Self-study: The geometry of exponential families

Tianpei Xie

Jun. 1st., 2015

Contents

1 The statistical manifold and the exponential families

1.1 Concepts

- A ***statistical manifold*** [Murray and Rice, 1993, Amari and Nagaoka, 2007]: Any parametrised family of probability distributions,

$$\mathcal{P} = \{p(\boldsymbol{x}; \boldsymbol{\theta})\}$$

with parameter $\boldsymbol{\theta}$ running over some open subset of $\mathbb{R}^d$, is automatically a ***manifold***, in which the probability distributions are the points of the manifold and the parameters are co-ordinate functions.

- **Definition** Let p, q be $\mathbb{R}^d \supset \mathcal{M}_0 :\rightarrow \mathbb{R}$ density functions and let $\alpha \in \mathbb{R} \setminus \{1\}$. The ***Rényi divergence*** of order α or $\underline{\alpha\text{−}\textbf{divergence}}$ of a distribution p from a distribution q is defined to be

$$\mathbb{D}^{\alpha}\left(p \,\|\, q\right) = \frac{1}{\alpha - 1} \log \left[\mathbb{E}_Q \left[\left(\frac{dP}{dQ} \right)^{\alpha} \right] \right] = \frac{1}{\alpha - 1} \log \left(\int_{\mathcal{M}_0} p^{\alpha}(x) q^{1-\alpha}(x) \; \mu(dx) \right) \tag{1}$$

- **Definition** Let P and Q be two probability distributions over a space Ω, such that $P \ll Q$, that is, P is ***absolutely continuous*** with respect to Q. Then, for a ***convex function*** f : $[0, +\infty) \rightarrow (-\infty, +\infty]$ such that $f(x)$ is finite for all $x > 0$, $\underline{f(1) = 0}$, and $\underline{f(0) = \lim_{t \rightarrow 0^+} f(t)}$ (which could be infinite), the ***f-divergence*** of P from Q is defined as

$$\mathbb{D}^{f}\left(P \,\|\, Q\right) = \mathbb{E}_Q \left[f \left(\frac{dP}{dQ} \right) \right] = \int_{\Omega} f \left(\frac{dP}{dQ} \right) dQ = \int_{\Omega} q(x) f \left(\frac{p(x)}{q(x)} \right) \mu(dx) \tag{2}$$

1.2 Exponential families

- The canonical representation of ***exponential famlity*** of distribution has the following form

$$\begin{aligned} p(x_1, \ldots, x_m) = p(\boldsymbol{x}; \boldsymbol{\eta}) &= \exp\left(\langle \boldsymbol{\eta} \,,\, \boldsymbol{\phi}(\boldsymbol{x}) \rangle - A(\boldsymbol{\eta}) \right) \mu(d\boldsymbol{x}) \\ &= \exp \left(\sum_{i=1}^{d} \eta_i \phi_i(\boldsymbol{x}) - A(\boldsymbol{\eta}) \right) \mu(d\boldsymbol{x}) \end{aligned} \tag{3}$$

where ϕ is a feature map and $\boldsymbol{\phi}(\boldsymbol{x})$ defines a set of ***sufficient statistics*** (or ***potential functions***). The normalization factor is defined as

$$A(\boldsymbol{\eta}) := \log \int \exp\left(\langle \boldsymbol{\eta} \,,\, \boldsymbol{\phi}(\boldsymbol{x}) \rangle \right) h(\boldsymbol{x}) \nu(d\boldsymbol{x}) = \log Z(\boldsymbol{\eta})$$

$A(\boldsymbol{\eta})$ is also referred as ***log-partition function*** or *cumulant function*. The parameters $\boldsymbol{\eta} = (\eta_{\alpha})$ are called ***natural parameters*** or *canonical parameters*. The canonical parameter $\{\eta_{\alpha}\}$ forms a **natural (canonical) parameter space**

$$\Omega = \left\{ \boldsymbol{\eta} \in \mathbb{R}^d : A(\boldsymbol{\eta}) < \infty \right\} \tag{4}$$

- The exponential family is the unique solution of ***maximum entropy estimation*** problem:

$$\min_{q \in \Delta} \quad \mathbb{KL}\left(q \,\|\, p_0\right) \tag{5}$$

$$\text{s.t.} \quad \mathbb{E}_q\left[\phi_\alpha(X)\right] = \mu_\alpha \quad \forall\, \alpha \in \mathcal{I} \tag{6}$$

where $\mathbb{KL}\left(q \,\|\, p_0\right) = \int \log(\frac{q}{p_0}) q dx = \mathbb{E}_q\left[\log \frac{q}{p_0}\right]$ is the relative entropy or the Kullback-Leibler divergence of q w.r.t. p_0.

