Generalized Geometric Structures on Statistical manifolds

M. Sitaramayya
School of Maths. & Stats.,
University of Hyderabd, Hyd.-500046

email: malladi_sr@yahoo.com
K.S. Subrahamanian Moosath & K.V. Harsha
Department of Mathematics ,
IIST, Valiamalai, Thiruvananthapuram 695547

Subject class [2010]:53C25, 53C50
Keywords: Semi-symmetric metric connection, curvature tensor, Ricci tensor.

1 Introduction

Information geometry is of recent origin started around 1985 and it is the result of applying non-euclidean geometry(Riemannian and Affine) to the probability theory in general and statistical inference and estimation in particular. This is very effective in applications to Neural networks and in Machine learning thru the gradient-descent method by the uniqueness of the Fisher metric.

Metrics and distances (or semidistances like divergences) between probability distributions play an important role in problems of statistical inference and estimation and also in practical applications to study affinities (or dissimilarities) among a given set of populations. A statistical model is specified by a family of probability distributions, normally described by a set of continuous parameters called a parameter space. The parameter space possesses some geometrical properties which are induced by the local information contents and structures of the distributions. Starting from the Fisher's pioneering work, in 1925 the study of these geometrical properties has received much attention in statistical literature. Then Rao [27] introduced a Riemannian metric in terms of the Fisher information matrix over the parameter space of a parametric family of probability distributions in 1945.

Since then many statistians have attempted to construct a geometrical theory in probability spaces. It was only in 1975, Efron [18] was able to introduce a new affine connection into the geometry of the parameter spaces and thereby elucidating the role of curvature in statistical problems. Significant contributions to Efron's work were added by Reeds [29] and Dawid [18]. In fact Dawid even suggested a geometrical foundation for Efron's approach as well as pointing out the possibility of introducing other affine connections into the geometry of parameter spaces (see Amari [1,3], Burbea-Rao [12,13,14].

Some work in this direction was also done by Cencov [16] and also Atkinson-Mitchell [4], independent of Cencov's work computed the Rao distance for a number of parametric families of important probability distributions of statistics. In this article we systematically study the geometry of statistical manifolds. The material is arranged as follows.

§2 gives a general setting for statistical models; in §3 statistical model as a smooth manifold and Fisher metric are introduced and illustrated with an example; §4 studies weighted Fisher metric and its associated geometry; in §5 Amari's family of α-affine connections were studied and illustrated with two examples; §6 investigates dual connections and dual flatness and their topological implications; §7 studies generalization of α-structures via) the divergence approach and skewness tensor approach and the existence of dual potential scheme on dually flat spaces is brought out; and also various equivalent ways of defining a statistical structure are given; in §8 and 9 the geometry of Amari (α-geometry) interpreted as geometry of special embeddings (α-embeddings) was further generalised as (F, G)-geometry on a statistical model and in §10 the special place of Amari's α-geometry among all possible (F, G)-geometries was established and also the invariant and non-inveriant properties of general (F, G)-geometries given. The presentation is elementary with clarity of concepts (through geometric computations are indicated sketchily) illustrated by many examples.

2 General Setting

Let μ be a σ -finite additive measure defined on a σ-algebra of the subsets of a measurable space χ. Then $\mathcal{M} = M(\chi; \mu)$ stands for the space of all μ-measurable functions onn χ and let $\mathcal{L} = \mathcal{L}(\chi; \mu)$ denote the space of all $p \in \mathcal{M}$ so that $\|p\| \equiv \int_\chi |p(x)| d\mu(x) = \int_\chi |p| d\mu < \infty$.

Let $\mathcal{M}_+ = \mathcal{M}_+(\chi; \mu)$ denote the set of all $p \in \mathcal{M}$ such that $p(x) \in R_+ = [0, \infty)$ for all μ-almost all $x \in \chi$ and define $\mathcal{L}_+ = \mathcal{L}_+(X; \mu)$ as $\mathcal{L}_+ = \mathcal{M}_+ \cap \mathcal{L}$ and we let $\mathbb{P} = \mathbb{P}(\chi; \mu)$ denote the set of all $p \in \mathcal{L}_+$ with $\|p\|_\mu = 1$. Clearly $\mathbb{P}$ is a convex subset of $\mathcal{L}_+$. In most applciations $\chi = \mathbb{R}^n$, $\mu =$ Lebesgue measure on $\mathbb{R}^n$. So any probability measure on χ can be represented in terms of density function with respect to Lebesgue measure μ.

Definition 1: A probability distribution on χ is a function $p : \chi \to \mathbb{R}$ such that $p(x) \geq 0$ $\forall x \in \chi$ and $\int_\chi p(x) dx = 1$ (cont. case) ($\sum_{x \in \chi} p(x) = 1$, χ is a discrete finite or countable set and μ is counting measure).

Let $\theta = (\theta^1, \theta^2, \cdots, \theta^n)$ be a set of real continuous parameters belonging to a parameter space, Θ a manifold embedded in $\mathbb{R}^n$ and let $F_\Theta = \{p(\cdot, \theta) \in \mathcal{L}_+ : \theta \in \Theta\}$ be a parametric family of positive distributions } and $P_\Theta = \{p(\cdot, \theta) \in \mathbb{P} : \theta \in \Theta\}$ be a parametric family of probability distributions on χ and P_Θ is a convex subfamily of F_Θ.

We denote by $S = P_\Theta = \{p_\theta = p(x, \theta) : \theta \in \Theta \subset \mathbb{R}^n\}$ and call S a n-dimensional statistical model or parametric model. We put some regularity conditions on S.

1. Θ is an open subset of $\mathbb{R}^n$ and the map: $\theta \in \Theta \to p(\cdot, \theta) \in \mathbb{R}$ is of class C^∞, and the map: $\theta \to p(x, \theta), x \in \chi$ is injective.
2. Let $l(x, \theta) = \log p(x, \theta)$ and $\partial_i = \frac{\partial}{\partial \theta^i}$. For every fixed θ, the n functions $\{\partial_i l(x, \theta\}_{i=1}^n$ of x on χ are linearly independent and these functions are called scores.
3. The order of integration and differentiation may be freely interchanged on χ.
4. The moments of scores exist upto necessary orders.
5. The support of the function $p_\theta : \chi \to \mathbb{R}$, does not vary with θ where $\mathrm{supp}(p_\theta) = \{x \in \chi | p(x, \theta) > 0\}$. So se can redefine χ to be $\mathrm{supp}(p_\theta)$ so that $p(x, \theta) > 0$ holds forall $\theta \in \Theta$ and all $x \in \chi$. Thus the statistical model S is a subset of $\mathbb{P}(\chi) = \{p : \chi \to \mathbb{R}|_{p(x)>0}, \ \forall x \in \chi, \ \int_\chi p(x) dx = 1\}$

3 Statistical Model S as smooth manifold, Fisher matrix on S

Definition 2: For a statistical model $S = \{p_\theta : \theta \in \Theta \text{ open} \subset \mathbb{R}^n\}$, the mapping $\varphi : S \to \mathbb{R}^n$ defined by $\varphi(p_\theta) = \theta$ allows us to consider $\varphi = (\theta^i)$ as a coordinate system for S and hence S becomes a smooth n-dimensional manifold called statistical manifold with a global coordiante system (θ^i). Note that any C^∞-diffeomorphism ψ of $\mathbb{R}^n$ to itself induces a reparametrization of S by $\theta \to \rho = \psi(\theta)$.

Example 3: Normal distribution $N(\mu, \sigma)$: Take $\chi = \mathbb{R}$, $n = 2$, $\theta = (\mu, \sigma) = (\theta^1, \theta^2)$ and $\Theta = \{(\mu, \sigma) \in \mathbb{R}^2 | -\infty < \mu < \infty \text{ and } 0 < \sigma < \infty\}$

and $N(\mu, \sigma) = \left\{ p(x, \theta) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\dfrac{(x-\mu)^2}{2\sigma^2} \right)_{\theta = (\mu,\sigma) \in \Theta} \right\}$

This is a 2-dimensional statistical manifold which can be identified with the upper half plane $H^+ = \{(\mu, \sigma) \in \mathbb{R}^2|^{\mu \in \mathbb{R}}_{\sigma \in \mathbb{R}^+}\}$ which has negative curvature with respect to Poincaré metric.

Definition 4: Let $l_\theta = \log p_\theta$. Regard $l_\theta : \chi \to \mathbb{R}$ sending $x \to l_\theta(x) = \log p(x, \theta)$. We define the Fisher information matrix $(g_{ij}(\theta))_{n \times n}$ of the manifold S at state θ by

$$g_{ij}(\theta) = E_\theta[\partial_i l(\theta) \partial_j l(\theta)] == \int_\chi \partial_i l(\theta) \partial_j l(\theta) p(x, \theta) dx \tag{3.1}$$

Note here E_θ is the expectation w.r.t p_θ of [ ].

Here we assume that the above integral exists for all $\theta \in \Theta$. Note that $G(\theta) = (g_{ij}(\theta))$ is a symmetric $n \times n$ matrix.

For any $c = [c^1, c^2, \cdots, c^n]^t \in \mathbb{R}^n$, $c^t G(\theta) c = \int_\chi \left\{ \sum_{i=1}^n c^i \partial_i l(x, \theta) \right\}^2 dx$ and so $G(\theta)$ is positive definite since $\{\partial_i l(\theta)\}_{i=1}^n$ are linearly independent. So the Fisher matrix $G(\theta)$ is a symmetric positive definite matrix and hence defines an inner product on the tangent space of the statistical manifold S denoted by $<,>_\theta$ called the Fisher metric on S and this $<,>$ gives a Riemannian metric (called Rao's metric) on the statistical manifold S, making (S, g) a Riemannian manifold.

Example 5 (continued) : $S = N(\mu, \sigma) = \left\{ p(x, \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right) \big| \theta = (\mu, \sigma) \in \Theta \subset \mathbb{R}^2 \right\}$ with parameter $\theta = (\mu, \sigma)$. Then $l(x, \theta) = \log p(x, \theta) = -\frac{(x-\mu)^2}{2\sigma^2} - \log(\sqrt{2\pi}\sigma)$.

Let $\partial_1 = \frac{\partial}{\partial \mu}$, $\partial_2 = \frac{\partial}{\partial \sigma}$. Then $\partial_1 l = \frac{(x-\mu)}{\sigma^2}$. $\partial_2 l = -\frac{(x-\mu)^2}{\sigma^3} - \frac{1}{\sigma}$.

Hence the Fisher information matrix

$$G(\theta) = (g_{ij}(\theta))_{2\times 2} \qquad \text{is given by} \qquad G(\theta) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix}$$

using formula (3.1).

