotient space of the unit sphere $S^n = \{p \in \mathbb{R}^{n+1} : \|p\| = 1\}$ by the equivalence relation $p \sim -p$ (identification of antipodal points). We can hence introduce another differentiable structure on the n-dimensional projective space, taking advantage from the parametrisation of S^n and then projecting it to $P^n(\mathbb{R})$ through the canonical projection $\pi : S^n \to P^n(\mathbb{R})$, see (Carmo, 1992) for details.

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 $\mathbb{R}^n$.

The directional derivative of a function $f : \mathbb{R}^n \supseteq A \to \mathbb{R}^n$, with A open, at $x_0 \in A$, in direction $v \in \mathbb{R}^n$ such that $\|v\| = 1$, is defined, if the limit exists and is finite, as

$$D_v f(x_0) := \lim_{t \to 0} \frac{f(x_0 + tv)}{t}$$

In the same way we can define the partial derivatives w.r.t. the $i-$th coordinate, as $\frac{\partial}{\partial x_i} f(x_0) := D_{e_i} f(x_0)$ and the gradient in of f is then

$$\nabla f(x) = \left(\frac{\partial}{\partial x_1} f(x), \dots, \frac{\partial}{\partial x_n} f(x) \right)$$

Recall the difference between *derivabilità [IT]* at a point and *differenziabilità [IT]* (differentiability).

With this identification of vectors with (directional) derivatives in mind, let us define the tangent spaces of a manifold M. Here, the charts allow us to carry over ideas of the "usual" differential calculus in the Euclidean space to our manifold.

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[2]

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

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_pM 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.

It can be proved that the space of all derivations on M at p coincides with T_pM, i.e. $\left\{\frac{\partial}{\partial \xi^i}\right\}_i$ is a basis for $\mathcal{D}_p(M)$. T_pM has the same dimension of the manifold.

The motivation for calling these spaces, *tangent spaces* comes from the following equivalent definition (Carmo, 1992)

Definition 1.3 (Tangent vector to a curve)**.** Let $\gamma : (-\epsilon, \epsilon) \to M$ be a differentiable curve on M such that $\gamma(0) = p \in M$. The tangent vector to γ at $t = 0$ is the linear map $\gamma'(0) : D(M) \to \mathbb{R}$ defined as follows

$$\gamma'(0)f := \left.\frac{d}{dt}(f \circ \gamma)\right|_{t=0}$$

for each $f \in D(M)$.

A tangent vector at $p \in M$ is then the tangent vector at $t = 0$ to some curve γ satisfying $\gamma(0) = p$.

[2] Einstein summation is use throughout these notes.

The equivalence between the two definitions of tangent spaces can be seen choosing a parametrisation $(U, \mathbf{x})$ at $\mathbf{x}(0) = p \in M$ and expressing both the function f and the curve γ in the parametrisation:

$$\begin{aligned} f \circ \mathbf{x} : &\quad (x^1, \dots, x^n) \mapsto f(x^1, \dots, x^n) \\ \mathbf{x}^{-1} \circ \gamma : &\quad t \mapsto \left(x^1(t), \dots, x^n(t)\right). \end{aligned}$$

Then, $f \circ \gamma = f \circ \mathbf{x} \circ \mathbf{x}^{-1} \circ \gamma = f\left(x^1(t), \dots, x^n(t)\right)$ and

$$\gamma'(0) f = \frac{d}{dt} f\left(x^1(t), \dots, x^n(t)\right)\bigg|_{t=0} = \dot{x}^i(0) \left.\frac{\partial}{\partial x^i}\right|_0 f$$

and again we find the vectors $\left.\frac{\partial}{\partial x^i}\right|_0$, which are here the tangent vectors at p of the coordinate curves $x^i \to \mathbf{x}(0, \dots, 0, x^i, 0, \dots, 0)$.

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 1.4 (Cotangent space)**.** The dual space T_p^*M is called cotangent space of M at p and its elements are called 1-forms in p, or covectors, or covariant vectors. If $[\xi^i]$ are local coordinates at p, the dual basis to $\left\{ \left.\frac{\partial}{\partial \xi^i}\right|_p \right\}$ is denoted by $\{d\xi^i|_p\}$ and it holds $d\xi^i\left(\frac{\partial}{\partial \xi^j}\right) = \delta^i_j$.

Definition 1.5 (Tangent and cotangent bundles)**.**

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

Definition 1.6 (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

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

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

Observe that $g \circ f \in D(M|\mathbb{R})$.

