Hittite Journal of Science and Engineering, 2018, 5 (4) 249-257
ISSN NUMBER: 2148-4171
DOI: 10.17350/HJSE19030000089

On Information Geometrical Structures

Fatma Muazzez Şimşir
Hitit University, Department of Mathematics, Çorum, TURKEY

Article History:
Received: 2017/10/06
Accepted: 2017/11/18
Online: 2018//12/31

***Correspondence to**: Fatma Muazzez Şimşir Hitit University, Department of Mathematics, Ulukavak Mahallesi, 19040, Çorum, Turkey*
E-mail: fmuazzezsimsir@hitit.edu.tr
Phone: +90 364 2277000-1663

ABSTRACT

Information geometry is a modern differential geometric approach to statistics, in particular theory of information. The main motivation for this expository survey article is the lack of compact material that mainly address to mathematical audience because of the interdisciplinary content. Information geometry simply described as applying the techniques of differential geometry to statistical models, represented as manifolds of probability distributions. This can be done either done by putting the concept of divergences on the center or the Fisher metric. This paper is motivated from the latter approach.

Keywords:
Information geometry; Dual affine connections; Statistical distributions

INTRODUCTION

Probability distributions are the basic tools of statistics and statistical inference. One of the main problems of mathematical statistics is finding a measure to distinguish one probability distribution from another. Moreover, in statistical inference a probability distribution is chosen from a set of candidates. This immediately brings up the question of what would happen if a neighbor distribution is selected. One way to answer such questions is to introduce a notion of ”distance” between probability distributions.

Information theory originated in 1940’s by Shannon, [25]. The earliest ideas of combining statistics and differential geometry goes back 1945’s to Rao and Jeffreys [12, 23] who used independently Fisher information as a Riemannian metric. However, it was by Efron that the role of differential geometry started to play an important role in statistics. He defined the statistical curvature for one-parameter statistical models in 1975, [11]. However, in Efron’s work tools of differential geometry was not used elaborately. The first step to use elegant differential geometry in the context of statistics was by Dawid, [10]. By using the notion of statistical curvature, he defined e-connection (exponential connection) on the space of positive probability distributions. Moreover, he also showed that on this space different connections may be defined. The (0)-connection (Levi Civita) and (m)-connection (mixture connection) of the Fisher metric were the examples. Since the space of positive probability distributions is infinite dimensional, it is not easy to see this space as a manifold. Amari [2] in 1980’s used tools of modern differential geometry and developed a systematic method to investigate informational theoretical concepts by taking projection of Efron’s model into finite dimensional models. He and Nagaoka considered (e)-connection and (m)-connection as a pair of dual connections which will later be on the center of information geometry, [4]. Actually, Chentsov [7] had already been defined (α)-connections from a different viewpoint, however, the article was in Russian and their relationship with statistical estimation was omit- ted in the article. Hence, his contributions are not well-known among statisticians. The standard references to get familiar with information geometry are [1, 2, 3, 4, 8, 14, 19, and 26].

Throughout this paper the close relationship between statistical models and differential geometry, in particular affine differential geometry is emphasized. Most of the times, different schools of geometry prefer to use different terminology for the same concepts which cause confusion for those that are not much familiar with the field. Therefore, such cases are highlighted and nuances is also tried to be explained, as well.

STATISTICAL MODELS

The probability distributions on a set will be represented as follows: If χ is a discrete set (with finite or countably infinite cardinality), then a probability distribution on χ is a function $P:\chi \to \mathbb{R}$ which satisfies.

$$p(x) \ge 0, \quad \forall x \in \chi \quad \text{and} \quad \sum_{x\in\chi} p(x) = 1 \qquad (2.1)$$

If $\chi = \mathbb{R}^n$ then it is a function $P:\chi \to \mathbb{R}$ which satisfies

$$p(x) \ge 0, \quad \forall x \in \chi \quad \text{and} \quad \int p(x)dx = 1. \qquad (2.2)$$

Consider a family S of probability distributions on χ Suppose that each element of S , a probability distribution, may be parametrized using an n real-valued variables $\left[\xi^1,\ldots,\xi^n\right]$ so that

$$S = \{ \ p_\xi = p(x;\xi) \mid \xi = \left[\xi^1,\ldots,\xi^n\right] \in \Xi \ \}$$

where Ξ is a subset of $\mathbb{R}^n$ and the mapping $\xi \to p_\xi$ is injective. Such S is called an *n*-dimensional statistical model, a parametric model, or simply a model on χ . Assumptions that we made on statistical models are:

- We may freely differentiate with respect to the parameters. Assume that Ξ is an open subset of $\mathbb{R}^n$ and $\forall x \in \Xi$ the function $\xi \to \mathbb{R}$ is C^∞ .

- The order of integration and the differentiation may freely be rearranged. For instance,

$$\int \partial_i p(x;\xi)dx = \partial_i \int p(x;\xi)dx = \partial_i 1 = 0.$$

where $\partial_i = \dfrac{\partial}{\partial \xi^i}$.

- The model S is a subset of

$$P(\chi) = \left\{ p:\chi \to \mathbb{R} \mid p(x) > 0, \forall x \in \chi and \int p(x)dx = 1 \right\}.$$

Some Examples of Statistical Models

- Normal Distribution

$$\chi = \mathbb{R},\ n = 2,\ \xi = [\mu,\sigma],\ \Xi = \left\{[\mu,\sigma] \mid \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+\right\}$$

$$p(x,\xi) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}$$

- Multivariate Normal Distribution

$$\chi = \mathbb{R}^k,\ n = k + \frac{k(k+1)}{2},\ \xi = [\mu,\Sigma]$$

$$\Xi = \left\{[\mu,\Sigma] \mid \mu \in \mathbb{R}^k, \Sigma \in \mathbb{R}^{k\times k} : \textit{positive definite}\right\}$$

$$p(x;\xi) = (2\pi)^{-n/2} (\det\Sigma)^{-1/2} \exp\left\{ -(x-\mu)^t \Sigma^{-1} (x-\mu) \right\}$$