Here $\boldsymbol{\mu} = (\mu_\alpha)_{\alpha \in \mathcal{I}}$ is a set of ***mean parameters***. The space of mean parameters $\mathcal{M}$ is a *convex polytope* spanned by potential functions $\{\phi_\alpha\}$.

$$\mathcal{M} := \left\{\boldsymbol{\mu} \in \mathbb{R}^d : \exists q \text{ s.t. } \mathbb{E}_q\left[\phi_\alpha(X)\right] = \mu_\alpha \quad \forall\, \alpha \in \mathcal{I}\right\} = \text{conv}\left\{\phi_\alpha(x),\ x \in \mathcal{X},\ \alpha \in \mathcal{I}\right\} \tag{7}$$

- Note that $A(\boldsymbol{\eta})$ is a convex function and its gradient $\nabla A : \Omega \to \mathcal{M}^{\circ}$ is a bijection between the natural parameter space Ω and the **interior** of $\mathcal{M}$, $\mathcal{M}^{\circ}$; $\nabla A(\boldsymbol{\eta}) = \boldsymbol{\mu}$ based on the following equation

$$\frac{\partial A}{\partial \eta_\alpha} = \mathbb{E}_{\boldsymbol{\eta}}\left[\phi_\alpha(X)\right] := \int_{\mathcal{X}^m} \phi_\alpha(\boldsymbol{x}) q(\boldsymbol{x}; \boldsymbol{\eta}) d\boldsymbol{x} = \mu_\alpha \tag{8}$$

- Moreover $A(\boldsymbol{\eta})$ has a variational form

$$A(\boldsymbol{\eta}) = \sup_{\boldsymbol{\mu} \in \mathcal{M}} \left\{\langle \boldsymbol{\eta}\,,\, \boldsymbol{\mu} \rangle - A^*(\boldsymbol{\mu})\right\} \tag{9}$$

where $A^*(\boldsymbol{\mu})$ is the conjugate dual function of A and it is defined as

$$A^*(\boldsymbol{\mu}) := \sup_{\boldsymbol{\eta} \in \Omega} \left\{\langle \boldsymbol{\mu}\,,\, \boldsymbol{\eta} \rangle - A(\boldsymbol{\eta})\right\} \tag{10}$$

It is shown that $A^*(\boldsymbol{\mu}) = -H(q_{\boldsymbol{\eta}(\boldsymbol{\mu})})$ for $\boldsymbol{\mu} \in \mathcal{M}^{\circ}$ which is the negative entropy. $A^*(\boldsymbol{\mu})$ is also the optimal value for the **maximum likelihood estimation** problem on p. The exponential family can be reparameterized according to its mean parameters $\boldsymbol{\mu}$ via backward mapping $(\nabla A)^{-1} : \mathcal{M}^{\circ} \to \Omega$, called **mean parameterization**.

- The *gradient of log-likelihood function* (***score functions***) for exponential family is

$$\nabla_{\boldsymbol{\eta}} \log p(\boldsymbol{x}; \boldsymbol{\eta}) = \boldsymbol{\phi}(\boldsymbol{x}) - \nabla_{\boldsymbol{\eta}} A(\boldsymbol{\eta}) = \boldsymbol{\phi}(\boldsymbol{x}) - \mathbb{E}_p\left[\boldsymbol{\phi}(\boldsymbol{X})\right] \tag{11}$$

- The ***Fisher information*** for exponential family is

$$\begin{aligned}
[\boldsymbol{I}(\boldsymbol{\eta})]_{i,j} &:= \mathbb{E}_{\boldsymbol{\eta}}\left[\left(\frac{\partial}{\partial \eta_i} \log p(\boldsymbol{x}; \boldsymbol{\eta})\right)\left(\frac{\partial}{\partial \eta_j} \log p(\boldsymbol{x}; \boldsymbol{\eta})\right)\right] \\
&= -\mathbb{E}_{\boldsymbol{\eta}}\left[\frac{\partial^2}{\partial \eta_i \partial \eta_j} \log p(\boldsymbol{x}; \boldsymbol{\eta})\right] && (12) \\
&= \frac{\partial^2 A(\boldsymbol{\eta})}{\partial \eta_i \partial \eta_j} = \mathbb{E}_{\boldsymbol{\eta}}\left[\phi_i(\boldsymbol{X}) \phi_j(\boldsymbol{X})\right] - \mathbb{E}_{\boldsymbol{\eta}}\left[\phi_i(\boldsymbol{X})\right] \mathbb{E}_{\boldsymbol{\eta}}\left[\phi_j(\boldsymbol{X})\right] && (13) \\
&:= \text{Cov}(\phi_i(\boldsymbol{X}), \phi_j(\boldsymbol{X}))
\end{aligned}$$

Note that $A(\boldsymbol{\eta})$ is convex, so the Fisher information matrix is positive definite $\boldsymbol{I}(\boldsymbol{\eta}) \succ \boldsymbol{0}$.