6. Remarks: 1) The Fisher metric allows us to study "nearness" of two probability distributions [20] and the geodesic distance between two points of S using differential geometric techniques.

2) Note that the Fisher metric $<,>_\theta$ is indeed a metric on the parameter space $\Theta \subset \mathbb{R}^n$ and the Fisher metric $g_{ij} = E_\theta[\partial_i l \partial_j l]$
measures the expectation of the 2nd order cummulants of the variations of the log-likelihood function $\log p(x, \theta)$.

3)$g_{ij} = E_\theta[\partial_i l \partial_j l] = -E_\theta[\partial_i \partial_j l]$ using

$$\int_\chi p(x)\partial_i l dx = \int_\chi \partial_i p dx = \partial_i \int_\chi p(x)dx = 0.$$

That is $E_\theta(\partial_i l) = 0$, $i = 1, 2, \cdots, n$.
Similarly $h_{ijk}(\theta) = E_\theta[\partial_i l, \partial_j l \partial_k l]$ tensor defines a richer structure on S (details later).

4) Such Fisher metric $g =<,>$ is unique having the two properties
(a) (g_{ij}) is invariant under reparametrization of the sample space χ.
(b) g_{ij} is covariant under reparametrization of the parameter space Θ [16,17].

5) Using the Fisher metric g, the Christoffel symbol of first kind is given by

$$\Gamma_{ij} = \frac{1}{2}[\partial_i g_{jk} + \partial_j g_{ki} - \partial_k g_{ij}\}$$

by standard Riemannian geometries [24].
Thus (S, g) is a Riemmanian manifold. We give later on more geometric structure on the statistical manifold S.

Remarks 7: Properties of the Fisher Metric g on S.
(i) Suppose our probability distributions are defined in terms of a random variable x taking values in $X \subseteq \mathbb{R}^n$. Then

$$g_{ij}(\theta) = \int_\chi \frac{1}{p_\theta(x)} \partial_i p_\theta(x) \partial_j p_\theta(x) dx \qquad (by(3.1))$$

We can re-express this in terms of another random variable y taking values in $Y \subseteq \mathbb{R}^n$, if we suppose that $y = f(x) : X \to Y$ is an invertible mapping. We clearly have

$$\tilde{p}_\theta(y) = \int_\chi p_\theta(x)\delta(y - f(x))dx \tag{3.2}$$

If f is invertible then using the relation

$$\delta(y - f(x)) = \frac{1}{|\frac{\partial f}{\partial x}|}\delta(f^{-1}(y) - x)$$

we can find that

$$\begin{aligned} \tilde{p}_\theta(y) &= \int_\chi p_\theta(x) \frac{1}{|\frac{\partial f}{\partial x}|}\delta(f^{-1}(y) - x) \\ &= \left[\frac{1}{|\partial f/\partial x|} p_\theta(x)\right]_{x=f^{-1}(y)} \qquad (3.3) \end{aligned}$$

Since $|\frac{\partial f}{\partial x}|$ does not depend on θ we see that

$$\begin{aligned}\int_Y \frac{1}{\tilde{p}_\theta(y)}\partial_i\tilde{p}_\theta(y)\partial_j\tilde{p}_\theta(y)dy &= \int_Y dy\left[\frac{1}{\frac{1}{|\frac{\partial f}{\partial x}|}p_\theta(x)}\frac{\partial_i p_\theta(x)}{|\frac{\partial f}{\partial x}|}\frac{\partial_j p_\theta(x)}{|\frac{\partial f}{\partial x}|}\right]_{x=f^{-1}(y)}\\ &= \int_X \frac{dx}{p_\theta(x)}\partial_i p_\theta(x)\partial_j p_\theta(x) = g_{ij}(\theta) \text{ by (3.1) using } \int_Y dy = \int_X dx|\frac{df}{\partial x}|\end{aligned}$$

Thus we proved. **Propsition 8:** (a) The Fisher metric $(g_{ij}(\theta))$ is invariant under transformation of the random variable.
(b) It is also covariant under reparametrization: proof of (b) Suppose $(\tilde{\theta}^i)$ is a new set of coordinates on $\Theta \subset \mathbb{R}^n$ and (θ^i) is the original set of coordinates on $\Theta \subset \mathbb{R}^n$. They are related by the invertible relationship $\tilde{\theta} = \tilde{\theta}(\theta)$. Defining $\tilde{p}_{\tilde{\theta}}(x) \equiv p_{\theta(\tilde{\theta})}(x)$, we are then able to compute $\tilde{g}_{ij}(\tilde{\theta}) \equiv \int_\chi dx \frac{1}{\tilde{p}_{\tilde{\theta}}(x)}\frac{\partial}{\partial\tilde{\theta}^i}\tilde{p}_{\tilde{\theta}}(x)\frac{\partial}{\partial\tilde{\theta}^j}\tilde{p}_{\tilde{\theta}}(x)$ in terms of $g_{ij}(\theta)$. For, since $\frac{\partial}{\partial\tilde{\theta}^i}\tilde{p}_{\tilde{\theta}} = \frac{\partial\theta^j}{\partial\tilde{\theta}^i}\frac{\partial}{\partial\theta^j}p_{\theta(\tilde{\theta})}$, we may directly conclude that

$$\tilde{g}_{ij}(\tilde{\theta}) = \left[\frac{\partial\theta^k}{\partial\tilde{\theta}^i}\frac{\partial\theta^l}{\partial\tilde{\theta}^j}g_{kl}(\theta)\right]_{\theta=\theta(\tilde{\theta})} \tag{3.4}$$

which is precisely the covariance of Fisher metric under reparametrization of the parameter space.
Example 9: On the space of normal distributions $N(\mu, \sigma^2)$ the Fisher metric is a special one.

$$\begin{aligned}N(\mu,\sigma^2) =&\{p(x,\theta) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-0\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R},\\ &\theta \in \Theta = \mathbb{R}\times\mathbb{R}^+ \quad \text{and} \quad \theta = (\mu,\sigma^2) \in \Theta, \quad \mu \in \mathbb{R},\ \sigma \in \mathbb{R}^+\end{aligned}$$

The Fisher metric is

$$ds^2 = \frac{1}{\sigma^2}\left[\left(\frac{d\mu}{\sqrt{2}}\right)^2 + (d\sigma)^2\right]$$

and its curvature is the Gauss curvature $\kappa = -\frac{1}{2}$.
We complexify the situation. Letting $\mu^* = \frac{\mu}{\sqrt{2}}$ and introducing the complex variable $z = \mu^* + i\sigma$ we can identify $\Theta \subset \mathbb{R}^2$ with the upper half plane $H^+ = \{z \in \mathbb{C} : Imz > 0\}$ of $\mathbb{C}$ and the Fisher metric is then.

$$\begin{aligned}ds^2 =& \frac{2}{\sigma^2}\left[\left(\frac{d\mu}{\sqrt{2}}\right)^2 + d\sigma^2\right] = \frac{2}{\sigma^2}\left[(d\mu^*)^2 + d\sigma^2\right]\\ =&\frac{2}{\sigma^2}dzd\bar{z} \quad \text{which is the standard Poincaré metric on} \quad H^+\end{aligned}$$

Hence the geodesic curves in the parameter space are the "semicircles" $z = a + re^{i\varphi}$, $r > 0$, $0 < \varphi < \pi$, $a \in \mathbb{R}$ constant with limiting case (as $r \to +\infty$) as half lines (vertical) $Rez =$ constant, $z \in$ Ⓗ

Remark 10: The corresponding multinormal distribution family on $\mathbb{R}^{2n} \approx \mathbb{C}^n$ $\{p(\underline{x}; \hat{\mu}, \Sigma) : x \in \mathbb{R}^{2n},$ Ⓗ $= \mathbb{R}^{2n} \times p(2n, \mathbb{R})\}$ has Fisher metric with rich information geometry which needs further investigation (cf. [30])[9]).

Here $\hat{\mu}$ is mean vector in $\mathbb{R}^{2n}$, Σ is a positive definite symmetric $2n \times 2n$ real matrix and $p(2n, \mathbb{R})$ is the set of all such Σs

4 Weighted Fisher metric and associated geometry

4.1. The Reimannian geometric aspects of the R manifold (S, g):

Let $p = p(x, \theta) \in S$. The $dp = dp(\cdot, \theta) = \sum_{i=1}^{n} (\partial_i p) d\theta_i$.

Define the metric element $ds^2(\theta) = E_\theta [d \log p]^2$

$$= \int_\chi p(\theta) d(\log p)^2 d\mu = \int_\chi p(\theta) \left(\sum_{i=1}^{n} \partial_i \log p d\theta_i \right) \left(\sum_{j=1}^{n} \partial_j \log p d\theta_j \right) d\mu$$

$$= \sum_{i,j=1}^{n} g_{ij}(\theta) d\theta_i d\theta_j \quad \text{where} \quad g_{ij}(\theta) = \int_\chi \frac{1}{p} \partial_i p(\theta) \partial_j p(\theta) d\mu \qquad (3.5)$$

which is in the standard form

Since $g_{ij}(\theta)$ gives $(i, j = 1 \text{to} n)$ a covariant symmetric tensor of order 2 forall $\theta \in$ Ⓗ $\subset \mathbb{R}^n$, $ds^2(\theta)$ is invariant under the admissible transformations of the parameters. Using the line element $ds = \sqrt{ds^2(\theta)}$ one can find the distance between two points $\theta^{(1)}, \theta^{(2)} \in$ Ⓗ along a curve γ given by $\theta(t)$ in Ⓗ as

$$(3.6) \qquad S(\theta^{(1)}, \theta^{(2)}) = \left| \int_{t_1}^{t_2} \frac{ds(\theta)}{dt} dt \right| = \left| \int_{t_1}^{t_2} \left\{ \sum_{i,j=1}^{n} g_{i,j}(\theta) \dot{\theta}_i \dot{\theta}_j \right\}^{1/2} dt \right|$$

where $\theta^{(1)} = \theta(t_1)$ and $\theta^2 = \theta(t_2)$ and $\dot{\theta}_i = \frac{d\theta_i}{dt}$.

Such curve γ joining $\theta^{(1)}$ and $\theta^{(2)}$ in Ⓗ such that $S(\theta^{(1)}, \theta^{(2)})$ is the shortest is called an information geodesic curve, and such distance is called the information geodesic distance or the Rao distance between $p_{\theta^{(1)}}$, $p_{\theta^{(2)}}$ distributions for the Fisher metric.