- Poisson Distribution

$$\chi = \{0,1,2,\ldots\},\ n = 1,\ \Xi = \{\xi \mid \xi > 0\}$$

$$p(x;\xi) = e^{-\xi} \frac{\xi^x}{x!}$$

$$P(X) \ \ \textit{for} \ \ \textit{finite} \ \ \chi$$

$$\chi = \{x_0, x_1, \ldots, x_n\}$$

$$\Xi = \left\{ \left[\xi^1, \xi^2, \ldots, \xi^n\right] \mid \xi^i > 0 \forall i, \sum_{i=1}^n \xi^i < 1 \right\}$$

$$p(x;\xi) = \begin{cases} \xi^i & 1 \le i \le n \\ 1 - \Sigma_{i=1}^n \xi^i & i = 0 \end{cases}$$

KAHLER AFFINE MANIFOLDS, STATISTICAL MANIFOLDS AND DUALLY FLAT STRUCTURES

An affine manifold is a differential manifold whose coordinate changes are affine transformations which immediately give rise to existence of a torsion-free connection with vanishing curvature. Note that affine transformations are made up of a linear transformation followed by a translation. Since an affine manifold is a differentiable manifold with affine charts one may define a two tensor $g_{ij} = \dfrac{\partial^2 F}{\partial x^i \partial x^j} dx^i \otimes dx^j$ where φ is a strictly convex function. Thus, *g* is symmetric and positive definite. Hence, it is a Riemannian metric on *M* which will be called a Kahler affine metric. Note that the coefficients of the metric tensor *g* is invariant under affine transformations, [15, 16]. Such structures are first introduced by Cheng and Yau, [6]. Kahler affine metrics are called Hessian metrics by Japanese school due to the fact that g_{ij} is the Hessian of a convex local potential *F*, [26].

An affine manifold equipped with a Kahler affine metric is called a Kahler affine manifold. One may recover dually flat connections from this structure. Conversely, given mutually flat connections one may obtain local potential functions. The flat affine connection D and its dual D^* are called dually flat connections with respect to the Kahler affine metric *g*. In other words, for all vector fields *X,Y* on *M*, $Xg(Y,Z) = g(D_X Y, Z) + g(Y, D_X^* Z)$. On the other hand, a statistical manifold is simply a Riemannian manifold (*M,g*) together with two torsion free connections ∇ and ∇^* that satisfies a duality relation with respect to the Riemannian metric . Two torsion free connections ∇ and ∇^* are called dual to each other with respect to a Riemannian metric *g* if for all vector fields X,Y on M, $Xg(Y,Z)=g(\nabla_X Y,Z)+g(Y,\nabla^*_X Z)$. If $\nabla=\nabla^*$ the geometry reduces to the Riemannian one. One may refer to the works of Lauritzen, Kurose and Noguchi [19, 17, 22] for a detailed study of statistical manifolds. There is a close relationship between the statistical manifolds and Kahler affine manifolds. It can be seen from definitions that every Kahler affine manifold is a statistical manifold. However, not all statistical manifolds are Kahler affine. Consider $\mathbb{R}^n$ with is standard affine coordinate system $\{x_1,\dots,x_n\}$ and let D be the canonical flat affine connection, i.e., $D_{\frac{\partial}{\partial x^i}}\frac{\partial}{\partial x^j}=0.$ Let $\Omega\subset\mathbb{R}^n$ be a domain and let φ be a strictly convex function on Ω. .With the Kahler affine metric $g=\frac{\partial^2\varphi}{\partial x^i\partial x^j}dx^i dx^j$, the triple (Ω,D,g) is a Kahler affine manifold. This triple is a flat statistical manifold. Conversely a flat statistical manifold is locally isometric to the Kahler affine manifold (Ω,D,g).

The α -Connection

Let M be a Kahler affine manifold for $-1\le\alpha\le 1$ the α -Connection is defined by

$$\Gamma^{(\alpha)}_{ijk}=\Gamma^{(0)}_{ijk}-\frac{\alpha}{2}\partial_i\partial_j\partial_k\varphi \tag{3.1}$$

where the Levi-Civita connection of g is denoted by $\Gamma^{(0)}_{ijk}$ and

$$\Gamma^{(0)}_{ijk}\left\langle \nabla^{(0)}_{\frac{\partial}{\partial x^i}}\frac{\partial}{\partial x^j},\frac{\partial}{\partial x^k}\right\rangle. \tag{3.2}$$

From (3.1) and (3.2)

$$\Gamma^{(0)}_{ijk}=\frac{1}{2}\partial_i\partial_j\partial_k\varphi, \tag{3.3}$$

and

$$\Gamma^{(\alpha)}_{ijk}=\frac{1}{2}(1-\alpha)\partial_i\partial_j\partial_k\varphi. \tag{3.4}$$

Since 3.4 is symmetric with respect to i and j, $\nabla^{(\alpha)}$ is torsion free. Moreover,

$\Gamma^{(\alpha)}_{ijk}+\Gamma^{(-\alpha)}_{ijk}=2\Gamma^{(0)}_{ijk},\nabla^{(\alpha)}$ and $\nabla^{(-\alpha)}$) is dual to each othei with respect to g. In other words, for all vector fields X,Y,Z

$$Zg(X,Y)=g(\nabla^{(\alpha)}_Z X,Y)+g(X,\nabla^{(-\alpha)}_Z Y). \tag{3.5}$$

Since $\Gamma^{(1)}_{ijk}=0,\nabla^{(1)}$ defines a flat structure and x-coordinates are an affine coordinate system for $\nabla^{(1)}$. Therefore, the connection $\nabla^{(-1)}$ is dual to the connection $\nabla^{(1)}$ and its Christoffel symbols in x-coordinates is of the form

$$\Gamma^{(-1)}_{ijk}=\partial_i\partial_j\partial_k\varphi.$$

The dual affine coordinate θ is

$$\theta_j=\partial_i\varphi, \tag{3.6}$$

and so also

$$g_{ij}=\partial_i\theta_j. \tag{3.7}$$

The corresponding local potential function can be calculated by a Legendre transform

$$\phi(\theta)=\max_x(x^i\theta_i-\varphi(x)),\quad \varphi(x)+\phi(\theta)-x.\xi=0, \tag{3.8}$$

and

$$x^j=\partial^j\phi(\theta),\quad g^{ij}=\frac{\partial x^j}{\partial\theta_i}=\partial^i\partial^j\phi(\theta). \tag{3.9}$$