2 The information geometry of exponential families

2.1 Information geometry and its related concepts

The field of ***information geometry*** [Murray and Rice, 1993, Amari and Nagaoka, 2007] refers to an interdisciplinary field that applies the techniques of ***differential geometry*** [do Carmo Valero, 1976, 1992, Lee, 2003.] to study ***families of probability distributions and statistics***.

Statistical inference concerns situations in which one knows or suspects that data are generated by sampling from a space according to a probability distribution which is a member of some known family $p(\theta)$. The problem is to infer facts about the distribution from the data. For example, one might want to know the parameter value of the distribution (*point estimation*), or simply whether or not this value lies in some *particular set* of parameters (*hypothesis testing*).

There are several **motivations** to study *information geometry*:

- Many of hypothesis tests and much of the theory of statistical inference depends on **the choice of parameters**. It is important to know how the theory depends on the parameters, either because one suspects that **it should not depend on the parameters at all** or because one would like to know if a particular choice of parameters may **simplify** matters. Differential geometry provides tools to do "***calculus on manifolds***".
- In information geometry, we think of ***families of probability distributions*** as *entities* ***independent*** of any particular ***parametrization***, and able to support a variety of ***geometries***. Information geometry involves studies of ***invariants*** under certain transformation on distributions
- In information geometry, we relate the ***statistical properties*** to the ***geometries*** of underlying space of probablity distributions.

We listed concepts in information geometry as follows:

- From **differential geometry**:
  - ***manifold***, *sub-manifold*, *affine subspace*, *diffeomorphism*, *coordinate systems*, *differential form*, *differential operator*,
  - *tangent space*, *(tangent) vector field*, *change of coordinates*, *Christoffel symbols*, *Gauss map*, *The first and second fundamental form of a regular surface*,
  - *curvature*, ***Riemannian metric***, ***isometry***, *intrinsic geometry of a surface*,
  - *covariant derivative*, ***parallel transport***, *parallel translation*, ***geodesics***, *exponential map*, ***connection***, *affine connection*, ***Riemannian connection***, *dual connection*,
- From **statistics and information theory**:
  - *parameterization*, ***log-likelihood***, ***Fisher information metric***, *sufficient statistics*, ***exponential families***, *log-partition functions/cumulant functions*, *maximum likelihood estimator*, *maximum entropy estimator*, *Cramér-Rao inequality*, *asymptotics*
  - *entropy*, ***statistical divergences***, *KL divergence*, *Rényi divergence*, *f-divergence*, *Wasserstein distance*,

2.2 Affine spaces

- An affine space is nothing more than a *vector space* whose **origin** we try to forget about, by adding **translations** to the linear maps. In an affine space, there is ***no*** *distinguished point* that serves as an ***origin***. Hence, no vector has a fixed origin and no vector can be uniquely associated to a point.

- **Definition** An ***affine space*** is defined as a set X together with a ***vector space*** V and a *transitive* and *free action* of the ***additive group*** of V on X. Explicitly, we define the ***translation*** action $+ : X \times V \to X$, so that $(p, v) \mapsto p + v$. The translation have to satisfy the following rules

  1. *Right identity:* $a + 0 = a$ for any $a \in X$ and $0 \in V$ is the zero vector;

  2. *Associativity:* $(p + v) + w = p + (v + w)$ for any point $p \in X$ and vectors $v, w \in V$;

  3. *Subtraction:* given any two points $p, q \in X$ there must be a unique translation that moves one to the other, i.e. $q = p + v$ for some unique $v \in V$

- An affine space is a ***principal homogeneous space*** for the action of the additive group of a vector space. The element of X is called *points* and the vector from the associated vector space V is called *free vectors*. The operation $p + v$ is called ***translation by*** v from p.

- **Definition** An ***affine subspace*** of X is the subset of X of the form

$$p + W = \{p + w \,|\, w \in W\}$$

  where $p \in X$ and $W \subseteq V$ is a linear subspace of V associated with X.

- The linear subspace associated with an affine subspace is often called its ***direction***, and two subspaces that *share the same direction* are said to be ***parallel***. Every translation $A \to A : a \mapsto a + v$ maps any affine subspace to a parallel subspace.