The geodesic curve $\theta = \theta(t)$ in Ⓗ from $\theta^{(1)}$ to $\theta^{(2)}$ may be determined from the Euler-Lagrange equations (a system of n nonlinear equations of second order)

$$\sum_{i=1}^{n} g_{ij} \ddot{\theta}_i + \sum_{i,j=1}^{n} \Gamma_{ijk} \dot{\theta}_i \dot{\theta}_j = 0 \quad (k = 1, 2, \cdots, n) \qquad (3.7)$$

with boundary conditions

$$\theta_i(t_j) = \theta_i^{(j)} \quad (i = 1, 2, \cdots, n, j = 1, 2) \qquad (3.8)$$

and Γ_{ijk} are the Christoffel symbols of first kind of the metric ds^2 given by

$$\Gamma_{ijk} = \frac{1}{2}[\partial_i g_{jk} + \partial_j g_{ki} - \partial_k g_{ij}]$$

and then the geodesic distance or information distance or Rao distance explicitly calculated. As observed by Rao, this distance gives a measure of dissimilarity between two distributions in a statistical model. For various models (discrete and continuous) Rao's distances were determined in the literature ([28],[4],[13],[14]).

4.2. Weighted Fisher Metric

Definition 11: Let f be a positive function on $\mathbb{R}_+$ of class (at least) $\mathbb{C}^2$ (Burbea takes cont. f), $S = \{p(x,\theta)|\theta \in \textcircled{H} \subset R^n,\ x \in \chi\}$ be parametrized family of distributions. Define the metric element (weighted by f)

$$ds_f^2(\theta) \equiv \int_\chi \frac{f(p)}{p}[dp]^2 d\mu \qquad (4.1)$$

$$= \sum_{i,j=1}^{n} g_{ij}^{(f)}(\theta) d\theta_i d\theta_j \qquad (4.2) \quad \text{with}$$

$$g_{ij}^{(f)}(\theta) = \int_\chi \frac{f(p)}{p}(\partial_i p)(\partial_j p) d\mu \qquad (4.3)$$

Then the weighted Fisher information $n \times n$ matrix $G(\theta) = (g_{ij}^{(f)}(\theta))$ is positive definite for every θ and hence ds_f^2 gives a Riemannian metric on the parameter space $\textcircled{H}$ called f-weighted Fisher metric on $\textcircled{H}$.
In the language of expectations this metric can be written as

$$(4.4) \qquad ds_f^2(\theta) = E_\theta[(f \circ p)(d \log p)^2] \qquad \text{with}$$

$$(4.5) \qquad g_{ij}^{(f)}(\theta) = E_\theta[(f \circ p)(\partial_i \log p)(\partial_j \log p)]$$

We can give a geometric meaning for the metric element $ds_f^2(\theta)$ in terms of (weighted) average of Fisher information between two states as follows:

Remark 12: (a) In information theory the quantity $\log p(\cdot,\theta)$ for $p(\cdot,\theta) \in S$ is known as the amount of "self-information" associated with the state $\theta \in \textcircled{H}$. The self-information at a nearby state say $\theta + \delta\theta \in \textcircled{H}$ is then $\log(\cdot, \theta + \delta\theta)$. Then the difference $\log p(\cdot, \theta + \delta\theta) - \log(\cdot, \theta)$ is approximated by the first order differential $d \log p = \sum_{i=1}^{n} (\partial_i \log p) d\theta_i$, and hence $ds_f^2(\theta)$ measures the weighted average of the square of this first order difference with the weight $f[p(\cdot,\theta)]$. For this reason, the metric ds_f^2 and the matrix $[g_{ij}^{(f)}(\theta)]$ are called the f-information metric and the f-information matrix on the parameter space $\textcircled{H}$.
The above usual Riemannian geometry of §3 can be calculated for $ds_f^2(\theta)$ and the f-geodesic distance between two distributions can be computed.

4.3. Further Interpretations of $ds_f^2(\theta)$:

(b) Regard $ds_f^2(\theta)$ as a functional of $p(\cdot,\theta) \in S$. This functional is convex in $p(\cdot,\theta) \in S$ iff the function $F(x) = \frac{x}{f(x)}$ is concave on $\mathbb{R}_+$.

In particular, since f is a C^2-function on $\mathbb{R}_+$, $F(x) = \frac{x}{f(x)}$ is concave on $\mathbb{R}_+ \Leftrightarrow FF'' \geq 2(F')^2$ on $\mathbb{R}_+$.
Example 13: Take $f(x) = x^{\alpha-1}$. Then we get the "α-order information metric" $ds^2_\alpha(\theta) = E_\theta[p^{\alpha-1}(d\log p)^2]$ and the corresponding "α-order information matrix" $G(\theta) = [g^{(\alpha)}_{ij}(\theta)]$ with $g^{(\alpha)}_{ij}(\theta) = E_\theta[p^{\alpha-1}(\partial_i \log p)(\partial_j \log p)]$.
Then $ds^2_\alpha(\theta)$ is convex in $p(\cdot,\theta)$ iff $1 \leq \alpha \leq 2$.
For $\alpha = 1$, we get $ds^2 =$ the Fisher metric, $[g_{ij}]$ =the Fisher matrix and geodesic distance is the usual Rao distance. For the explicit computations of Rao distances for several well known families of distributions of statistics refer to Rao-Burbea [12],[13],[14].
(c) The burbea metric $ds^2_f(\theta)$ arises as the second order differential of certain entropy or divergence functional along the direction of the tangent space of the parameter space $\Theta \subset \mathbb{R}^n$ at $\theta \in \Theta$.
For example, let $F(\cdot,\cdot)$ be a C^2-function on $\mathbb{R}_+ \times \mathbb{R}_+$ and consider the "F-divergence" defined by

$$D_F(p,q) \equiv \int_\chi F[p(x),q(x)]d\mu(x) \quad (p,q,\in S)$$

Assume (i) $F(x,\cdot)$ is strictly convex on $\mathbb{R}_+$ for $\forall\, x \in \mathbb{R}_+$ (ii) $F(x,x) = 0\ \forall\, x \in \mathbb{R}_+$ (iii) $\partial_y F(x,y)|_{y=x}$ is constant for every $x \in \mathbb{R}_+$.
Then for $p(\cdot,\theta^{(1)})$ and $p(\cdot,\theta^{(2)}) \in S = \mathbb{P}_\Theta$ we write

$$D_F(\theta^{(1)},\theta^{(2)}) \equiv D_F[p(\cdot,\theta^{(1)}),p(\cdot,\theta^{(2)})] \qquad (\theta^{(1)},\theta^{(2)} \in \Theta)$$

Then, for $p(\cdot,\theta) \in S = \mathbb{P}_\Theta$ and $\theta \in \Theta$

$$D_F(\theta,\theta) = 0, \quad dD_F(\theta,\theta) = \int_\chi \partial_y F(p,y)|_{y=p}(dp)d\mu = 0$$

and $d^2D_F(\theta,\theta) = ds^2_f(\theta)$
where $f(x) = x\partial^2_y F(x,y)|_{y=x}(x \in \mathbb{R}_+)$
Hence it follows that to the second order infinitesimal displacements $D_F(\theta.\theta+\delta\theta) = \frac{1}{2}ds^2_f(\theta)$.

So the Burkea metric ds^2_f is realized as a second order differential of a divergence on S.

§5. Amari's α-family of affine connections on (S,g)

Let $S = \{p_\theta|\theta \in \Theta \subset \mathbb{R}^n\}$ be a statistical model and g is the Fisher metric. Let $\Gamma_{ijk} = \frac{1}{2}\{\partial_i g_{jk} + \partial_+ j g_{ik} - \partial_k g_{ij}\}$ define the Levi-Civita connection $\nabla^{(0)}$ on S which is unique affine (linear) connection such that its covariant derivative on the metric g vanishes.
Definition 14: On $G(S,g)$ define 3-covariant totally symmetric tensor T (called Skewness tensor) by $T_{ijk} = E_\theta[\partial_i l \partial_j l \partial_k l]$ (5.1)
Let $\alpha \in \mathbb{R}$. Define the affine α-connection $\nabla^{(\alpha)}$ on (S,g) by giving its connection coefficients as

$$\Gamma^\alpha_{ijk} = \Gamma_{ijk} - \frac{\alpha}{2}T_{ijk} \tag{5.2}$$

(equivalently in the language of expectations)

$$\Gamma^{\alpha}_{ijk} = E_{\theta}[(\partial_i \partial_j \log p)\partial_k \log p)] + \frac{1-\alpha}{2} T_{ijk} \tag{5.3}$$

Thus on (S, g) we have a one parameter family of affine connections $(\nabla^{\alpha})_{\alpha \in \mathbb{R}}$. Note that $\nabla^{(0)}$ is the metric connection the only one in the family. Each connection $\nabla^{(\alpha)}$ induces a parallellism on Θ.

Remarks 15: (1) The 1-connection ($\alpha = 1$) is given by Efron and studied its geometry [18] and the $\alpha = -1$ connection by Dawid [18].

We give two examples here and summation convention used in them.

Example 16: Consider an exponential family $S = \mathbb{P}_{\Theta}$ of distributions given by (5.4) $p(x, \theta) = \exp\{T(x) + T_i(x)\theta_i - \psi(\theta)\}$ $\quad (x \in \chi)$ with

$$(5.5) \qquad e^{\psi(\theta)} = \int_{\chi} e^{T_i(x)\theta_i} \cdot e^{T(x)} d\mu(x) \qquad (\theta \in \Theta \subset \mathbb{R}^n)$$

parameterized by the natural parameters $\theta = (\theta^1, \theta^2, \cdots \theta^n) \in \Theta$.

Hence ψ is a C^2-function on $\Theta \subset \mathbb{R}^n$ and $T, T_1, T_2, \cdots T_n$ are smooth functions (measurable enough in general) on χ. Then $\partial_i l = \partial_i \log p = T_i(x) - \partial_i \psi(\theta)$ $\partial_i \partial_j l = -\partial_i \partial_j \psi$.

Hence the Fisher metric $g_{ij} = \partial_i \partial_j \psi$ (5.6) and (5.7) $E_{\theta}[(\partial_i \partial_j l)\partial_k l[= 0$ and $\Gamma^{\alpha}_{ijk} = \frac{1-\alpha}{2} T_{ijk}$ (5.8) (from 5.3).

We note from (5.8) $\Gamma^{\alpha}_{ijk} \equiv 0$ for $\alpha = 1$ and hence this manifold (S, g) is 1-flat or the exponential family constitutes an uncurved space with respect to the ($\alpha = 1$) Efron connection or exponential connection.