Therefore, a Kahler affine metric together with the flat affine connection yields a dual structure. Such structures constitutes the foundations of the information geometry. Conversely, from such a dually flat structure, the local potential functions or the Kahler affine structure can be obtained. Let $D=\nabla^{(1)}$ and $D^*=\nabla^{(-1)}$ be dually flat connections and let $\{x^1,\dots,x^n\}$ be the affine coordinates that is obtained from the flat connection D.

Hence, the vector fields $\partial_i=\frac{\partial}{\partial x^i}$ are parallel. We define ∂^j

$$g(\partial_i,\partial^j)=\delta_i^j=\begin{cases}1 & i=j\\ 0 & i\neq j\end{cases}$$

and for every vector field X

$$Xg(\partial_{i,}\partial^j)=g(D_X\partial_i,\partial^j)+g(\partial_i,D^*_X\partial^j).$$

∂_i is parallel for D so is ∂^j for D^*. Since D^* is torsion free, $\left[\partial^j,\partial^k\right]=0$ for all j,k. Hence, the affine coordinates θ_j where $\partial^j=\frac{\partial}{\partial\theta_j}$ is obtained.

Note that, when passing from x-coordinates to θ -coordinates ∂_i transforms contravariantly, whereas ∂^j transforms covariantly. The transformation rule between x and θ coordinates is given by $\partial^j=(\partial^j x^i)\partial_i$ and $\partial_i=(\partial_i\theta_j)\partial^j$. In the following sequel the metric tensor g in x and θ -coordinates will be calculated. Since $g_{ij}:=g(\partial_i,\partial_j)$, $g^{ij}=g(\partial^i,\partial^j)$ and $g(\partial_i,\partial^j)=\delta_i^j$

$$g_{ij}=\frac{\partial\theta_j}{\partial x^i}$$

$$g^{ij}=\frac{\partial x^i}{\partial\theta_j}.$$

We would like to find the local potential functions $\varphi(x)$ and $\phi(\theta)$ that satisfy $\theta_i = \partial_i \varphi(x)$, $x^i = \partial^i \phi(\theta)$. The first equation can be solved locally if $\partial_i \theta_j = \partial_j \theta_i$ using the fact that the metric tensor g is symmetric. Hence, $g_{ij} = \partial_i \partial_j \varphi$ and φ strictly convex function. $\phi := x^i \theta_i - \varphi$ is defined from the duality to get

$\partial^i \phi = x^i + \frac{\partial x^j}{\partial \theta_i} \theta_j - \frac{\partial x^j}{\partial \theta_i} \frac{\partial}{\partial x^j} \varphi = x^i$. Since φ and ϕ are strictly convex functions, they relate to each other by a Legendre transform

$$\phi(\theta) = \max_x \left(x^i - \theta_i - \varphi(x) \right) \tag{3.10}$$

$$\varphi(x) = \max_\theta \left(x^i - \xi_i - \phi(\theta) \right). \tag{3.11}$$

Moreover, Christoffel symbols of the metric g^{ij} in x-coordinates are

$$\Gamma'^{ijk} = -\Gamma_{ijk} = -\frac{1}{2} \partial_i \partial_j \partial_k \varphi, \tag{3.12}$$

$$\Gamma'^{(\alpha)ijk} = \Gamma'^{ijk} - \frac{\alpha}{2} \partial_i \partial_j \partial_k \varphi = -\Gamma_{ijk}^{(-\alpha)}. \tag{3.13}$$

Consequently, $\Gamma'^{(1)} = -\Gamma^{(-1)}$. Hence, $-\Gamma^{(-1)} = 0$ in θ -coordinates.

GEOMETRY OF STATISTICAL MODELS

In particular, why differential geometry is useful for statistics? A statistical model is a set of probability distributions to which we believe that the true distribution belongs. It is a subset of all possible probability distributions. One often uses a statistical model to carry out statistical inference, assuming that the true distribution is in the model. However, the true distribution may not be in the model but only close to it. Therefore, in order to evaluate statistical inference procedures, it is important to know what part the statistical model occupies in the entire set of probability distributions and what shape the statistical model has in the entire set of models. Hence, it is expected that a fundamental role is played in statistics by the geometric quantities such as the distance or divergence of two probability distributions, the flatness or the curvature of the statistical model. Statistical inference may be carried out more and more precisely as the number of observations increases so that one could construct a universal asymptotic theory of the statistical inference in the regular case. Since the estimated probability distribution lies very close to the true distribution in this case, it is sufficient when evaluating statistical procedures to take account of only local structure of the model in a small neighborhood of the true or estimated distribution. Thus, one can locally linearize the model at the true or estimated distribution even if the model is curved in the entire set.