2.3 Log-likelihood and affine structure

- **Definition** Denote $\mathcal{M}$ as the families of **non-negative measures** on $\mathcal{X}$ which are ***absolutely continuous*** *with respect to each other*. Let $\mathcal{R}_{\mathcal{X}}$ be the **vector space** of measurable functions f on $\mathcal{X}$. Define ***translation operation*** $+ : \mathcal{M} \times \mathcal{R}_{\mathcal{X}} \to \mathcal{M}$ as $d\mu + f = e^{f} d\mu$. This operation satisfies the *Right identity*, *Associativity*, *Subtraction* conditions above. That is, $\mathcal{M}$ forms an ***affine space*** and the vector space $\mathcal{R}_{\mathcal{X}}$ is associated with $\mathcal{M}$. The operation $+f$ is called the ***translation by*** f. Every $f \in \mathcal{R}_{\mathcal{X}}$ is a random variable on $\mathcal{X}$.

- **Definition** If $\mu \in \mathcal{M}$ is the base measure, we denote by $\ell : \mathcal{M} \to \mathcal{R}_{\mathcal{X}}$ the map

$$\ell(p d\mu) := \log(p). \tag{14}$$

  When p is a probability density with respect to μ, $\ell(p d\mu)$ is called ***log-likelihood***.

- Expressing *measures* as ***densities*** with respect to a *base measure*, and considering the **log-likelihoods** of these densities, amounts precisely to *choosing an origin* for $\mathcal{M}$ and identifying points of $\mathcal{M}$ with their *translation vectors* from the origin.

- Denote by $\mathcal{P}$ the space of all probability measures in $\mathcal{M}$. Notice that probability measures cannot form an affine space inside $\mathcal{M}$, since translation operation will likely destory the finite total mass condition. However, probability measures can also be regarded as non-negative measures ***up to scale***. Regarding a probability measure as a finite measure up to scale in effect treats it as an ***equivalence class of measures***, with two measures being considered equivalent if they are rescalings of each other.

- It follows that the measures in a measure class $\mathcal{M}$, when **identified up to scale**, form an **affine space** whose translation vectors are measurable functions f identified up to the addition of a constant. The space of all probability measures $\mathcal{P}$ is a *subset* of this affine space, namely the set corresponding to finite measures up to scale.

- For a ***finite dimension affine subspace*** $\mathcal{P} \subset \mathcal{M}$, it is spanned by a set of basis $\phi_1, \ldots, \phi_d \in \mathcal{R}_{\mathcal{X}}$ and if μ is one of the measures (up to scale), then the measures in such a family have the form

$$p = \mu + \sum_{\alpha=1}^{d} \eta_\alpha \phi_\alpha = \exp\left(\sum_\alpha \eta_\alpha \phi_\alpha - A\right) d\mu,$$

  where constant A controls the scale of the measure. Choosing A as the cumulant function $A(\boldsymbol{\eta})$ will make sure p is a proper probability measure.

- From above, we see that ***exponential families*** have a geometrical characterisation in terms of the ***finite-dimensional affine subspaces*** of **measures** *up to scale*, with their natural ***log-likelihood affine structure***. From section below, we see that the reverse is true.

- The *canonical parameters* $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_d)$ are ***affine coordinates*** for exponential families.

2.4 Geometrical criterion for exponential families

$$P = \{p(\boldsymbol{x}; \boldsymbol{\eta})\} \subset \mathcal{P}$$

is a parametrised family of probability measures and is considered as a surface in $\mathcal{P} \subset \mathcal{M}$. Assume that the log-likelihood function $\ell(\boldsymbol{\eta})$ is a differentiable function of $\boldsymbol{\eta}$

- The affine subspace $P \subset \mathcal{P}$ has the form

$$P = \{p + f = \exp(f)p : f \in V\} \tag{15}$$

  where $p \in P$ is some point and $V \subseteq \mathcal{R}_{\mathcal{X}}$ is a subspace of random variables . We shall call the vector space V the **tangent space** to P at p and denote it by T_pP.

- **Proposition 2.1** *The **tangent space** T_pP has a **basis** as the gradient of log-likelihood functions*

$$\nabla_{\boldsymbol{\eta}} \ell(p) = \left(\frac{\partial \ell}{\partial \eta_1}(p), \ldots, \frac{\partial \ell}{\partial \eta_d}(p)\right)$$

  *This basis $\nabla_{\boldsymbol{\eta}} \ell(p)$ is referred to as **score vectors**. We denote $\ell_i(\boldsymbol{\eta}) := \frac{\partial \ell}{\partial \eta_i}$.*