Example 17: Consider the parametric family $S = P_{\Theta} \equiv P_{\Theta}(q_1, q_2, \cdots, q_{n+1})$ of distributions $p(\cdot, \theta)$ given by a mixture of $n + 1$ prescribed linearly independent probability distributions on χ as

$$(5.9) \quad p(x, \theta) = q_i(x)\theta_i + q_{n+1}(x)\theta_{n+1} \quad (x \in \chi)$$

where $\theta_{n+1} = 1 - (\theta_1 + \theta_2 + \cdots + \theta_n)$ and $\theta \in \Theta$
with $\Theta = \{\theta = (\theta_1, \theta_2, \cdots \theta_n) \in \mathbb{R}^n_+ \colon \theta_{n+1} > 0\}$. Then $\partial_i l = \partial_l \log p = p^{-1}(q_i - q_{n+1})$, $\partial_i \partial_j l = -(\partial_i l)(\partial_j l)$ and

$$E_{\theta}[(\partial_i \partial_j l)\partial_k l] = -T_{ijk} \qquad (5.10) \quad \text{and}$$

$$\Gamma^{\alpha}_{ijk} = -\left(\frac{1+\alpha}{2}\right) T_{ijk} \qquad (5.11)$$

Hence, since $\Gamma^{\alpha}_{ijk}(\theta) \equiv 0$ for $\alpha = -1$, this mixture family of distributions constitutes an uncurved space (or -1-flat space) with respect to -1-connection called Dawid connection or mixture connection [18].

Remarks 18: (1) Each of these affine connections ∇^{α} on (S, g) or on Θ gives rise to the corresponding geometry including curvature R^{α}_{ijkl}.

(2) For the exponential family above $P_{\textcircled{H}}$ the α-curvature and metric curvature are related by

$$R^{\alpha}_{ijkl} = (1-\alpha^2)R_{ijkl} \qquad (5.12)$$

and so the ± 1-connections render the parameter space $\textcircled{H}$ "flat".
For the geometry of important families of statistical distributions and for curvature computations and interrelation see Burbea [14].
(3) The skewness tensor $T = (T_{ijk})$ given by (5.1) gives rich statistical structure on (S, g) or $\textcircled{H}$ The expectation meaning of (5.1) is that it measures the expectation of the third order cummulants of the variation of the log-likelihood function on χ. We discuss this tensor later.
(4) Thus the statistical manifold (S, g) or $\textcircled{H}$ has a 1-parameter family of affine connection $\{\nabla^{\alpha}\}$.

5 Dual Connections and dual flatness

Let $(S, g =<>)$ be a Statistical manifold.
Definition 19: Let $\nabla = (\Gamma_{ijk})$ and $\nabla^* = (\Gamma^*_{ijk})$ be two affine connections on S. We say they are dual to each other w.r.t. g if

$$X^k \partial_k \langle Y, Z\rangle = \langle \nabla_X Y, Z\rangle + \langle Y, \nabla_X Z\rangle \qquad (6.1)$$

holds for $\forall$ vector fields X, Y, Z on S.
Note that in local coordinate vector fields $X = \partial_i$, $Y = \partial_j$, $Z = \partial_k$ (6.1) becomes

$$\partial_i g_{jk} = \Gamma_{ijk} + \Gamma_{ikj} \qquad (6.2)$$

Remarks 21: The α-connection and $-\alpha$-connection on (S, g) are dual.
Proof: We know

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

Since T is completely symmetric we set

$$\Gamma^{\alpha}_{ijk} + \Gamma^{(-\alpha)}_{ikj} = \partial_i g_{jk}$$

and hence by (6.2), $\Gamma^{(\alpha)}$ and $\Gamma^{(-\alpha)}$ are dual w.r.t. the Fisher metric on (S, g).
Corollary 22: The Efron and Dawid connections are dual w.r.t. Fisher metric.
Remarks 23: α-flatness implies $-\alpha$-flatness. Infact, more generally, we have $R^{(\alpha)}_{ijkl} = -R^{(-\alpha)}_{ijlk}$
Definition 24: By a dual structure on a manifold M we mean a triple (g, ∇, ∇^*) consisting of a Riemannaian metric g and a pair of affine connections which are dual w.r.t. g. Parametric statistical models $\{P_{\textcircled{H}},\ g,\ \nabla^{(\alpha)},\ \nabla^{(-\alpha)}\}$ are having dual structure.

A statistical manifold is defined abstractly as a Riemannian manifold M endowed with a dual structure i.e. a quadruple (M, g, ∇, ∇^*).
Any parametric statistical model with Fisher metric g and dual connection $\nabla^{(\alpha)}, \nabla^{(-\alpha)}$ is a statistical manifold.

Definition 25: A manifold is flat w.r.t. an affine connection ∇ if there exists a local coordinate system such that $\Gamma^i_{jk} = 0$.
Then the torsion T and the curvature R of flat ∇ above vanish; conversely, if ∇ is an affine connection with vanishing curvature and torsion on a manifold M then M is " locally" flat.
Definition 26: A statistical manifold (M, g, ∇, ∇^*) is dually flat if both the dual affine connections are flat i.e. $T = T^* = 0$ and $R = R^* = 0$.
Remarks 27: (1) When a manifold is flat w.r.t. an affine connection ∇ then it is also flat w.r.t. its dual connection ∇^*.
(2) If (S, g, ∇, ∇^*) is a dually flat manifold then in any local coordinate system the metric coefficients g_{ij} are constant. (use $\partial_k \partial_{ij} = \Gamma_{ijk} + \Gamma^*_{ikj}$).
Definition 28: On a dually flat manifold there exist two special coordinate systems namely the affine flat coordinates for each of the connections. These coordinate systems are related to one another by a duality relation of their own. They are called the dual coordinate systems.
More precisely, two coordinate system (θ^α) and $(\tilde{\theta}_{\tilde{\nu}})$ are said to be dual to one another when their coordinate basis vectors satisfy: $< \hat{e}_\mu, \tilde{e}^{\tilde{\nu}} >= \delta^{\tilde{\nu}}_\mu$ where $\hat{e}_\mu$ and $\tilde{e}^{\tilde{\nu}}$ are the coordinate basis vectors for the θ or $\tilde{\theta}$ systems respectively.
Remark 29: In general manifolds there is no guarantee that a pair of dual coordinate systems exists. Thus dually flat structure is special feature for a statistical manifold.

6.1. Relation of the metrics for the dual coordinate systems (θ), $(\tilde{\theta})$

Write $\theta = \theta(\tilde{\theta})$ and $\tilde{\theta} = \tilde{\theta}(\theta)$. Then the coordinate basis vectors are related by $\hat{e}_\mu = \frac{\partial \tilde{\theta}_{\tilde{\nu}} \tilde{e}^{\tilde{\nu}}}{\partial \theta^\mu}$, $\tilde{e}^{\tilde{\mu}} = \frac{\partial \theta^\nu}{\partial \tilde{\theta}_{\tilde{\mu}}} \hat{e}_\nu$. Using these relations we may write the metrics as

$$(6.3) \qquad \begin{aligned} g_{\mu\nu} \equiv& \langle \hat{e}_\mu, \hat{e}_\nu \rangle = \frac{\partial \tilde{\theta}_{\tilde{\nu}}}{\partial \theta^\mu} \langle \tilde{e}^{\tilde{\nu}}, \hat{e}_\nu \rangle \\ =& \frac{\partial \tilde{\theta}_{\tilde{\nu}}}{\partial \theta^\mu} \delta^{\tilde{\nu}}_\nu \end{aligned}$$

and

$$(6.4) \qquad \begin{aligned} \tilde{g}^{\tilde{\mu}\tilde{\nu}} \equiv& \langle \tilde{e}^{\tilde{\mu}}, \tilde{e}^{\tilde{\nu}} \rangle \\ =& \frac{\partial \theta^\mu}{\partial \tilde{\theta}_{\tilde{\mu}}} \langle \hat{e}_\mu, \tilde{e}^{\tilde{\nu}} \rangle = \frac{\partial \theta^\mu}{\partial \tilde{\theta}_{\tilde{\nu}}} \delta^{\tilde{\mu}}_\mu \end{aligned}$$

Noting that the Jacobians $[J_{\mu\tilde{\nu}}] \equiv \left[\frac{\partial \tilde{\theta}}{\partial \theta^\mu}\right]$ and $[J^{\tilde{\mu}\nu}] \equiv \left[\frac{\partial \tilde{\nu}}{\partial \tilde{\theta}_{\tilde{\mu}}}\right]$ are each matrix inverse of the other, we find that $[\tilde{g}_{\tilde{\nu}\tilde{\mu}}]$ and $[g_{\mu\nu}]$ are also inverses of each other. Since the matrix inverse of $[g_{\mu\nu}]$ is known to be the contravariant form of the metric, $[g^{\mu\nu}]$, we find that

$$\tilde{g}^{\tilde{\mu}\tilde{\nu}} = \delta^{\tilde{\mu}}_\mu \delta^{\tilde{\nu}}_\nu g^{\mu\nu} \quad \text{(relation for 2-tensors.)} \tag{6.5}$$

More generally for any tensor T expressed in θ-coordinates as $T^{\mu_1, \mu_2 \cdots \mu_m}_{\nu_1 \nu_2 \cdots \nu_n}$ may be reexpressed in $\tilde{\theta}$-coordinates as

$$T^{\tilde{\mu}_1 \tilde{\mu}_2 \cdots \tilde{\mu}_m}_{\tilde{\nu}_1 \tilde{\nu}_2 \cdots \tilde{\nu}_n} = \delta^{\tilde{\mu}_1}_{\mu_1} \delta^{\tilde{\mu}_x}_{\mu_2} \cdots \delta^{\tilde{\mu}_m}_{\mu_m} \delta^{\nu_1}_{\tilde{\nu}_1} \delta^{\nu_2}_{\tilde{\nu}_2} \cdots \delta^{\nu_n}_{\tilde{\nu}_n} T^{\mu_1 \cdot \mu_m}_{\nu_1 \cdots \nu_n}. \tag{6.6}$$

6.2: The functional form of $\theta = \theta(\tilde{\theta})$ and $\tilde{\theta} = \tilde{\theta}(\theta)$ can be determined from the following:
Theorem 30: When (θ^i) and $(\tilde{\theta}_i)$ are dual coordinate systems there exist potential functions $Ⓗ(\theta)$ and $\tilde{Ⓗ}(\tilde{\theta})$ such that

$$\theta^i = \tilde{\partial}^i \tilde{Ⓗ}(\tilde{\theta}) \quad \text{and} \quad \tilde{\theta}_i = \partial_i Ⓗ(\theta) \tag{6.7}$$

Then it follows that $g_{ij} = \partial_i \partial_j Ⓗ(\theta)$
and
$\tilde{g}^{ij} = \tilde{\partial}^i \tilde{\partial}^j \tilde{Ⓗ}(\tilde{\theta})$ (6.8)
and also $Ⓗ(\theta) + \tilde{Ⓗ}(\tilde{\theta}) = \theta^i \tilde{\theta}_i$ (6.9)
Conversely, when a potential function $Ⓗ(\theta)$ exists such that $g_{ij} = \partial_i \partial_j Ⓗ(\theta)$,(6.7) yields a coordinate system $(\tilde{\theta}_i)$ which will be dual to (θ^i) and (6.9) yields to derive the other potential function $\tilde{Ⓗ}(\tilde{\theta})$.