Let $S = \{ p_\xi \mid \xi \in \Xi \}$ be an n-dimensional statistical model. Given a point ξ, the The Fisher Information Matrix of at ξ is an nxn matrix $\left[g_{ij}(\xi) \right]$ is defined by

$$g_{ij}(\xi) = E_\xi \left[\partial_i l_\xi \partial_j l_\xi \right] = \int \partial_i l_\xi \partial_j l_\xi p_\xi dx \tag{4.1}$$

where $l_\xi = l(x;\xi) = \log p(x;\xi)$ and is called log likelihood in statistics. Although there are some models in which the above integral diverges, we assume that g_{ij} is finite for all *i,j* and that $g_{ij} : \Xi \to \mathbb{R}$ is C^∞. Note that one can write g_{ij} as:

$$g_{ij}(\xi) = -E_\xi \left[\partial_i \partial_j l_\xi \right]. \tag{4.2}$$

Another important representation is

$$g_{ij}(\xi) = 4 \int \sqrt{p(x;\xi)} \sqrt{p(x;\xi)} dx. \tag{4.3}$$

In finite case, it becomes

$$\Sigma_{k=1}^n \frac{1}{p_k} \frac{\partial p_k}{\partial \xi^i} \frac{\partial p_k}{\partial \xi^j}. \tag{4.4}$$

which is the Shashahani metric in mathematical biology, [13, 14]. This is simply the metric obtained on the simplex Σ^{n-1} when identifying it with the spherical sector S_+^{n-1} via the map $p = q^2$, $q \in S_+^{n-1}$. If the second derivatives vanish, i.e., if $p(x;\xi)$ is linear in ξ then

$$\Sigma_{k=1}^n \frac{1}{p_k} \frac{\partial p_k}{\partial \xi^i} \frac{\partial \xi^i}{\partial \xi^j} p_k = \frac{\partial^2}{\partial \xi^i \partial \xi^j} \Sigma_{k=1}^n p_k \log p_k$$

where $\Sigma_{k=1}^n p_k \log p_k$ is the entropy. As will be discussed later, negative of the entropy is a potential for the metric. The Fisher metric then induces a metric on any smooth family of probability measures on Ξ. Families of the form

$$p(x;\theta) = \exp(\gamma(x) + \Sigma_{i=1}^n f_i(x)\theta^i - \varphi(\theta))$$

where $\theta = (\theta^1, ..., \theta^n)$ is an n-dimensional parameter and $\gamma(x)$ and $f_1(x), ..., f_n(x)$ are functions on Ω are called exponential families. Of course the family is defined only for those θ for which $\int \exp(\gamma(x) + \Sigma f_i(x)\theta^i) dx < \infty$.

Fisher Metric Calculations For Some Distributions

Example 1 The normal distribution

The normal distribution $\frac{1}{\sqrt{2\pi\sigma}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$ on $\mathbb{R}$ with parameters μ and σ can easily be written in this form by putting,

$$\gamma(x) = 0,\ f_1(x) = x,\ f_2(x) = x^2,\ \theta^1 = \frac{\mu}{\sigma^2},\ \theta^2 = -\frac{1}{2\sigma^2}$$

$$\varphi(\theta) = \frac{\mu^2}{2\sigma^2} + \log\sqrt{2\pi\sigma} = -\frac{(\theta^1)^2}{4\theta^2} + \frac{1}{2}\log(-\frac{\pi}{\theta^2}).$$

For an exponential family, we have

$$\frac{\partial}{\partial\theta^i}\log p(x;\theta) = f_i(x) - \frac{\partial}{\partial\theta^i}\varphi(\theta)$$
$$\frac{\partial^2}{\partial\theta^i\partial\theta^j}\log p(x;\theta) = -\frac{\partial^2}{\partial\theta^i\partial\theta^j}\varphi(\theta)$$

This expression no longer depends on *x*, but only on the parameter θ. Therefore, the Fisher metric on such a family is given by

$$g_{ij}(p) = -E_p(\frac{\partial^2}{\partial\theta^i\partial\theta^j}\log p(x;\theta)) =$$
$$\int\frac{\partial^2}{\partial\theta^i\partial\theta^j}p(x;\theta)dx = \frac{\partial^2}{\partial\theta^i\partial\theta^j}\varphi(\theta)$$

For the normal distribution, we compute the metric in terms of θ^1 and θ^2, using $\frac{\partial^2}{\partial\theta^i\partial\theta^j}\varphi(\theta)$ and transform the result to the variables μ and σ to obtain

$$g(\frac{\partial}{\partial\mu},\frac{\partial}{\partial\mu}) = \frac{1}{\sigma^2}g(\frac{\partial}{\partial\mu},\frac{\partial}{\partial\sigma}) = 0g(\frac{\partial}{\partial\sigma},\frac{\partial}{\partial\sigma}) = \frac{2}{\sigma^2}.$$

As the Fisher metric invariant under diffeomorphism of $\Omega = \mathbb{R}$, and since $x \to x - \mu$ is such a diffeomorphism, it suffices to perform the computation at $\mu = 0$. The metric computed there, however, up to a simple scaling is the hyperbolic metric of the half plane

$$H := \{(\mu,\sigma); \mu \in \mathbb{R}, \sigma > 0\}.$$

Therefore, the Fisher metric on the family of normal distributions is the hyperbolic metric. To summarize, the Fisher metric is constructed as the natural Riemannian metric on a projective space over a linear space. In the finite case, this projective space is simply a spherical sector. In particular, this metric is the standard metric on the sphere, and it therefore has sectional curvature $\kappa = 1$. This fact is valid for the general case, as well. As a consequence, the Fisher metric is not Euclidean.

These type of calculations can be carried out faster and with less pain without a coordinate change by using the properties of the ex-pected value function. The expected value function for the discrete and continuous models are defined as follows:

- $E[x] = \sum_{x\in X} xf(x)$, discrete model.
- $E[x] = \int_{-\infty}^{\infty} xf(x)dx$, continuous model, where $f(x)$ represents the related probability density function.

Let $X, X_1, X_2, ..., X_n$ be real valued random variables with a common mean μ and $a_1, a_2, ..., a_n$, c be arbitrary constants.In this case, some useful properties of the expected value function is listed below:

1. $E[c] = c$,
2. $E\left[\Sigma_{i=1}^n a_i X_i\right] = \Sigma_{i=1}^n a_i E[X_i]$, Linearity.
3. $E[X_i] = \mu$ for all $i = 1,...,n$ and $Y = \Sigma_{i=1}^n X_i$ then $E[Y] = n\mu$.
4. $E[X_i X_j] = E[X_i]E[X_j]$ if X_i and X_j
5. $E[X-\mu] = E[X] - \mu = \mu - \mu = 0$ since $E[X] = \mu$.