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 $[\xi^i]$. Each vector field X may be written as $X_p = X^i_p \partial_i|_p$, where $\partial_i := \frac{\partial}{\partial \xi^i}$ and X^i_p, for $i = 1, \dots, n$, are the scalar components of X w.r.t. the coordinate system $[\xi^i]$. The change of coordinates, here, is the same as in (Equation 1.1).

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^*_pM, and $\mathbb{R}$, is said to by a tensor field.

[TBD] Recap on: - multi-linear maps - tensor products of vector spaces - tensor algebra generated by a vector space, its dual space, and its field.

In the framework of tensors, each $\left.\frac{\partial}{\partial \xi^i}\right|_p$ is a tensor of T_pM, i.e. a contravariant vector. Tensors of the dual tangent space T^*_pM are covariant vectors. The canonical bases of $\mathcal{A}_{\mathbb{R}}(T_pM)$ are given by tensor products of $\left\{ \left.\frac{\partial}{\partial \xi^i}\right|_p \right\}$ and $\{d\xi^i|_p\}$.

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

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

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

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

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 $X_p(f) := X^i(p) \left.\frac{\partial f}{\partial \xi^i}\right|_p$. In general, every smooth vector field X defines a linear mapping from $D(M)$ to $D(M)$ by $f \mapsto X(f)$, where $X(f)(p) =: X_p(f)$ for 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 covariant 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 $g : p \mapsto \langle\ ,\ \rangle_p \in [T_pM]^0_2$ is a covariant tensor field of order 2. Equivalently, assume we have a smooth covariant tensor field g 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 a **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.

Examples

(1) The canonical metric in the Euclidean space $\mathbb{R}^n$ is $g_{ij} = \langle e_i, e_j\rangle = \delta_{ij}$ so that the matrix representation of the metric is the identity and, in Cartesian coordinates, $g = \delta_{ij}dx^i dx^j = \sum_{i=1}^n \left(dx^i\right)^2$.

(2) The product metric. Let M_1, M_2 be Riemannian manifolds and consider their product $M_1 \times M_2$ with the natural projections $\pi_i : M_1 \times M_2 \to M_i$ for $i = 1, 2$. Then a Rimannian metric on $M_1 \times M_2$ can be introduced by

$$\langle u, v\rangle_{(p,q)} = \langle d\pi_1(u), d\pi_1(v)\rangle_p + \langle d\pi_2(u), d\pi_2(v)\rangle_q\,,$$

where we use the differentials of the projections to push forward the derivations/vectors in the tangent space $T_{(p,q)}M_1 \times M_2$ to M_1 and M_2 accordingly.

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. To proof this result one needs to introduce a smooth partition of the unity. We refer the interested reader to (Lang, 2012; Sernesi, 1994).

□

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 (also called *local representation of the Riemannian metric* in the chart) 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 $v = v^i \partial_i|_p, w = w^i \partial_i|_p \in T_pM$;
- and the norm of any $v^i \partial_i|_p = v \in T_pM$ 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

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

and its **energy**

$$E_g(\gamma) = \int_I |g(\gamma'(t), \gamma'(t))| \, dt.$$

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 $v \in TM$.

Those who are familiar with exterior calculus will notice that the metric one to transform a 1-form, the differential, into a 1-vector, the gradient.

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 $\mathbb{A}^n$.

Definition 1.7 (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 1.6.

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 $Y_p \mapsto \left(\nabla_{Y_p} X\right)_p$ (and a known result which guarantees that, for $p \in M$, if $t \in \mathcal{A}_{\mathbb{R}}(T_pM)$ then there exists a differentiable tensor field Ξ in M such that $\Xi_p = t$ [ADD CITATION HERE]) defines a tensor $(\nabla X)_p \in T^*_pM \otimes T_pM$ 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...