Proof: Symmetry of the metric $g_{ij} = \partial_i \tilde{\theta}_j(\theta)$ show that $\partial_i \tilde{\theta}_j - \partial_j \tilde{\theta}_i = 0$ from which we conclude that, at least locally, $\tilde{\theta}(\theta)$ is the derivative of some function i.e., there exist a function $Ⓗ(\theta)$ such that $\tilde{\theta}_i = \partial_i Ⓗ$ (6.8) follows directly from inserting (6.7) into (6.3) and (6.4). Finally (6.9) is a general fact about Legendre transform.
The other direction is easy: when $g_{ij} = \partial_i \partial_j Ⓗ(\theta)$, we see that $(\tilde{\theta}_i) \equiv (\partial_i Ⓗ(\theta))$ is dual to (θ) from the fact that

$$\tilde{e}^{\nu} = \left(\left[\frac{\partial \tilde{\theta}}{\partial \theta} \right]^{-1} \right)^{ji} \hat{e}_i = \left([g_{\mu\nu}]^{-1} \right)^{ij} \hat{e}_{\mu}$$

where $< \hat{e}^j, \hat{e}_i >= \delta^{\nu}_{\mu}$, proving the duality. On a dually flat manifold there exists a pair of dual coordinate systems as in the following:
Theorem 31: When a manifold M is flat with respect to a dual pair of torsion-free connection ∇ and ∇^*, there exists a pair of dual coordinate systems (θ^i) and $(\tilde{\theta}_i)$ such that (θ^i) is ∇-affine and $(\tilde{\theta}_i)$ is ∇^*-affine.
Proof: ∇-flatness allows us to introduce a coordinate system (θ^i) in which $\Gamma_{ijk} = 0$. According to (6.2) this means over $\Gamma^*_{ijk} = -\partial_i g_{jk}$ in θ-coordinates. Since we assume that ∇^* is torsion-free, we have $\Gamma^*_{ijk} = \Gamma^*_{jik}$ and hence $\partial_i g_{jk} = \partial_j g_{ik}$ combining this with the fact that $g_{ij} = g_{ji}$ we may conclude that (again, at least locally) a potential function Ⓗ exists such that $g_{ij} = \partial_i \partial_j Ⓗ$ This allow us to introduce a coordinate system $(\tilde{\theta}_i)$ dual to (θ^i) defined by $\tilde{\theta}_j = \partial_j Ⓗ(\theta)$.

In order to show that $(\tilde{\theta}_i)$ is a ∇^*-affine coordinate system as claimed, we note that for any $i : \partial_i < \hat{e}_j, \tilde{e}^k >= \partial_i \delta^k_j = 0$, since $(\tilde{\theta}_i)$ is dual to (θ^i).

On the other hand, (6.1) shows that

$$\begin{aligned} \partial_i < \hat{e}_j, \tilde{e}^k >&= < \nabla_{\partial_i} \hat{e}_j, \tilde{e}^k > + < \hat{e}_j, \nabla^*_{\partial_i} \tilde{e}^k > \\ &= g^{kl} < \nabla_{\hat{e}_i} \hat{e}_j \, \hat{e}_l > + g_{jl} g_{im} < \tilde{e}^l, \nabla^*_{\tilde{e}^k}, \tilde{e}^m > \\ &= g^{kl} \Gamma_{ijl} + g_{jl} g_{im} (\Gamma^*)^{mkl} \end{aligned}$$

where $(\Gamma^*)^{mkl}$ is the connection coefficients corresponding to ∇^*.
Since both the left hand side and the first term in the right are zero, we conclude that $(\Gamma^*)^{mkl} = 0$ proving dual $(\tilde{\theta}_i)$ is a ∇^* affine coordinate system.
Remark 32: (1) The dual flat structure constitutes a fundamental mathematical concept of information geometry.
(2) Abstractly a dually flat manifold is a smooth Riemannian manifold (M, g) equipped with a pair of flat torsion-free affine connections ∇ and ∇^* which are dual to each other in the sense that forall vectors holds X, Y, Z on M.

$$Xg(Y, Z) = g(\nabla_X Y, Z) + g(Y, \nabla^*_X Z).$$

There are global topological obstructions for the existence of dually flat structure on a statistical manifold M.
Theorem A: (Ay [8].) Let (M, g, ∇, ∇^*) be a dually flat manifold. If M is compact then necessarily the first fundametnal group $\pi_1(M)$ of M has infinite order.
Consequently compact Riemannian manifolds with trivial or finite fundamental group never admit dually flat structures.
(3) In a compact manifold M with a complete affine connection ∇ it is false that any two points of M could be joined by a ∇-geodesic. However, dually flat manifolds enjoy this "geodesic property" namely.
Theorem B: Let (M, g, ∇, ∇^*) be a dually flat manifold. If one of the two connections on M is complete say ∇ then any two points in M can be joined by a ∇-geodesic. Further the topological classification of such dually flat manifolds can be obtained from a corresponding structure theorem due to Hicks [21] as in
Theorem C: Let (M, g, ∇, ∇^*) be a dually flat manifold of dimension m. Assume say ∇ is complete. Then there exists a connection-preserving diffeomorphsm $\Phi : (M, \nabla) \to (\mathbb{R}^m/\Gamma, \nabla_\Gamma)$ where $\Gamma \cong \pi_1(M)$ is a sub group of the group $\mathbb{R}^m \propto GL(m, \mathbb{R})$ of affine motions of $\mathbb{R}^m$ which acts freely and properly discontinuously in $\mathbb{R}^m$, and where ∇_Γ denotes the connection on $\mathbb{R}^m/\Gamma$ induced from the canonical flat affine connection on $\mathbb{R}^m$.
As a consequence, such dually flat manifolds M have their universal covering space $\tilde{M}$ diffeomorphitic to $\mathbb{R}^m$ and its first fundamental group $\pi_1(M)$ is isomorphic to a subgroup of the group of affine motions of $\mathbb{R}^m$ and hence higher homotopy groups $\pi_i(M)$, $2 \leq i \leq m$ necessarily vanish.

6 Generalizations of Amari's α-structures

Statistical manifolds are geometrical abstractions of statistical models (parametric or not). One way of defining statistical manifolds are Riemannian manifolds endowed with a pair of torsion-free dual connections (Lauritzen [22]).

Secondly, a divergence or contrast function gives rise to a statistical structure as its Riemannian metric is given by 2nd order derivatives of the divergence and a pair of dual connections by its 3rd order derivatives.

Thirdly, a statistical structure can be realized by starting with a skewness tensor C on a Riemannian manifold.
7.1 Divergence approach:

Firstly note that on a manifold with dual connections we can define the divergence between two points

$$D(p,q) = Ⓗ(\theta_p) + \tilde{Ⓗ}(\tilde{\theta}_q) - \theta_p^i \tilde{\theta}_{qi}$$

then D has the properties: a) $D(p,q) \geq 0$ and is 0 iff $p = q$
b) $\frac{\partial}{\partial \theta_p^i} D(p,q)|_{p=q} = \frac{\partial}{\partial \theta_q^i} D(p,q)|_{p=q} = 0$.
c) $\frac{\partial}{\partial \theta_p^i} \frac{\partial}{\partial \theta_p^j} D(p,q)|_{p=q} = g_{ij}(p)$
d) For three points p, q, r given, we have $D(p,r) \gtreqless D(p,q) + D(q,r)$ according as the angle φ between the tangent vectors at q of the ∇-geodesic joining p and q and the ∇^*-geodesic joinging q and r at point q is $\gtreqless \pi/2$.

Conversely starting with a divergence D we have seen earlier how its 2nd derivatives defines a Riemannian metric g as $g_{ij}(\theta) = \partial_{\theta^i} \partial_{\theta'^j} D(\theta, \theta')|_{\theta=\theta'}$ In fact, more geometry can be gotten from D.

Definition 33: Let S be an n-dimensional statistical model with global coordinate system $\theta = (\theta^i)_{i=1}^n$. Define a divergence function $D(\cdot,\cdot) : S \times S \to \mathbb{R}$ is a smooth function satisfying:
a) $D(p,q) \geq 0$ for $\forall\, p, q \in S$ and equality iff $p = q$.
b) $\partial_i D(p,q)|_{p=q} = 0$ and $\partial_i' D(p,q)|_{p=q} = 0$.
c) $\partial_i \partial_j' D(p,q)|_{p=q}$ is negative definite.
where $\partial_i = \frac{\partial}{\partial \theta^i}$ and $\partial_i' = \frac{\partial}{\partial \theta'^i}$.
Such D defines a unique Riemannian metric g^D and the affine connections ∇^D and ∇^{*D} (Eguchi [20])

$$g_{ij}^D(\theta) =< \theta_i, \theta_j >_\theta^D = -\partial_i \partial_j' D(p,q)|_{p=q} \tag{7.1}$$

and

$$\Gamma_{ij,k}^D(\theta) =< \nabla_{\partial_i}^D \partial_j, \partial_k >_\theta^D = -\partial_i \partial_j \partial_k' D(p,q)|_{p=q} \tag{7.2}$$

and its dual divergence D^* by $D^*(p,q) = D(q,p)$ and then $g^{D^*} = g^D$ and $\Gamma_{ijk}^{D^*} = -\partial_{i'} \partial_{j'} \partial_k D(p,q)|_{p=q}$ is the dual affine connection defined by D. ∇^D and ∇^{D^*} are dual w.r.t g^D.
Hence a divergence function D induces a dualistic structure $(g^D, \nabla^D, \nabla^{D^*})$ uniquely on a statistical model.

Conversely, Matumoto [23] proved that every torsion-free dualistic structure is induced from a globally defined divergence though not a unique one.