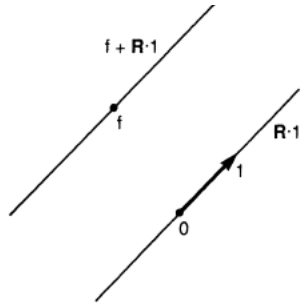

Figure 1.1

Figure 1: The quotient space $\mathcal{R}_{\mathcal{X}}/\mathcal{R}.1$ contains all lines in $\mathcal{R}_{\mathcal{X}}$ in parallel to $\mathcal{R}.1$. [Murray and Rice, 1993]

- Note that score vectors have **zero mean**:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\eta}}\left[\nabla_{\boldsymbol{\eta}}\ell(p)\right] &= \left[\int p(\boldsymbol{x};\boldsymbol{\eta})\frac{\partial \log p(\boldsymbol{x};\boldsymbol{\eta})}{\partial \eta_i}d\boldsymbol{x}\right]_i \\ &= \left[\int p(\boldsymbol{x};\boldsymbol{\eta})\frac{1}{p(\boldsymbol{x};\boldsymbol{\eta})}\frac{\partial p(\boldsymbol{x};\boldsymbol{\eta})}{\partial \eta_i}d\boldsymbol{x}\right]_i \\ &= \left[\int \frac{\partial p(\boldsymbol{x};\boldsymbol{\eta})}{\partial \eta_i}d\boldsymbol{x}\right]_i \\ &= \mathbf{0} \end{aligned} \tag{16}$$

  The last equality holds because $\int p(\boldsymbol{x};\boldsymbol{\eta})d\boldsymbol{x} = 1 \Rightarrow \partial_i \int p(\boldsymbol{x};\boldsymbol{\eta})d\boldsymbol{x} = \int \partial_i p(\boldsymbol{x};\boldsymbol{\eta})d\boldsymbol{x} = 0$.

- In order to deal with probability measures as positive measures ***up to scale*** it will be convenient to *extend* any such family to a family $\widetilde{P}$ of **positive measures** defined by

$$\widetilde{P} = \{\exp(\lambda)p \,|\, \lambda \in V,\, p \in P\}$$

  We extend the local co-ordinates η_i by defining $\eta_i(\exp(\lambda)p) := \eta_i(p)$ and define a **new co-ordinate** by $\eta_0(\exp(\lambda)p) = \lambda$, i.e. the bias term.

- Let us denote by $\mathcal{R}.1$ the one-dimensional vector subspace of *constant* random variables. The 1 in this notation is the random variable which is everywhere equal to 1 and the $\mathcal{R}$ denotes that we want to consider all scalar multiples of 1, that is, the line in $\mathcal{R}_{\mathcal{X}}$ containing 1. The space of random variables ***up to addition of constants*** forms a ***quotient space*** $\mathcal{R}_{\mathcal{X}}/\mathcal{R}.1$. $\mathcal{R}_{\mathcal{X}}/\mathcal{R}.1$ contains all lines in $\mathcal{R}_{\mathcal{X}}$ in parallel to $\mathcal{R}.1$. The line through the random variable f is denoted $f + \mathcal{R}.1$. See Figure 1.

- **Proposition 2.2** *(**Geometrical criterion for exponential families**) [Murray and Rice, 1993]*
  P is an affine subspace of $\mathcal{P}$ generated by a vector space $V \subset \mathcal{R}_{\mathcal{X}}/\mathcal{R}.1$ if and only if $\widetilde{P}$ is an affine subspace of $\mathcal{M}$ generated by

$$\widetilde{V} = \{f \in \mathcal{R}_{\mathcal{X}} : f + \mathcal{R}.1 \in V\}$$

  *Hence P is an **exponential family** if and only if $\widetilde{P}$ is an **affine subspace** of $\mathcal{M}$.*

- Since the ***derivatives of the scores lie in the span of the scores*** at each point is characteristic of affine subspaces, it follows that P is an exponential family, if and only if

$$\widetilde{\ell}_{i,j}(\boldsymbol{\eta}) = \sum_{k=0}^{d} \Gamma_{i,j}^{k}(\boldsymbol{\eta})\widetilde{\ell}_k(\boldsymbol{\eta})$$

where $\widetilde{\ell}(\exp(\lambda)p) = \widetilde{\ell}(\boldsymbol{\eta}) = \ell(\boldsymbol{\eta}) + \lambda$ is the log-likelihood defined via $\widetilde{V}$. Also $\widetilde{\ell}_0(\boldsymbol{\eta}) = 1$ and $\widetilde{\ell}_{0,j}(\boldsymbol{\eta}) = 0$. The linear cooefficients $\Gamma_{i,j}^{k}(\boldsymbol{\eta})$ are called the ***Christoffel symbol***.