Moreover, the n^{th} moment *E(X^n)* and n^{th} central moment $E[(X-\mu)^n]$ provides simplicity to the calculations. Since $E[x-\mu] = 0$ and the 3rd central moment $E[(x-\mu)^3] = 0$

$$g_{11}(\xi) - E_\xi[\partial_1\partial_1 l_\xi] = \int\frac{1}{\sigma^2}\frac{1}{\sqrt{2\pi\sigma}}\exp(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2})dx$$

let $\frac{x-\mu}{\sigma} = t$ and use the Gaussian integral $\int_{-\infty}^{\infty}\exp(-x^2)dx = \sqrt{\pi}$ to get $g_{11}(\xi) = \frac{1}{\sigma^2}$.

Similarly,

$$g_{12}(\xi) = E[\partial_1 l_\xi \partial_2 l_\xi] = E[-\frac{(x-\mu)}{\sigma^3} + \frac{(x-\mu)^3}{\sigma^5}] =$$
$$-\frac{1}{\sigma^3}E[(x-\mu)] + \frac{1}{\sigma^5}E[(x-\mu)^3] = 0.$$

As a Riemannian metric Fisher information metric is symmetric hence $g_{21}(\xi) = 0$.

One may compute $g_{22}(\xi)$ from the definition of variance $E[(x-\mu)^2] = \sigma^2$

$$g_{22}(\xi) = E[\partial_2 l_\xi \partial_2 l_\xi] = -E[\partial_2^2 l_\xi]$$
$$= -E[\frac{1}{\sigma^2} - \frac{3(x-\mu)^2}{\sigma^4}] = \frac{3}{\sigma^4}E[(x-\mu)^2] - \frac{1}{\sigma^2} =$$
$$\frac{3}{\sigma^2} - \frac{1}{\sigma^2} = \frac{2}{\sigma^2}$$

Example 2 Multivariate normal distribution

In the following sequel the sample space will be $\mathbb{R}^n$ in its standard vector coordinates $\{x^i\}_{1\le x^i \le n}$. Let $\mu \in \mathbb{R}$ be a mean vector and $\Sigma = [\sigma_{ij}]$ be an *nxn* symmetric, positive definite covariance matrix. Hence, the multivariate Gaussian distribution in mean and covariance parameters can be written as

$$p(x;\mu;\Sigma) = (2\pi)^{-1/2}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

For the multivariate Gaussian distribution, components of the Fisher information matrix is as follows:

$$I_{(\mu,\Sigma)} = (\partial_i \mu)^T \Sigma^{-1} (\partial_j \mu) + \frac{1}{2} Tr(\Sigma^{-1}(\partial_i \Sigma)\Sigma^{-1}(\partial_i \Sigma))$$

Since the parameter space of multivariate Gaussian Distribution is $n+\frac{n(n+1)}{2}$ dimensional with the Fisher metric it becomes a $n+\frac{n(n+1)}{2}$ dimensional Riemannian manifold. On the other hand, we can express the same distribution in its natural or in other words exponential parameters. In this case the probability density function and the Fisher metric can be expressed in terms of a potential function ϕ. One can rewrite the distribution in exponential or in other words natural coordinates defining

$$\theta = \Sigma^{-1}\mu$$

$$\Theta = -\frac{1}{2}\Sigma^{-1}$$

Representing ϑ as $\vartheta = (\theta;\Theta)$ we may write our probability density function in ϑ coordinates as

$$p(x;\vartheta) = \exp(\theta^T x + x^T \Theta x - \phi(\vartheta))$$

where $\phi(\vartheta) = \frac{1}{2}\left(n\log 2\pi - \frac{1}{2}\theta^T\Theta^{-1}\theta - \log(-2)^n |\Theta| \right)$

The detailed computations for the components of the Fisher metric in different coordinates can be found in [28] and [20].

Example 3 Poisson distribution

$$\chi = \{0,1,2,\ldots\}, n=1, \Xi = \{\xi \mid \xi > 0\}$$

$$p(x,\xi) = \exp(-\xi)\frac{\xi^x}{x!}$$

Since $n=1$ coefficients of the Fisher metric is represented by 1x1 matrix.

$$\log p(x;\xi) = -\xi + x\log(\xi) - \log(x!)$$

$$\frac{\partial \log p(x;\xi)}{\partial \xi} = -1 + \frac{x}{\xi}$$

$$G(\xi) = E[(\frac{\partial}{\partial\xi}\log p(x;\xi))^2] = -E[\frac{\partial^2}{\partial\xi^2}\log p(x;\xi)] =$$

$$-E_\xi[\frac{-x}{\xi^2}] = \frac{1}{\xi^2}E[x]$$

where

$$E[x] = \sum_{x=0}^{\infty} x\exp(-\xi)\frac{\xi^x}{x!} = \sum_{x=1}^{\infty} x\exp(-\xi)\frac{\xi^x}{x!} =$$

$$\sum_{x=1}^{\infty}\frac{\xi^{x-1}}{(x-1)!}\xi\exp(-\xi)$$

Since $\sum_{x=1}^{\infty}\frac{\xi^{x-1}}{(x-1)!} = \exp(\xi)$, $E[x]=\xi$. Therefore, the coefficients of the Fisher metric is $G(\xi) = \frac{1}{\xi}$.