7.2 Skewness Tensor approach:

A divergence D gives rise to a Riemannian metric g^D and a pair of dual connections $\nabla^{(D)}, \nabla^{(D^*)}$ and also generates the skewness tensor C defined by

$$\begin{aligned} C^{(D)}(X,Y,Z) &= g(\nabla_X^{(D^*)} Y - \nabla_X^{(D)} Y, Z) \\ &= (X_{(\xi_1)} Y_{(\xi_1)} Z_{(\xi_2)} - X_{(\xi_2)} Y_{(\xi_2)} Z_{(\xi_1)}) D(\xi_1, \xi_2)|_{\xi_1=\xi_2} \end{aligned} \tag{7.4}$$

(here $p = p_{\xi_1}$, $q = q_{\xi_2}$), X, Y, Z vector fields on S.
In local coordinates this becomes

$$\begin{aligned} C_{ijk}^{(D)} &= \Gamma_{ij,k}^{(D^*)} - \Gamma_{ij,k}^{(D)} = \partial_{\xi_1^i} \partial_{\xi_1^j} \partial_{\xi_2^k} D(\xi_1, \xi)|_{\xi_1=\xi_2} \\ &\quad - \partial_{\xi_2^i} \partial_{\xi_2^j} \partial_{\xi_1^k} D(\xi_1, \xi_2)|_{\xi_1=\xi_2}. \end{aligned} \tag{7.5}$$

Then C is a 3-covariant completely symmetric tensor called the Skewness tensor arisen from Divergence D.

Let D_S be a contrast function on manifold S. Let $p, q \in S$ with coordinates ξ_1, ξ_2, respectively. Denote $\Delta\xi^i = \xi_2^i - \xi_1^i$. Then the third order approximation of $D_S(p, q)$ about p is given by

$$(7.6) \quad \begin{aligned} D_S(p,q) =& D_S(p,p) + \partial_{\xi_2^i} D(\xi_1,\xi_2)|_{\xi_1=\xi_2=\xi}\Delta\xi^i \\ &+ \frac{1}{2} D_{\xi_2^i}\partial_{\xi_2^j} D(\xi_1,\xi_2)|_{\xi_1=\xi_2=\xi}\Delta\xi^i\Delta\xi^j \\ &+ \frac{1}{6}\partial_{\xi_2^i}\partial_{\xi_2^j}\partial_{\xi_2^k} D(\xi_1,\xi_2)|_{\xi_1=\xi_2=\xi}\Delta\xi^i\Delta\xi^j\Delta\xi^k + o(||\Delta\xi||^2) \end{aligned}$$

where $o(\|\Delta\xi\|^2)$ is a term which converges to zero faster than $\|\Delta\xi\|^2$ as $p \to q$. The first two terms are zero by definition of contrast function and so.

$$(7.7) \quad D_S(p,q) = \frac{1}{2} g_{ij}(\xi_1)\Delta\xi^i\Delta\xi^j + \frac{1}{6} h_{ijk}(\xi_1)\Delta\xi^i\Delta\xi^j\Delta\xi^k + o(\|\Delta\|^2)$$

where g_{ij} is the induced Riemannian metric. Then

$$(7.8) \quad h_{ijk} = \Gamma^*_{ij,k} + \Gamma_{jk,i} + \Gamma^*_{ik,j} = \partial_j g_{ik} + \Gamma^*_{ik,j} = \partial_k g_{ij} + \Gamma^*_{ij,k}$$

Also we have $h_{ijk} = \partial_i g_{kj} + \Gamma^*_{jk,i}$ and hence h_{ijk} is totally symmetric in i, j, k i.e. $h_{ijk} = h_{ikj} = h_{kji} = h_{jik}$. Note that if D_S induces a dually flat statistical structure with $\Gamma = \Gamma^* = 0$ then $h_{ijk} = 0$.
Conversely given (g, ∇, ∇^*) a dualistic structure define a divergence by

$$(7.9) \quad D(p,q) = \frac{1}{2} g_{ij}(p)\Delta\xi^i\Delta\xi^j + \frac{1}{6} h_{ijk}(p)\Delta\xi^i\Delta\xi^j\Delta\xi^k$$

where $\Delta\xi^i = \xi^i(q) - \xi^i(p)$ and $h_{ijk} = \partial g_{kj} + \Gamma^*_{jk,i}$
which is only locally defined (not globally) or alternatively, defined by
(7.10) $D(p,q) = \frac{1}{2} g_{ij}(p)\Delta\xi^i\Delta\xi^j - \frac{1}{2} h^*_{ijk}(p)\Delta\xi^i\Delta\xi^j\Delta\xi^k$ with $h^*_{ijk} = \partial_i g_{jk} + \Gamma^*_{jk,i}$ locally. Note that there exists a global divergence D_S for given (g, ∇, ∇^*) dual structure (Matumoto [23]).
7.3 Special Skewness tensor Let S be a statistical manifold. Assume that on S there exists a local coordinate system w.r.t which the contrast function D_S is induced locally by a convex function φ by

$$(7.11) \quad \begin{aligned} D(\xi_o, \xi) =& \varphi(\xi) - \varphi(\xi_o) - \sum_j \partial_j \varphi(\xi)(\xi^j - \xi_o^j) \\ =& \varphi(\xi) - \varphi(\xi_o) - < \partial\varphi(\xi_o), \xi - \xi_o > \end{aligned}$$

where $\varphi : \Theta \subset \mathbb{R}^k \to \mathbb{R}$ is a strictly convex function. Such a divergence is called Bregman type divergence. Then the metric induced by this divergence is given by

$$(7.12) \qquad g_{ij}(\xi) = \partial_{\xi^i}\partial_{\xi^j}\varphi(\xi)$$

which is the Hessian of φ and the components of the induced dual connections $\nabla^{(D)}$ and $\nabla(D^*)$

$$(7.13) \quad \Gamma_{ijk}^{(D)} = 0 \quad \text{and} \quad \Gamma_{ijk}^{(D^*)} = \partial_{\xi^i}\partial_{\xi^j}\partial_{\xi^k}\varphi(\xi) \qquad (7.14)$$

and their Riemann curvature tensors are $R = R^* = 0$ i.e the connections are dually flat. Then the skewness tensor C defined by this divergence D_S is given by the third order derivative as

$$C_{ijk}^{(D)} = \partial_{\xi^i}\partial_{\xi^j}\partial_{\xi^k}\varphi(\xi) \tag{7.12}$$

Such special divergences induce on S what is called a Hessian geometry.

Remakrs 34: Such schematic divergence as in (7.11) occurs on dually flat statistical manifolds (S, g, ∇, ∇^*). Then there exist affine coordinates θ and η of the connections ∇ and ∇^* respectively and also there exist potential function $\psi(\theta)$ and $\varphi(\eta)$ corresponding to θ and η respectively. The dual coordinate η is the Legendre transform of the convex function $\psi(\theta)$ given by

$$\eta_i = \partial_i \psi(\theta) \tag{7.15}$$

In terms of θcoordinates.

$$g_{ij}(\theta) = \partial_i\partial_j\psi(\theta)\,; \quad \Gamma_{ijk}(\theta) = 0 \tag{7.16}$$

and

$$\Gamma^*_{ijk} = \partial_j\partial_j\partial_k\psi(\theta) \tag{7.17}$$

where $\partial_i = \frac{\partial}{\partial\theta^i}$.

The dual potential function $\varphi(\eta)$ is given by

$$\phi(\eta) = \max_{\theta}\{\theta\cdot\eta - \psi(\theta\} \tag{7.18}$$

and we have

$$\theta^i = \partial^i\varphi(\eta) \tag{7.19}$$

In dual coordinates η we have

$$\tilde{g}_{ij}(\eta) = \partial^i\partial^j\varphi(\eta)\,, \tilde{\Gamma}^*_{ijk}(\eta) = 0$$

and

$$\tilde{\Gamma}_{ijk}(\eta) = \partial^i\partial^j\partial^k\varphi(\eta) \tag{7.20}$$

Remark 35: Unlike dualistic manifolds S, the dually flat ones possess a canonical divergence defined in

$$D(p,q) = \psi(p) + \varphi(q) - \sum \theta^i(p)\eta_i(q) \tag{7.21}$$

and this one is unique.

Another equivalent way of defining a statistical structure:

Definition 36: A statistical structure on a manifold is triple (M, g, C) where C is a 3-covariant totally symmetric tensor on the Riemannian manifold (M, g) and C is called the skewness tensor.

7.4 Derivation of the usual (M, g, ∇, ∇^*) statistical manifold from (M, g, C).

Let $\nabla^{(0)}$ be the metric connection of g on M. given in the Koszul formula

$$\begin{aligned} (7.22) \quad 2g(\nabla^{(0)}_X Y, Z) =& X(g(Y,Z)) + Y(g(Z,X)) - Z(g(X,Y)) \\ &+ g([X,Y],Z) - g([Y,Z],X) - g([X,Z],Y) \end{aligned}$$

In local coordinates the connection components are then

$$(7.23) \qquad \Gamma^l_{ij} = \frac{1}{2} g^{kl}(\partial_{x^i} g_{kj} + \partial_{x^j} g_{ik} - \partial_{x^k} g_{ij})$$

which are the usual Christoffel symbols.

Then the geometric objects ∇ and ∇^* are defined by

$$\left.\begin{aligned} & g(\nabla_X Y, Z) = g(\nabla^{(0)}_X Y, Z) - \tfrac{1}{2} C(X,Y,Z) \\ \text{and} \quad & g(\nabla^*_X Y, Z) = g(\nabla^{(0)}_X Y, Z) + \tfrac{1}{2} C(X,Y,Z) \end{aligned}\right\} \quad (7.24)$$

are torsion-free dual connections and ∇, ∇^* and C are related by

$$(7.5) \quad g(\nabla^*_X, Y, Z) = g(\nabla_X Y, Z) + C(X,Y,Z)$$

Also Amari's α-connection $\nabla^{(\alpha)}$ is given by

$$(7.26) \quad g(\nabla^{(\alpha)}_X Y, Z) = g(\nabla^{(0)}_X Y, Z) + \frac{\alpha}{2} C(X,Y,Z)$$

and also

$$(7.27) \quad \nabla g = C, \qquad \nabla^* g = -C \quad \text{and} \quad \nabla^{(\alpha)} g - \alpha C$$

7 α-geometry and its generalization

Let $S = \{p_\theta / \theta \in \Theta^{\text{open}} \subset \mathbb{R}^n\}$ be a n-dimensional statistical model on sample space χ. Let $(\theta^i)^n_{i=1}$ be global coordinate system on S. Then the tangent space $T_\theta(S)$ to S at θ (i.e. at p_θ) is given by

$$T_\theta(S) = \left\{ \sum_{i=1}^n \alpha^i \partial_i | \alpha_i \in \mathbb{R} \right\} \tag{8.1}$$

with $\partial_l = \frac{\partial}{\partial \theta^i}$. Then $l = \log : S \to \mathbb{R}$ sending $p \in S$ to $l(p) = \log p(x, \theta) \in \mathbb{R}$, $x \in \chi$, $\theta \in \Theta \subset \mathbb{R}^n$ is an embedding map. Equivalently $l = \log p(x,\theta) : \mathbb{R}^+ \to \mathbb{R}$ is monotone function.