2.5 Computational criterion for exponential families

- To show that $\widetilde{P}$ is an affine subspace in $\mathcal{M}$, it suffices to know if the $\ell_{i,j}$ are in the tangent space to P for each $i, j = 1, \ldots, d$.
- **Definition** We can define an ***inner product*** in *tangent space* T_pP via expectation operation

$$\langle f \, , \, g \rangle_p = \mathbb{E}_p \left[f \, g \right] \tag{17}$$

on the subspace of p square-integrable random variables f, i.e. those random variables satisfying $\mathbb{E}_p \left[f^2 \right] < \infty$.

In general it suffices to assume that the ***scores*** $\widetilde{\ell}_k(p)$ at p and the *second derivatives* $\widetilde{\ell}_{i,j}(p)$ are p square-integrable.

- **Definition** The ***Fisher information matrix*** is defined as

$$g_{i,j}(p) := \langle \ell_i \, , \, \ell_j \rangle_p = \mathbb{E}_p \left[\ell_i \, \ell_j \right], \quad \text{for } i, j = 1, \ldots, d. \tag{18}$$

The Fisher information matrix is just the matrix of the ***inner product*** *with respect to the* ***basis*** in T_pP defined by the scores. It is also called the ***first fundamental form*** of regular surface in differential geometry.

- This inner product on $\mathcal{R}_{\mathcal{X}}$ defines a ***normal space*** N_p to $T_p\widetilde{P}$ such that

$$\mathcal{R}_{\mathcal{X}} = N_p \oplus T_p\widetilde{P}. \tag{19}$$

if $f \in \mathcal{R}_{\mathcal{X}}$ is a random variable its ***normal component*** in N_p is

$$\Pi_p(f) = f - \sum_{m,n} g^{m,n} \mathbb{E}_p \left[f \, \ell_m \right] \, \ell_n - \mathbb{E}_p \left[f \right] \tag{20}$$

where $g^{m,n}$ is the ***inverse*** of the Fisher information matrix.

Note

$$\begin{aligned} \langle \Pi_p(f) \, , \, 1 \rangle_p &= \mathbb{E}_p \left[f \right] - \sum_{i,j} g^{i,j} \mathbb{E}_p \left[f \, \ell_i \right] \, \mathbb{E}_p \left[\ell_j \right] - \mathbb{E}_p \left[f \right] = 0 \\ \langle \Pi_p(f) \, , \, \ell_k \rangle_p &= \mathbb{E}_p \left[f \, \ell_k \right] - \sum_{i,j} g^{i,j} \mathbb{E}_p \left[f \, \ell_i \right] \, \mathbb{E}_p \left[\ell_j \ell_k \right] = \mathbb{E}_p \left[f \, \ell_k \right] - \sum_{i,j} g^{i,j} g_{j,k} \mathbb{E}_p \left[f \, \ell_i \right] = 0 \end{aligned}$$

where $\mathbb{E}_p \left[\ell_j \right] = 0$ for all j.

We can rewrite f as

$$f = (f - \Pi_p(f)) + \Pi_p(f)$$

and $(f - \Pi_p(f))$ is a *linear combinations of the scores* and the constant random variables so in $T_p\widetilde{P}$.

- **Proposition 2.3** ***(Computational criterion for exponential families)*** *[Murray and Rice, 1993]*
  The family is ***exponential*** *if and only if the functions $\ell_{i,j}$ are always tangential to $\widetilde{P}$. This is equivalent to the* ***normal component of each $\ell_{i,j}$ vanishing****, that is, to*

$$\alpha_{i,j}(p) = \Pi_p(\ell_{i,j}) = \ell_{i,j} - \sum_{m,n} g^{m,n}\mathbb{E}_p\left[\ell_{i,j}\,\ell_m\right]\,\ell_n - \mathbb{E}_p\left[\ell_{i,j}\right] = 0 \tag{21}$$

  The quantity $\alpha_{i,j}$ is called the ***second fundamental form*** *of the family. It is also called the* ***imbedding curvature*** *in [Amari and Nagaoka, 2007].*

  This proposition states that the **second fundamental form vanishing** characterises affine subspaces.

- The *geometric interpretation* for above proposition is the following. $\widetilde{\ell}_k$ is tangent to $\widetilde{P}$ and $\widetilde{\ell}_{i,j}$ measures the rate of change of $\widetilde{\ell}_k(p)$ as the point p moves around $\widetilde{\mathcal{P}}$. These changes are to be regarded as due to *two causes*.

  1. The first cause is that the ***lines*** in $\widetilde{\mathcal{P}}$ on which the parameters are *constant* may be ***bending around***;
  2. The second cause is that the ***surface*** $\widetilde{\mathcal{P}}$ itself may be ***bending around*** inside $\mathcal{M}$.