Example 4 $P(\chi)$ for finite χ

$$\chi = \{x_0, x_1, \ldots, x_n\}$$

$$\Xi = \{[\xi^1,\ldots,\xi^n] \mid \xi^i > 0, \quad \forall i), \sum_{i=1}^{n}\xi^i < 1\}$$

Then,

$$p(x_i;\xi) = \begin{cases} \xi^i & 1 \le i \le n \\ 1 - \Sigma_{i=1}^n \xi^i & i = 0 \end{cases}$$

By using similar computations, on can shoe that the coefficients of the Fisher metric for finite discrete distribution is the *nxn* matrix

$$\begin{bmatrix} \frac{1}{1-\Sigma_{i=1}^n\xi^i} + \frac{1}{\xi^1} & . & . & \frac{1}{1-\Sigma_{i=1}^n\xi^i} \\ . & \frac{1}{1-\Sigma_{i=1}^n\xi^i} + \frac{1}{\xi^2} & . & . \\ . & . & . & . \\ \frac{1}{1-\Sigma_{i=1}^n\xi^i} & . & . & \frac{1}{1-\Sigma_{i=1}^n\xi^i} + \frac{1}{\xi^n} \end{bmatrix}$$

Connections

$p(x;\xi)$ be a n-dimensional smooth family of probability density functions depending on the parameter ξ. The components of the Fisher metric is of the following form:

$$g_{ij}(\xi) = E_\xi[(\frac{\partial}{\partial\xi^i}\log p(.;\xi)\frac{\partial}{\partial\xi^j}\log p(.;\xi))] \qquad (4.5)$$

$$= \int \frac{\partial}{\partial\xi^i}\log p(x;\xi)\frac{\partial}{\partial\xi^j}\log p(x;\xi)p(x;\xi) \qquad (4.6)$$

The Levi-Civita connection of the Fisher metric can be computed from the following formula:

$$\Gamma_{ij}^k = \frac{1}{2}g^{kl}(g_{il,j} + g_{jl,i} - g_{ij,l}) \qquad (4.7)$$

and note that $\Gamma_{ijk} = g_{il}\Gamma_{jk}^l$ where

$$\Gamma_{ijk}^{(0)} = E_\xi(\frac{\partial^2}{\partial\xi^i\partial\xi^j}\log p + \frac{1}{2}\frac{\partial}{\partial\xi^k}\log p + \frac{1}{2}\frac{\partial}{\partial\xi^i}\log p\frac{\partial}{\partial\xi^j}\log p\frac{\partial}{\partial\xi^k}\log p) \qquad (4.8)$$

yields the Levi-Civita connection $\nabla^{(0)}$ for the Fisher metric. More generally, a family of connections depending on a parameter $\alpha \in \mathbb{R}$ can be defined as follows

$$\Gamma_{ijk}^{(\alpha)} = E_\xi[(\frac{\partial^2}{\partial\xi^i\partial\xi^j}\log p + \frac{1-\alpha}{2}\frac{\partial}{\partial\xi^k}\log p + \frac{1}{2}\frac{\partial}{\partial\xi^i}\log p\frac{\partial}{\partial\xi^j}\log p\frac{\partial}{\partial\xi^k}\log p)]$$

$$\Gamma^{(\alpha)}_{ijk} = \Gamma^{(0)}_{ijk} - \frac{\alpha}{2} E_p[\frac{\partial}{\partial \xi^i}\log p \frac{\partial}{\partial \xi^j}\log p \frac{\partial}{\partial \xi^k}\log p] \quad (4.9)$$

In particular, since this expression is symmetric in indices *i* and *j*, all the connections $\nabla^{(\alpha)}$ are torsion-free and

$$\Gamma^{(\alpha)}_{ijk} + \Gamma^{(-\alpha)}_{ijk} = \Gamma^{(0)}_{ijk}$$

Therefore, the connections $\nabla^{(\alpha)}$ and $\nabla^{(-\alpha)}$ are dual to each other.

Example 5 The exponential family

$$p(x;\theta) = \exp(\gamma(x) + f_i(x)\theta^i - \varphi(\theta))$$

$\Gamma^{(1)}_{ijk} = 0$ is obtained by substituting $\alpha = 1$ in equation 4.9. Thus, θ yields an affine coordinate system for the connection $\nabla^{(1)}$. This flat affine connection is called the exponential connection and abbreviated as e-connection.

Example 6 The mixture family

$$p(x;\eta) = c(x) + \Sigma^{d}_{i=1} g^i(x)\eta_i$$

is an affine family of probability density functions. A simple computation yields $\Gamma^{(-1)}_{ijk} = 0$. In other words, η is an affine coordinate system for the connection $\nabla^{(-1)}$ so called the mixture or m-connection. One can find local potential functions of the Fisher metric in *x* and *n*-coordinates as in section 3.1. It is important to note that $\nabla^{(-1)}$ is not flat in *n*-coordinates. Therefore, two flat affine connections $\nabla^{(1)}$ and $\nabla^{(-1)}$ that are dual to each other with respect to the Fisher metric is obtained. To summarize, the space of probability measures can be viewed as a linear space in two different manners:

On one hand, as in the finite case, it can be repre-sented as a simplex in a vector space. Thus, any probability meas ure can be represented as a convex linear combination of certain extremal measures which is called the mixture representation.

On the other hand, space of probability measures can be represented as the exponential image of a linear tangent space which gives the so-called the exponential representation.

As it is discussed throughout the sections 3 and 4, these two structures are dual to each other, in the sense that each of them is the underlying affine structure for some connection, and the two corresponding connections are dual with respect to the Fisher metric. Of course, neither of these connections can be the Levi-Civita connection of the Fisher metric as the latter does not have vanishing curvature, [14]. Note that these results are valid locally for the global behavior one may refer to [5].

CURVATURE COMPUTATIONS AND SECOND ORDER ESTIMATION

As it is mentioned in sections 3 and 4, dually flat structures are defined through the affine structure by mutually dual flat affine connections, namely the exponential and the mixture connections. Since these connections are not Levi-Civita connection of the Fisher metric unless the underlying statistical manifold or distribution is flat Riemannian. The dually flat structure allows to define dual parallel transports, dual potential functions and geodesics by means of which alternative Taylor approximations of a function can be defined. On the other hand, probability distributions can be considered as Riemannian manifolds equipped with the Fisher information metric. Thus, curvatures of the induced geometry may be computed and used for the purposes of inference. Both perspective has advantages and disadvantages.