1.2 Information geometry of statistical models

Let Ω be a set, we will now consider probability functions on Ω, i.e.

$$p : \Omega \to \mathbb{R}$$

such that $p(x) \geq 0$ for all $x \in \Omega$ and

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

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

Now, we have to distinguish different case:

i. If Ω is finite, then we can consider the real algebra of functions $f : \Omega \to \mathbb{R}$, which is a finite dimensional vector space over $\mathbb{R}$, denoted also by $\mathbb{R}^\Omega$, with the additional product operation (and additional axioms). Linear forms over $\mathbb{R}^\Omega$ are interpreted as signed measures (usual duality between functions and measures) and, since in $\mathbb{R}^\Omega$ there exists a preferred basis of function—$e_x(y) \in \{0, 1\}$ if x is different or equal to y, respectively—there is a natural isomorphism between the dual space of signed measures and $\mathbb{R}^\Omega$. Here we will introduce the probability simplex in section Section 1.6.
ii. If Ω is an infinite sample space, we will need also the $\sigma-$algebra $\mathcal{B}$ of subsets of Ω. In this case, $\mathbb{R}^\Omega$ is an infinite-dimensional functional space and we can consider either finite dimensional manifolds, through charts to/parametrisations from $\mathbb{R}^n$, or consider the infinite-dimensional non-parametric case. We will focus on the first.

Consider now a family M of probability distributions parametrised over a set of parameters $\Xi \subset \mathbb{R}^k$

$$M = \{p_\xi = p(\cdot; \xi) : \xi = [\xi^1, \dots, \xi^k] \in \Xi\}$$

where the mapping $p : \xi \mapsto p_\xi = p(\cdot; \xi)$ is injective (and we will also assume to have some regularity property). M is called an $k-$dimensional (parametric) **statistical model** on Ω. p plays the role of a parametrisation of the differentiable structure of (Carmo, 1992) and the $[\xi^i]$ define a (global) coordinate system for M. Each re-parametrisation of the model $\psi : \xi \to \psi(\xi) \subset \mathbb{R}^k$, where ψ is a smooth diffeomorphism, provides another equivalent (global) coordinate system for M, i.e. $\rho = \psi(\xi)$ and $M = \{p_{\psi^{-1}(\rho)} : \rho \in \psi(\xi)\}$. We can then consider M as a differentiable manifold, called a **statistical manifold**.

In the following we will mainly write $\xi^i(p) \in \mathbb{R}^k$ for $p \in M$, so that the ξ^is are the coordinates of the chart, instead of the parametrisation, to use a consistent notation with (Amari, 2016; Amari & Nagaoka, 2000). We will also, as it is commonly done in the field, denote by $T_\xi M$ the tangent space of M at p_ξ, and simply say "at ξ".

Observations

i. As we have defined it $p(\cdot,\xi)$ is a density function, the measure is $\mathbf{p}(\xi) = p(\cdot,\xi)\nu$ and ν does not depend on the parameter.
ii. $p : \Omega \times \Xi \to \mathbb{R}$ and we need some regularity assumptions w.r.t. x to compute, e.g., expected values. One possibility is $p_\xi \in L^1(\Omega,\nu)$ for all ξ. See (Ay et al., 2017) for more details.

Before moving on, some motivations. A typical problem in statistics is:

- 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 M, and the problem becomes a parameter estimation problem (assuming, possibly approximately, *faithfulness*).

Hence, it seems reasonable to focus on geometric properties of such **models**. By means of tangent spaces and their geometry (e.g. the Riemannian metric), we can study local properties of a statistical model. Affine connections, on the other hand, allow us to say something about the global geometry (e.g. curvature) of the model, since the establish a 1-to-1 affine correspondence between tangent spaces at different points. Originally, the works by Chentsov (1972), Efron (1975, 1978) and also Amari (1980, 1982) were motivated by the interest in higher-order asymptotic properties of inference. Nowadays, the affine/differential geometric approach of IG shows its usefulness in many applications, e.g. the natural gradient is well-known in optimisation and machine learning.

Assumptions on M

We assume that for each $x \in \Omega$ the parametrisation $\xi \to p(x;\xi)$ from the open (parameter) set $\Xi \subseteq \mathbb{R}^n$ to $\mathbb{R}$ is smooth (C^∞). This allows us to differentiate w.r.t. the parameter. We also assume that the order of integration and differentiation may be swapped, so that

$$\int_\Omega \frac{\partial}{\partial \xi_i} p(x;\xi)\nu(dx) = \frac{\partial}{\partial \xi_i} \int_\Omega p(x;\xi)\nu(dx) = 0 \tag{1.7}$$

Finally, we also assume that support of p, $\operatorname{supp}(p) = \{x \in \Omega : p(x;\xi) > 0\}$ does not depend on ξ. This means that, if we choose $\Omega = \operatorname{supp}(p)$, then $M \subset \mathcal{P}_>(\Omega)$, where

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

1.3 Fisher information

Recall the definition of the Fisher information matrix of a statistical model M.