  The tangential and normal components of $\widetilde{\ell}_{i,j}$ measure these two types of bending respectively. So the vanishing of the normal component of $\widetilde{\ell}_{i,j}$ corresponds to the fact that **the surface $\widetilde{P}$ is not bending**.

- **Definition** We can define a scalar quanitity γ from $\alpha_{i,j}$ so that $\gamma = 0$ if and only if $\alpha_{i,j} = 0$ for all i, j.

$$\gamma(p) = \sum_{i,j,k,l} g^{i,j}(p)g^{k,l}(p)\mathbb{E}_p\left[\alpha_{i,k}(p)\,\alpha_{j,l}(p)\right] \tag{22}$$

  where $g^{i,j}$ is the ***inverse*** of the Fisher information matrix. The function γ, in the case of a *one-dimensional family* is Efron's ***statistical curvature***.

  Note that since $g^{i,j}$ is positive definite, $\gamma = 0$ if and only if $\alpha_{i,j} = 0$ for all i, j.

- **Proposition 2.4** *The family P is* ***exponential*** *if and only its* ***statistical curvature*** *$\gamma(p) = 0$ for all $p \in P$.*

  This is equivalent to say that the exponential families are *flat*.

2.6 Parameter independence

In this section, we discuss how $\alpha_{i,j}$ depends on the choice of coordinates. We have **four equivalent criteria** to decide if a family of probability distribution is exponential.

1. The subset of positive measures $\widetilde{\mathcal{P}}$ is an ***affine subspace*** in $\mathcal{M}$;
2. The *second order derivatives* of log-likelihood $\partial_i \partial_j \ell$ are in the span of the ***scores*** $\{\partial_i \ell\}$ and the constants;
3. The ***second fundamental form*** $\alpha_{i,j}$ are vanishing;
4. The ***statistical curvature*** γ is vanishing.

Under reparameterization from $\boldsymbol{\eta}$ to $\boldsymbol{\theta}$, we compute these quantities.

- $$\frac{\partial^2 \ell}{\partial \eta_i \partial \eta_j} = \Gamma^k_{i,j} \frac{\partial \ell}{\partial \eta_k} + \Gamma^0_{i,j} \tag{23}$$

  Here we use the '***Einstein summation convention***' that any index which occurs *both raised and lowered* is *summed* over.

  If we change coordinates, by the chain rule,

  $$\begin{aligned} \frac{\partial \ell}{\partial \theta_k} &= \frac{\partial \eta_i}{\partial \theta_k} \frac{\partial \ell}{\partial \eta_i} \\ \frac{\partial \ell}{\partial \eta_i} &= \frac{\partial \theta_k}{\partial \eta_i} \frac{\partial \ell}{\partial \theta_k} \end{aligned} \tag{24}$$

  and

  $$\frac{\partial^2 \ell}{\partial \theta_k \partial \theta_l} = \frac{\partial^2 \eta_i}{\partial \theta_k \partial \theta_l} \frac{\partial \ell}{\partial \eta_i} + \frac{\partial^2 \ell}{\partial \eta_i \partial \eta_j} \frac{\partial \eta_i}{\partial \theta_k} \frac{\partial \eta_j}{\partial \theta_l} \tag{25}$$

  Substituting (23) and (24) into (25), we have

  $$\begin{aligned} \frac{\partial^2 \ell}{\partial \theta_k \partial \theta_l} &= \frac{\partial^2 \eta_i}{\partial \theta_k \partial \theta_l} \frac{\partial \ell}{\partial \eta_i} + \frac{\partial^2 \ell}{\partial \eta_i \partial \eta_j} \frac{\partial \eta_i}{\partial \theta_k} \frac{\partial \eta_j}{\partial \theta_l} \\ &= \frac{\partial^2 \eta_i}{\partial \theta_k \partial \theta_l} \frac{\partial \ell}{\partial \eta_i} + \Gamma^q_{i,j} \frac{\partial \ell}{\partial \eta_q} \frac{\partial \eta_i}{\partial \theta_k} \frac{\partial \eta_j}{\partial \theta_l} + \Gamma^0_{i,j} \frac{\partial \eta_i}{\partial \theta_k} \frac{\partial \eta_j}{\partial \theta_l} \\ &= \left(\frac{\partial^2 \eta_i}{\partial \theta_k \partial \theta_l} + \Gamma^i_{m,n} \frac{\partial \eta_m}{\partial \theta_k} \frac{\partial \eta_n}{\partial \theta_l} \right) \frac{\partial \ell}{\partial \eta_i} + \Gamma^0_{i,j} \frac{\partial \eta_i}{\partial \theta_k} \frac{\partial \eta_j}{\partial \theta_l} \\ &= \left(\frac{\partial^2 \eta_i}{\partial \theta_k \partial \theta_l} \frac{\partial \theta_r}{\partial \eta_i} + \Gamma^i_{m,n} \frac{\partial \theta_r}{\partial \eta_i} \frac{\partial \eta_m}{\partial \theta_k} \frac{\partial \eta_n}{\partial \theta_l} \right) \frac{\partial \ell}{\partial \theta_r} + \Gamma^0_{i,j} \frac{\partial \eta_i}{\partial \theta_k} \frac{\partial \eta_j}{\partial \theta_l} \end{aligned}$$