On an Kahler affine manifold, the differential D_γ of the difference tensor $\nabla - D$, where ∇ is the Levi-Civita connection of the Kahler affine metric, is called the affine(Hessian) curvature tensor and denoted by Q. Unlike the Riemannian curvature tensor on a Kahler affine manifold, affine curvature reflects the affine structure since its defined through the flat affine connection. Affine curvature tensor is of type (1, 3) and in local affine coordinates its components are of the form

$$Q^i_{ijk} = \frac{\partial \gamma^i_{jl}}{\partial x^k}, \quad (5.1)$$

[26, 27]. One can easily observe that for the Kahler affine metric

$$g_{ij} = \frac{\partial^2 \varphi}{\partial x^i \partial x^j}$$

$$Q_{ijkl} = \frac{1}{2}\frac{\partial^4 \varphi}{\partial x^i \partial x^j \partial x^k} - \frac{1}{2} g^{rs} \frac{\partial^3 \varphi}{\partial x^i \partial x^k \partial x^r}\frac{\partial^3 \varphi}{\partial x^j \partial x^l \partial x^s} \quad (5.2)$$

and its relation between Riemannian curvature tensor is given by

$$2R_{ijkl} = Q_{ijkl} - Q_{ijkl} \quad (5.3)$$

where the components of the Riemannnian curvature tensor is given by

$$R_{ijkl} = \frac{1}{4} g^{mn}\left(\frac{\partial^3 \varphi}{\partial x^m \partial x^l \partial x^i}\frac{\partial^3 \varphi}{\partial x^n \partial x^k \partial x^m} - \frac{\partial^3 \varphi}{\partial x^m \partial x^n \partial x^j}\frac{\partial^3 \varphi}{\partial x^n \partial x^k \partial x^i}\right) \quad (5.4)$$

It is remarkable that the Riemannian curvature of a Kahler affine metric depends only on the derivatives of the potential function to order at most three, whereas one would expect fourth derivatives of it to appear. Duistermaat gives some explanation for this phenomenon [9]. This property of the Fisher information metric allows us to avoid prolix Riemannian curvature computations in case of multivariate Gaussian distributions. On contrary to Riemannian curvature, affine curvature does not have this property which make its computation lengthy. Similary, one may define affine scalar curvature in local coordinates taking metric trace of the affine Ricci curvature tensor $Q_{ij}=Q_{ikj}^{k}$ as follows:

$$Q_{scal}=g^{ij}Q_{ij} \tag{5.5}$$

The concept of sectional curvature can also be carried to Kahler affine manifolds, [8, 26]. In this case, constant affine (Hessian) sectional Kahler affine manifolds can be constructed. There is a relationship between constant affine sectional Kahler affine manifolds and Riemannian manifolds of constant sectional curvature. If a Kahlerian manifold (*M,g*) is of constant affine sectional curvature *c* then (*M,g*) as a Riemannian manifold has constant sectional curvature $-\frac{c}{4}$.

Example 7 Curvature computations for normal (univariate Gaussian) distribution

Consider the normal distribution $p(x;\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}$ with mean μ and varience σ. In this case, Ricci tensor is of the form

$$R_{ij}=\begin{bmatrix}-\frac{1}{2\sigma^2} & 0\\ 0 & -\frac{1}{\sigma^2}\end{bmatrix} \tag{5.6}$$

Being the metric trace of Ricci curvature scalar curvature of the normal distribution is equal to -1. On the other hand, normal distribution in its natural coordinates is an exponential family with the coordinates $\theta^1=\frac{1}{2\sigma^2}$ and $\theta^2=\frac{\mu}{\sigma^2}$ where $\Theta=[\theta^1,\theta^2]$. For $\theta\in\Theta=\{[\theta^1,\theta^2]\,|\,\theta^1\in\mathbb{R}^+,\theta^2\in\mathbb{R}\}$ normal distribution in its natural coordinates is of the form

$$p(x;\theta^1,\theta^2)=\exp(\gamma_1(x)\theta^1+\gamma_2(x)\theta^2-\varphi(\Theta))$$

where $\gamma_1(x)=-x^2,\ \gamma_2(x)=x,\ \varphi(\theta)=\frac{(\theta^2)^2}{4\theta^1}+\frac{1}{2}\log(\frac{\pi}{\theta^1})$.

Then, the components of the Fisher metric and that of its inverse are as follows:

$$[g_{ij}(\theta)]=\begin{bmatrix}\frac{(\theta^2)^2}{2(\theta^1)^3}+\frac{1}{2(\theta^1)^2} & \frac{-\theta^2}{2(\theta^1)^2}\\ \frac{-\theta^2}{2(\theta^1)^2} & \frac{1}{2\theta^1}\end{bmatrix}, \tag{5.7}$$

$$[g^{-1}(\theta)]=\begin{bmatrix}2(\theta^1)^2 & 2\theta^1\theta^2\\ 2\theta^1\theta^2 & 2\theta^1+2(\theta^2)^2\end{bmatrix}. \tag{5.8}$$

Therefore,

$$R_{ij}(\theta)=\begin{bmatrix}-\frac{\theta^1+(\theta^2)^2}{4(\theta^1)^3} & \frac{\theta^1+\theta^1\theta^2+(\theta^2)^2}{4(\theta^1)^3}\\ \frac{\theta^1+\theta^1\theta^2+(\theta^2)^2}{4(\theta^1)^3} & \frac{1}{4\theta^1}\end{bmatrix} \tag{5.9}$$

is the Ricci tensor of the normal distribution in its natural coordinates and hence the scalar curvature is

$$R=\frac{\theta^1\theta^2+\theta^1(\theta^2)^2+(\theta^2)^3}{(\theta^1)^2} \tag{5.10}$$

The Riemannian curvature can be extended to α-connections. In section 3.1, α-connections are defined by 3.1 as $\Gamma_{ijk}^{(\alpha)}=\Gamma_{ijk}^{(0)}-\frac{\alpha}{2}\partial_i\partial_j\partial_k\varphi$.