Definition 1.8 (Fisher information)**.** Let $M = \{p(x;\xi) : \xi \in \Xi\}$ be an $k-$dimensional statistical model. The Fisher information matrix (FIM) of M at $\xi \in \Xi$ is the $k \times k$ matrix $I(\xi)$, whose $ij-$element is given by

$$g_{ij}(\xi) = \mathbb{E}_\xi\left[\frac{\partial}{\partial \xi^i}\ell(\xi;x)\frac{\partial}{\partial \xi^j}\ell(\xi;x)\right] = \int \partial_i \ell(\xi;x)\partial_j \ell(\xi;x)p(x;\xi)\nu(dx) \tag{1.8}$$

where we use the notation $\partial_i = \frac{\partial}{\partial \xi^i}$ and $\ell(\xi;x)$ denotes the log-likelihood, or $\ell(\xi;x) = \log p(x;\xi)$, depending on whether we have a random variable x or a random sample $(x_1, \dots, x_n)$.

Remarks

i. It is possible to write $g_{ij}(\xi)$ in other forms:
   - $g_{ij}(\xi) = -E_\xi\left[\partial_i \partial_j \ell(\xi;x)\right]$—use eq. Equation 1.7 to prove;
   - $g_{ij}(\xi) = 4\int \partial_i \sqrt{p(x;\xi)}\partial_j \sqrt{p(x;\xi)}\nu(dx)$.

ii. The FIM $I(\xi)$ is symmetric and positive semi-definite, i.e. for $\mathbb{R}^k \ni v = v^i e_i$ with components:
$$v^t I(\xi) v = \int_\Omega \left\{v^i \partial_i \ell(\xi;x)\right\}^2 p(x;\xi)\nu(dx) \geq 0,$$

iii. Assuming $I(\xi)$ be positive definite, means $v^i \partial_i \ell(\xi;x) \neq 0$, i.e. the vectors $\partial_i \ell(\xi;x) = \frac{\partial_i p(x;\xi)}{p(x;\xi)}$ be linearly independent, i.e. $\partial_i p(x;\xi)$ are linearly independent.

iv. For $g_{ij}(\xi)$ to be finite, we need at least $\partial_i \ell(\xi;\cdot) \in L^2(\Omega,\nu)$ for all ξ. We also assume also that $g_{ij} : \Xi \to \mathbb{R}$ is smooth (i.e. that we have a smooth covariant tensor of order 2).

Now, each $\xi \mapsto p(\cdot;0,\dots,0,\xi^i,0,\dots,0)$ is both, a one-dimensional model, and a curve on M, which can be used as coordinate curves on the statistical manifold. Let us re-write it in a more compact form: consider a differentiable curve from an open set $\mathbb{R} \supset I \ni \theta \mapsto p(\theta)$ to M and the tangent vector to the curve at $\theta = \theta_0$,

$$\dot{p}(\theta_0) = \lim_{h\to 0}\frac{p(\theta_0 + h) - p(\theta_0)}{h}.$$

This tangent vector measures the variation w.r.t. the base measure ν. If, otherwise we consider $\ell(\theta;\cdot) = \log p(\theta;\cdot)$ as a coordinate curve, then

$$\frac{d}{d\theta}\log p(\theta;\cdot) = \frac{\dot{p}(\theta;\cdot)}{p(\theta;\cdot)}$$

the tangent vector is Fisher's score, which measures the relative variation of p w.r.t. itself. That is, by assuming the existence of Fisher's score, we are assuming that $\dot{p}$ is absolutely continuous w.r.t. p ($\dot{p}(\theta) \ll p(\theta)$). We will formalise it in the next chapter.

By Remark-iii. we have that both systems of tangent vectors are good candidates for a basis of the tangent space of M at ξ. Choosing as basis the one provided by Fisher's score, we obtain that the FIM is the matrix representation of **the Fisher metric**, which is a Riemannian metric for Remark-ii. It can be seen that the Fisher metric is invariant over the choice of coordinate system. Observe that $\mathbb{E}_{\xi}[\partial_i \ell(\xi; x)] = 0$, this gives us an intuition about the probabilistic/statistical meaning of tangent vectors: tangent vectors are random variables, which are centred w.r.t. the distribution at ξ.