More generally for each $\alpha \in \mathbb{R}$, define $l_\alpha : \mathbb{R}^+ \to \mathbb{R}$ by

$$u \to l_\alpha(u) = \frac{2}{1-\alpha} u^{\frac{1-\alpha}{2}} \ , \ \alpha \neq 1 \, ; = \log u \quad \text{for} \quad \alpha = 1 \tag{8.2}$$

Note that this l_α is of class C^2 and its derivative is nonzero on $(0,\infty)$ and interpreting l_α on S by

$$\begin{aligned} l_\alpha(p) =& \frac{2}{1-\alpha} p^{\frac{1-\alpha}{2}} \quad (\alpha \neq 1) \\ &= \log p (\alpha = 1) \qquad\qquad (8.3) \end{aligned}$$

gives an embedding of S into the infinite dimensional vector space $\mathbb{R}_\chi = \{X : \chi \to \mathbb{R}$ "nice" function called Random variable$\}$ of random variables on χ, for each α in $\mathbb{R}$. Thus we have a α-parameter family $\{l_\alpha\}_{\alpha\in\mathbb{R}}$ of embeddings of S into $\mathbb{R}_\chi$. Each $l_\alpha = l_\alpha(p(x,\theta))$ is called the α-representation of the density function $p(x,\theta)$. For $\alpha = 0$, $l_o(p) = (p)^{1/2}$; For $\alpha = 1$, $l_1(p) = \log p$ and for $\alpha = -1$, $l_{-1}(p) = p$ are special representations. We write simply $l_\alpha(x,\theta)$ for $l_\alpha(p(x,\theta))$.
We know the set of scores $\{\partial_i l(x,\theta)\}_{i=1}^n$ is linearly independent by assumption and also $\{\partial_i l_\alpha(x,\theta)\}_{i=1}^n$ is a linearly independent set for each $\alpha \in \mathbb{R}$.
Define an n-dimensional vector space spanned by the scores (or generalized scores) as in

$$T_\theta^1(S) = \{A(x) | A(x) = \sum_{i=1}^n A^i \partial l(x,\theta)\}$$

(or more generally,

$$T_\theta^\alpha(S) = \{A_\alpha(x) | A_\alpha(x) = \sum_{i=1}^n A^i \partial_i l_\alpha(x,\theta) | A^i \in \mathbb{R}\}$$

which are called the 1-representation (respectively α-representation) of the tangent space $T_\theta(S)$. Infact the map: $\partial_i \to \partial_i l$ or $\partial_i l_\alpha (i = 1, 2 \cdot n)$ gives a natural isomorphism between $T_\theta(S)$ and $T_\theta^1(S)$ (or $T_\theta^\alpha(S)$).
Note that under this isomorphism the α-representation of a tangent vector $A = \sum A^i \partial_i \in T_\theta(S)$ is the random variable $A(X) = \sum A^i \partial_i l(x,\theta) \in T_\theta^1(S)$ for $\alpha = 1$ (respectively $A_\alpha(x) = \sum_{i=1}^n A^i \partial_i l_\alpha(x,\theta) \in T_\theta^\alpha(S)$. The passage among them is given by the formulae

$$(8.4) \qquad \partial_i l_\alpha = p^{\frac{1-\alpha}{2}} \partial_i l$$

and

$$(8.5) \quad \partial_i \partial_j l_\alpha = p^{\frac{1-\alpha}{2}} \{\partial_i \partial_j l + \frac{1-\alpha}{2} \partial_i l \partial_j l\}$$

Remark 37: Note that for each $A(x) \in T_\theta^1(S)$, $E_\theta[A(x)] = 0$ since $E_\theta[\partial_i l(x,\theta)] = 0$, $i = 1, 2, \cdots, n$. Infact, this give a characterization of $T_\theta^1(S)$ as $T_\theta^1(S) = \{X :$ random variable on χ s.t. $E_\theta(X) = 0\}$. This expectation w.r.t.$p(x,\theta)$ defined by $E_\theta(f) = \int_\chi f(x)p(x,\theta)dx$ defines an inner product on $T_\theta(S)$ (or on $T_\theta^1(S)$, $T_\theta^\alpha(S)$ or even on vector space $\mathbb{R}_\chi$) in a natural way: Let $A, B \in T_\theta(S)$ with corresponding r.v.s $A(x)$, $B(x) \in T_\theta^1(S)$ (respectively $A_\alpha(x), B_\alpha(x) \in T_\theta^\alpha$). Then the inner product $g =<,>$ is defined by

$$g(A,B)(\theta) =< A, B >_\theta = E_\theta[A(x)B(x)] = \int_\chi A(x)B(x)p(x,\theta)dx$$

$$(8.6)\ (\text{respectively} g(A_\alpha, B_\alpha) = < A_\alpha, B_\alpha >_\theta = E_\theta[A_\alpha(x)B_\alpha(x)] \\ = \int_\chi A_\alpha(x)B_\alpha(x)\{p(x,\theta)\}^\alpha dx \quad (\alpha\text{-expectation})$$

In particular on basis tangent vectors ∂_i and ∂_j

$$g_{ij}(\theta) = < \partial_i, \partial_j >_\theta = E_\theta[\partial_i l(x,\theta)\partial_j l(x,\theta)] \tag{8.7}$$

$$\text{respectively} \qquad g^\alpha_{ij}(\theta) = < \partial_1, \partial_j)^\alpha_\theta = \int_\chi \partial_i l_\alpha \partial_j l_\alpha p^\alpha dx \tag{8.8}$$

Remarks 38: (1) Note that g_{ij} in (8.7) is the Fisher metric given earlier.
(2) using formulae (8.5), $g^\alpha_{ij}(\theta) = \int_\chi \partial_i l_\alpha \partial_j l_{-\alpha} dx = \int_\chi \partial_i l \partial_j l p dx = g_{ij}(\theta)$ for all α.
(3) Hence each member of Amari's α-embedding $\{l_\alpha\}_{\alpha\in\mathbb{R}}$ induces the Fisher metric only on S.
(4) As before the n^3 Christoffel functions Γ^1_{ijk} defined by

$$\Gamma^1_{ijk}(\theta) = E_\theta[\partial_i\partial_j l(x,\theta)\partial_k l(x,\theta)] \tag{8.9}$$

giving the unique affine connection ∇^1 on S called the Efron (or exponential) or 1- connection by

$$\Gamma^1_{ijk}(\theta) = < \nabla^1_{\partial_i}\partial_j, \partial_k >_\theta \tag{8.10}$$

(5) Using the α-representation of the density function define the n^3 functions Γ^α_{ijk} for each $\alpha \in \mathbb{R}$ by

$$\Gamma^\alpha_{ijk}(\theta) = \int \partial_i\partial_j l_\alpha(x,\theta)\partial_k l_{-\alpha}(x,\theta) p^\alpha dx$$

$$(8.11) \qquad = \int \left\{\partial_i\partial_j l(x,\theta) + \left(\frac{1-\alpha}{2}\right)\partial_i l \partial_j l\right\}\partial_k l p dx$$

these functions uniquely determine an affine connection ∇^α on S, called the Amari's α-connection as

$$\Gamma^\alpha_{ijk}(\theta) = < \nabla^\alpha_{\partial_i}\partial_j, \partial_k > \tag{8.12}$$

thus we get a one-parameter family $\{\nabla^\alpha\}_{\alpha\in\mathbb{R}}$ of affine connections on the statistical manifold S.
Note that for $\alpha = 0$, $\nabla^{(0)}$ is the Levi-civita connection and for $\alpha = -1$ it is Dawid's or mixture connection, $\alpha = 1$ it is Efron connection or exponential connection. The differential geometric study of $(S, g; \nabla^\alpha)$ is called the α-geometry of S.
(6) Amari conjectured that among all possible embeddings of S, the α-family is the only one which is invariant under (i) reparametrizations (ii) smooth 1-1 transformation of random variables [Amari-Nagaoka [3]. (8.13)
We answer this in next sections.

8 General F-geometry on a statistical manifold S:

We have noted for studying the geometry of families of important distributions of statistics Burbea and Rao introduced a weighted Fisher metric ds_f^2 where f is a positive C^2-function (atleast cont.). Using a general embedding F (instead of Amari's l_α) and taking a weight function G, a smooth positive function, we define a general (F, G)-geometry on a statistical manifold S.

Let $F : \mathbb{R}^+ \to \mathbb{R}$ be a function of class atleast C^2 and that $F'(u) \neq 0$ for each $u \in (0, \infty)$. Then F gives an embedding of S into $\mathbb{R}_\chi$ taking $p(x, \theta)$ into $F(p(x, \theta))$. Denote $F(p(x, \theta))$ simply by $F(x, \theta)$. Then

$$\partial_i F(x, \theta) = p(x, \theta) F'(p) \partial_i l(x, \theta) \tag{9.1}$$

Note that for every θ, the n functions $\{\partial_i F(x, \theta)\}_{i=1}^n$ are linearly independent functions of x since $\{\partial_i l\}_{i=1}^n$ are linearly independent functions of x.