Then, α-curvature tensor can be calculated from

$$R_{jkm}^{(\alpha)i}=\Gamma_{jm,k}^{(\alpha)i}-\Gamma_{jk,m}^{(\alpha)i}+\Gamma_{nk}^{(\alpha)i}-\Gamma_{jm}^{(\alpha)n}-\Gamma_{nm}^{(\alpha)i}-\Gamma_{jk}^{(\alpha)n}. \tag{5.11}$$

Note that original Riemannian curvature tensor is obtained for $\alpha=0$. Dual affine connections corresponds to the cases $\alpha=1$ and $\alpha=-1$, respectively.

For the normal distribution, only independent nonzero component of the α-curvature tensor is given by

$$R_{1212}^{(\alpha)}=\frac{1-\alpha^2}{\sigma^4} \tag{5.12}$$

Furthermore, only non-zere component of the α-Ricci tensor is

$$Ric_{11}^{(\alpha)}=\frac{\alpha^2-1}{2\sigma^2} \tag{5.13}$$

Note that more computations on different families of probability distributions one may refer to [1] and [24].

ACKNOWLEDGEMENTS

The author expresses her deep gratitude to TÜBİTAK (Turkish Scientific and Technological Research Council) since this work is supported under the project number 113F296 entitled as ”Harmonic maps, affine manifolds and their applications to information geometry”. The author is also grateful to Hitit University for the partial support by Hitit BAP: FEF01.13.006, “Afin Manifoldlarin Geometrisi”.

References

1. K. Arwini, C. T. J. Dodson, Information Geometry: Near Ran− domness and Near Independence, Lecture Notes in Mathematics, 1953, Springer, 2008.
2. S. Amari, Differential Geometrical Methods in Statistics, Springer Lecture Notes in Statistics 28, Springer−Verlag, Berlin, 1985.
3. S. Amari, Information Geometry and Its Applications, Applied Mathematical Sciences, 194, Springer, 2016.
4. S. Amari, H. Nagaoka, Methods of information geometry, Transl.Math. Monogr. 191, AMS & Oxford Univ. Press, 2000.
5. N.Ay, W.Tuschmann, Dually flat manifolds and global informa− tion geometry, Open Sys.& Information Dyn.9, 195−200, 2002.
6. S.Y.Cheng, S.T.Yau, The real Monge−Ampere equation and affine flat structures, in: S.S.Chern, W.T.Wu (eds.), Differential geom→etry and differential equations, Proc. Beijing Symp.1980, pp.339− 370, 1982.
7. N.N.Chentsov, Statistical decision rules and optimal inferences, AMS, 1982 (Translation of the Russian version, Nauka, Moscow, 1972).
8. O. Calin, C. Udriste, Geometric Modelling in Probability and Statistics, Springer,2014.
9. J. Duistermaat, Hessian Riemannian Structures, Assian Jour. Math, 5, 79−91, 2001.
10. A. P. Dawid, Invited discussion of Defining the statistical prob→lem (with applications to second−order efficiency, Ann. Statist., 3, 1231−1234, 1975.
11. B. Efron, Defining the Curvature of a Statistical Problem (with applications to second order efficiency), Ann. Statist., 3, 1189→1242, 1975.
12. H. Jeffreys, An invariant form for the prior probability in estima− tion problems, Proceedings of Royal Society of London, Series A, Mathematical and Physical Sciences, 186, 453−461, 1946.
13. J. Hofbauer, K. Sigmund, Evolutionaty games and population dynamics, Cambridge Univ Press, 1998.
14. J.Jost, Information geometry, Lecture Notes.
15. J. Jost, F. M. Şimsir, Affine harmonic maps, Analysis, 29, 185→197, 2009.
16. J. Jost, F. M. Simsir, Nondivergence harmonic maps, Harmonic maps and differential geometry, Contemp. Math., 542, 231−238, 2011.
17. T. Kurose, Dual connections and affine geometry, Mathematische Zeitschrift, 203, pp. 115−121, 1990.
18. R. E. Kass, P.W. Vos, Geometrical Foundations of Asymptotic Inference, Wiley Series in Probability and Statistics, New York, 1997.
19. S.L. Lauritzen, Statistical manifolds, in: Differential Geometry in Statistical Inference, Institute of Mathematical Statistics Lecture Notes, 10, pp. 163−218, Berkeley 1987.
20. L. Malago, G. Pistone, Information geometry of the Gaussian Dis→tribution in view of stochastic optimization, in: FOGA'15 Pro→ceedings of the 2015 ACM Conference on Foundations of Genetic Algorithms XIII, 150−162, Wales 2015.
21. M. K. Murray, J. W. Rice, Differential Geometry and Statistics, Chapman Hall,1993.
22. M. Noguchi, Geometry of Statistical Manifolds, Differantial Ge− ometry and its Applications 2, pp. 197−222, 1992.
23. C. R. Rao, Information and accuracy attainable in the estimation of statistical parameters, Bulletin of the Calcutta Mathematical Society , 37, 81−91, 1945.
24. Y. Sato, K. Sugawa and M. Kawaguchi, The geometrical structure of parameter space of the two dimensional normal distribution, Division of Information Engineering, Hokkaido Univ., Sapporo, Japan, 1977.
25. C. E. Shannon, A mathematical theory of communication, Bell. Syst. Tech. J., 27, 379−423 and 623−656, 1948.
26. H. Shima, The Geometry of Hessian Structures, World Scientific, 2007.
27. H. Shima, Hessian manifolds of constant Hessian sectional curva− ture, J. Math. Soc. Japan, 47(4), 735−753, 1994.
28. L. T. Skovgaard, Riemannian geometry of the multivariate nor→mal model, Scandinavian Journal of Statistics, 11(4), pp. 211−223, 1984.