Recall the definition of **sufficient statistic**: let $\mathbf{x} = (x_1, \dots, x_n)$ be a random sample and let κ be a statistic. Then κ is sufficient for ξ if the density of the vector $\mathbf{x}$ conditioned on the value of the statistic, $p(\mathbf{x}|\kappa(x); \xi)$ does not depend on ξ. Then, from Neyman's factorisation theorem we have the characterisation

$$\begin{aligned}\kappa : \Omega \to \Omega' \text{ sufficient for } \xi \;\Leftrightarrow\; & \exists\, s : \Omega' \times \Xi \to \mathbb{R}, \exists\, t : \Omega \to \mathbb{R} \text{ s. t.} \\ & p(x; \xi) = s(\kappa(x); \xi) t(x),\ \forall x, \xi.\end{aligned}$$

There is also a characterisation of a sufficient statistic in terms of the FIM, that is $I(\xi)$ is invariant under sufficient-statistic transformation, see e.g., (Amari & Nagaoka, 2000, thm. 2.1).

Other well-known results are * Monotonicity, or chain rule: $I_\kappa(\xi) \le I(\xi)$ i.e. the difference matrix $I(\xi) - I_\kappa(\xi)$ is positive semi-definite. * Rao-Cramér inequality: $\mathbb{V}(\hat{\xi}(X)) \ge I(\xi)^{-1}$, where $\hat{\xi}$ is an estimator of ξ. When the equality holds, $\hat{\xi}$ is an efficient estimator of ξ. * An efficient estimator does not always exist, unless we impose additional assumptions on the model M and on the parametrisation $\xi \mapsto p(\cdot; \xi)$. * There always exists a sequence of asymptotically efficient estimators. The matrix $I(\xi)^{-1}$ quantifies the fluctuations of the asymptotic efficient estimator around the true value ξ.

1.4 The exponential and mixture families

Recall that a probability density p on Ω belongs to an the exponential family with k parameters if there exist functions $C, \{B_i\}_{i=1}^k$ on Ω and ψ on Θ such that

$$p(x; \theta) = \exp\{C(x) + \theta^i B_i(x) - \psi(\theta)\}.$$

(Remember that we use the Einstein summation over repeated indices convention).

$[\theta^i]$ are called natural, or canonical, parameters.

On the other hand, if p can be expressed as

$$p(x;\theta) = C(x) + \theta^i B_i(x)$$

than M is a mixture family with mixture parameters $[\theta^i]$.

When Ω is finite, $\mathcal{P}_>(\Omega)$ is a mixture family (Amari & Nagaoka, 2000, pag. 35).

1.5 Exponential map

Geodesics (exponential families) and mixture models.

1.6 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, 2020).

- 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 $p \in \mathbb{R}^X : \sum_{x \in X} p(x) = 1$.

2 Summary

In summary, this book has no content whatsoever.

References

Amari, S. (2016). *Information geometry and its applications* (Vol. 194). Springer. http://doi.org/10.1007/978-4-431-55978-8

Amari, S., & Nagaoka, H. (2000). *Methods of information geometry* (Vol. 191). American Mathematical Soc. http://doi.org/10.1090/mmono/191

Ay, N., Jost, J., Vân Lê, H., & Schwachhöfer, L. (2017). *Information geometry* (Vol. 64). Springer. http://doi.org/10.1007/978-3-319-56478-4

Carmo, M. P. do. (1992). *Riemannian geometry.* Boston, Mass. [etc: Birkhäuser. http://doi.org/10.1007/978-1-4757-2201-7

Chentsov, N. N. (1982). Statiscal decision rules and optimal inference. *Monog*, *53*.

Efron, B. (1975). Defining the Curvature of a Statistical Problem (with Applications to Second Order Efficiency). *The Annals of Statistics*, *3*(6), 1189–1242. http://doi.org/10.1214/aos/1176343282

Lang, S. (2012). *Differential and riemannian manifolds* (Vol. 160). Springer Science & Business Media.

Petersen, P. (2006). *Riemannian geometry* (Vol. 171). Springer.

Pistone, G. (2013, July). Nonparametric Information Geometry. arXiv. Retrieved from https://arxiv.org/abs/1306.0480

Pistone, G. (2020). Information geometry of the probability simplex: A short course. *Nonlinear Phenomena in Complex Systems*, *23*(2), 221–242. http://doi.org/10.33581/1561-4085-2020-23-2-221-242

Rao, C. R. (1945). Information and the accuracy attainable in the estimation of statistical parameters. *Reson. J. Sci. Educ*, *20*, 78–90.

Sernesi, E. (1994). Geometria 2 bollati boringhieri. Torino.