  Therefore if $\partial_i \partial_j \ell(\boldsymbol{\eta})$ are in the span of the scores $\{\partial_i \ell(\boldsymbol{\eta})\}$, so also is $\partial_i \partial_j \ell(\boldsymbol{\theta})$.

- The *second fundamental form* $\alpha_{k,l}$ under $\boldsymbol{\theta}$ can be computed by considering it as the projection of $\frac{\partial^2 \ell}{\partial \theta_k \partial \theta_l}$ to the orthogonal space of $T_p \widetilde{P} = \operatorname{span}\{\partial_i \ell(\boldsymbol{\eta})\}$.

  $$\begin{aligned} \alpha_{k,l} &= \Pi \left(\frac{\partial^2 \ell}{\partial \theta_k \partial \theta_l} \right) \\ &= \Pi \left(\frac{\partial^2 \ell}{\partial \eta_i \partial \eta_j} \frac{\partial \eta_i}{\partial \theta_k} \frac{\partial \eta_j}{\partial \theta_l} \right) = \Pi \left(\frac{\partial^2 \ell}{\partial \eta_i \partial \eta_j} \right) \frac{\partial \eta_i}{\partial \theta_k} \frac{\partial \eta_j}{\partial \theta_l} \\ &= \alpha_{i,j} \frac{\partial \eta_i}{\partial \theta_k} \frac{\partial \eta_j}{\partial \theta_l} \end{aligned} \tag{26}$$

The second equation holds since $\Pi\left(\frac{\partial \ell}{\partial \eta_i}\right) = 0$. It follows that $\alpha_{k,l}$ vanishes precisely when $\alpha_{i,j}$ vanishes.

- Note that the ***Fisher information matrix*** under ***reparameterization*** $\boldsymbol{\theta}$ can be computed as

$$\begin{aligned} g_{k,l}(\boldsymbol{\theta}) &= \mathbb{E}_p\left[\frac{\partial^2 \ell}{\partial \theta_k \partial \theta_l}\right] \\ &= \frac{\partial^2 \eta_i}{\partial \theta_k \partial \theta_l}\mathbb{E}_p\left[\frac{\partial \ell}{\partial \eta_i}\right] + \mathbb{E}_p\left[\frac{\partial^2 \ell}{\partial \eta_i \partial \eta_j}\right]\frac{\partial \eta_i}{\partial \theta_k}\frac{\partial \eta_j}{\partial \theta_l} \\ &= g_{i,j}(\boldsymbol{\eta})\ \frac{\partial \eta_i}{\partial \theta_k}\frac{\partial \eta_j}{\partial \theta_l} \end{aligned} \tag{27}$$

  since $\mathbb{E}_p\left[\frac{\partial \ell}{\partial \eta_i}\right] = 0$ for all i.

  From (13), we see that the ***intrinsic properties*** of curvature are ***unchanged*** under different parametrizations. In general, the Fisher information matrix provides a ***Riemannian metric*** (more precisely, the Fisher-Rao metric) for the manifold of thermodynamic states.

- Substituting results in (26) and (27) into (22), we see that the Jacobians are all cancelled out. The following proposition readily follows:

  Proposition 2.5 *The statistical curvature γ is an **invariant** of the family of distributions P. That is, it is a function on P that is **independent** of the choice of coordinates.*

- From the proposition above, we see that both the geometric criterion (*affine subspace*) and the computational criterion (*vanishing statistical curvature*) for *exponential families* are **independent** of choice of coordinates.

References

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2007.

Manfredo Perdigao do Carmo Valero. *Differential geometry of curves and surfaces*, volume 2. Prentice-hall Englewood Cliffs, 1976.

Manfredo Perdigao do Carmo Valero. *Riemannian geometry*. Birkhäuser, 1992.

John Marshall Lee. *Introduction to smooth manifolds*. Graduate texts in mathematics. Springer, New York, Berlin, Heidelberg, 2003. ISBN 0-387-95448-1.

Michael K Murray and John W Rice. *Differential geometry and statistics*, volume 48. CRC Press, 1993.