Let $T_{F(p_\theta)}(F(S))$ be the n-dimensional vector space spanned by $\{\partial_i F\}_{i=1}^n$. Thus

$$T_{F(p_\theta)}(F(S)) = \{A^F(x) \in \mathbb{R}_\chi | A^F(x) = \sum_{i=1}^{n} A^i \partial_i F(x, \theta) : A^i \in \mathbb{R}\} \tag{9.2}$$

denoted simply by $T_\theta^F(S)$ which is naturally isomorphic with the tangent space $T_\theta(S)$ of S at p_θ by the map $\partial_i \to \partial_i F(x, \theta)$. The vector space $T_\theta^F(S)$ given by (9.2) is called the F-representation of the tangent space $T_\theta(S)$. Thus the F-representation of a tangent vector

$$A = \sum A^i \partial_i \in T_\theta(S) \text{ is the r.v. } A^F(x) = \sum_{i=1}^{n} A^i \partial_i F \in T_\theta^F(S) \tag{9.3}$$

Definition 39: Let $G : (0, \infty) \to \mathbb{R}$ be a positive smooth function (weight function) and let F be an embedding function. Then (following Burbea) the (F, G)-expectation of a r.v.f w.r.t the distribution $p(x, \theta)$ is defined as

$$E_\theta^{F,G}(f) = \int_\chi f(x) \frac{1}{p(F'(p))^2} G(p) dx \tag{9.4}$$

Using this, define an inner product on the vector space $\mathbb{R}_\chi$ of r-variables on χ by

$$< f, g >_\theta^{F,G} = E_\theta^{F,G}[f(x)g(x)] \tag{9.5}$$

which induces a Riemannian metric on S given by

$$< A.B >_\theta^{F,G} = E_\theta^{F,G}[A^F(x) B^F(x)] \quad \text{with } A, B \in T_\theta(S) \text{ using } (9.3) \tag{9.6}$$

In terms of basis vectors we have

$$(9.7) \qquad < \partial_i, \partial_j >_\theta^{F,G} = \int_\chi \partial_i F \partial_j F \frac{G(p)}{p(F'(p))^2} ds$$
$$= \int_\chi \partial_i l \partial_j l G(p) p dx \quad \text{using}(9.1)$$

which is independent of the embedding F (just like the case of α-embedding above).
We call this metric as the G-metric on S with weight function G (similar to f-weighted Fisher metric of Burbea-Rao discussed earlier), denote it by $g^G =<,>^G$ and its components by $g_{ij}^G(\theta) =< \partial_i, \partial_j >_\theta^G = \int_\chi \partial_i l \partial_j l G(p) p dx$ and the matrix $(g_{ij}^G(\theta))$ is called the G-matrix. This is the most possible generalization of Fisher information metric and matrix.
Definition 40: Let $\pi_{|p_\theta}^{F,G} : \mathbb{R}_\chi \to T_\theta^F(S)$ be the projection map.
The affine connection induced by this map on S called the (F,G)-connection $\nabla^{F,G}$ is defined as

$$\nabla_{\partial_i}^{F,G} \partial_j = \pi_{|p_\theta}^{F,G} \left(\frac{\partial^2 F}{\partial \theta^i \partial \theta^j} \right) = \sum_n \sum_m g^{G(m,n)} \langle \frac{\partial^2 F}{\partial \theta^i \partial \theta^j}, \frac{\partial F}{\partial \theta^m} \rangle_\theta^{F,G} \partial_n \tag{9.8}$$

where $[g^{G(m,n)}(\theta)]$ is the inverse of the G-matrix $[g_{mn}^G(\theta)]$.
Note that the (F,G)-connections are symmetric i.e. torsion-free.
Then we can express this (F,G)-connection interms of score function as
(a)

$$\nabla_{\partial_i}^{F,G} \partial_j = \sum_n \sum_m g^{G(m,n)} E_\theta \left[\left\{ \partial_i \partial_j l + \left(1 + \frac{pF''(p)}{F'(p)} \right) \partial_i l \partial_j l \right\} \partial_m l G(p) \right] \partial_n \tag{9.9}$$

and (b)

$$\Gamma_{ijk}^{F,G}(\theta) = \int_\chi \left\{ \partial_i \partial_j l + \left(1 + \frac{pF''(p)}{F'(p)} \right) \partial_i l \partial_j l \right\} \partial_k l G(p) p \, dx \tag{9.10}$$

(connection coefficients) by direct computation.
Theorem 41: Let F and H be two embedding of S into $\mathbb{R}_\chi$ via such functions from $\mathbb{R}^+ \to \mathbb{R}$. Let G be a positive smooth weight function on $(0,\infty)$. Then the (F,G)-connection $\nabla^{F,G}$ and the (H,G)-connection $\nabla^{H,G}$ are dual w.r.t G-metric if an only if the functions F and G satisfy the relation

$$H'(p) = \frac{G(p)}{p} \frac{1}{F'(p)} \tag{9.11}$$

For details see [32].
Remarks 42: (1) we call such H if it exists, a G-dual embedding of F. In general for embedding F such dual H won't exist. Zhang [34] showed for strict convexity of such F, such dual H exists.
(2) The components of the dual connection $\nabla^{H,G}$ if it exists are given by

$$\begin{aligned} \Gamma_{ijk}^{H,G}(\theta) &= \int \left\{ \partial_i \partial_j l + \left(1 + \frac{pH''(p)}{H'(p)} \right) \partial_i l \partial_j l \right\} \partial_k l G(p) p dx \qquad (9.12) \\ &= \int \left\{ \partial_i \partial_j l + \left(\frac{pG'(p)}{G(p)} - \frac{pF''(p)}{F'(p)} \right) \partial_i l \partial_j l G(p) p \right\} dx \end{aligned}$$

This is a straight forward computation and using (9.11).
(3) The Amari α-geometry is a spacial case of our (F,G)-geometry on S. For, taking

$F(p) = l_\alpha(p)$ and $G(p) = 1$ constant function, we get $\Gamma^{F,G}_{ijk}(\theta) = \Gamma^{\alpha}_{ijk}(\theta)$ and $\Gamma^{H,G}_{ijk}(\theta) = \Gamma^{-\alpha}_{ijk}(\theta)$ and $g^G_{ij}(\theta) = \int \partial_i l \partial_j l p dx = g_{ij}(\theta)$.
(4) The Levi-civita connection ∇^G of the G-metric is an (F, G)-connection $\nabla^{F,G}$ with F given by

$$F'(p) = \sqrt{\frac{G(p)}{p}} \tag{9.13}$$

and for $G(p) = 1$, ∇^G reduces to Amari's 0-connection $\nabla^{(0)}$ with embedding function $F(p) = \sqrt{p}$ which is an integrated version of (9.13).
Example 43: Take $F(x) = x \ln x - x$, $G(x) = \ln(x)$. Then the G-dual embedding H of F is given by $H'(x) = \frac{G(x)}{xF'(x)} = \frac{\ln x}{x \ln x} = \frac{1}{x}$ and hence $H(x) = \ln x$.
Then the G-metric is given by

$$g^G_{ij}(\theta) = \int \partial_i l \partial_j l \ln p \, p dx \,;$$

and its H dual is

$$\begin{aligned}\Gamma^{H,G}_{ijk} =& \int \partial_i \partial_j l \partial_k l \ln p \, p dx. \\ =& E_\theta[\partial_i \partial_j l \partial_k l G(p)]\end{aligned}$$

which agrees with Burbea's formula for the f-geometry.

9 Invariance of (F, G)-geometry

For the stastical manifold $S = \{p(x, \theta) | \theta \in \Theta \subset \mathbb{R}^n\}$ the intrinsic goemetrical properties of S should be independent of the label θ of each point p. There are two kinds of invariance of the geometric structures namely (a) covariance under reparametrizations of the parameter space of S. (b) invariance under smooth 1-1 transformations of the random variable. Now we answer the Amari's questions in (8.13). We show that the α-geometry on S is the only invariant geometry among all the possible (F, G)-geometries.
Definition 44: Let (θ^i) and $\eta_j)$ be two coordinate systems on S which are related by an invertible transformation $\eta = \eta(\theta)$. Let the coordinate expressions for the metric g w.r.t (θ^i) and (η_j) be $g_{ij} =< \partial_i, \partial_j >$ and $\tilde{g}_{ij} =< \partial^i, \partial^j >$ where $\partial_i = \frac{\partial}{\partial \theta^i}$ and $\partial^j = \frac{\partial}{\partial \eta_j}$. Let the components of the connection ∇ w.r.t (θ^i) and (η_j) be given by $\Gamma_{ijk}, \tilde{\Gamma}_{ijk}$ respectively. Then the covariance under the reparametrization of the metric and the connection is defined as

$$\tilde{g}_{ij} = \sum_m \sum_n \frac{\partial \theta^m}{\partial \eta_i} \frac{\partial \theta^n}{\partial \eta_j} g_{mn} \tag{10.1}$$

and

$$\tilde{\Gamma}_{ijk} = \sum_{m,n,h} \frac{\partial \theta^m}{\partial \eta_i} \frac{\partial \theta^n}{\partial \eta_j} \frac{\partial \theta^h}{\partial \eta_k} \Gamma_{mnh} + \sum_{m,h} \frac{\partial \theta^h}{\partial \eta_k} \frac{\partial^2 \theta^m}{\partial \eta_i \partial \eta_j} g_{mh} \tag{10.2}$$

That is, the covariance under reparametrizations actually means the metric and the connection are coordinate independent.
Definition 45: Let $S = \{p(x,\theta)|\theta \in \Theta \subset \mathbb{R}^n\}$ be a statistical manifold defined on the sample space χ. Let x and y be random variable defined on sample space χ and $\mathcal{Y}$ respectively and φ be a smooth $1-1$ transformation of x to y. Assume this transformation induces a statistical model.
$\tilde{S} = \{q(y,\theta)|\theta \in \Theta \subset \mathbb{R}^n\}$ on $\mathcal{Y}$. Let $\lambda : S \to \tilde{S}$ be a diffeomorphism defined as $\lambda(p_\theta) = q_\theta$. Let $g, \tilde{g}$ be Riemannian metrics and $\nabla, \tilde{\nabla}$ be affine connection on S and $\tilde{S}$ respectively. Then the invariance under smooth 1-1 transformations of the random variable is defined as follows:

$$g(X,Y)_p = \tilde{g}(\lambda_*(X), \lambda_*(Y))_{\lambda(p)} \tag{10.3}$$

and

$$\lambda_*(\nabla_X Y) = \tilde{\nabla}_{\lambda_*(X)}\lambda_*(Y), \forall\ X, Y \in T_\theta(S) \tag{10.4}$$

where λ_* is the differential of map λ defined by

$$\lambda_*(X)_{\lambda(p)} = (d\lambda)_p(X)$$

Note that the Fisher information metric g and the α-connections are invariant under smooth 1-1 transformations of the random variable and also covariant under reparametrization (refer prop. 8(a) and (b) of §2).
More generally for (F,G)-geometries we have.
Theorem 46: The G-metric g^G is covariant under reparametrization.
proof: direct computation using above definition 44.
Theorem 47: The (F,G)-connection $\nabla^{F,G}$ is covariant under reparametrization.
Proof: analogous computation to α-connection of Amari [3].
Theorem 48: The (F,G)-geometric structure, the G-metric and the (F,G)-connection are *not* invariant under smooth 1-1 transformation of r.v in general.
Corollary 49: The only (F,G)-geometry which is invariant under smooth 1-1 transformations of the r.v. is the Amari's α-geometry.
Proof: Using Euler's homogenous function theorem, we get F' is a positive homogeneous function in p of degree k. Hence $F'(\lambda p) = \lambda^k F'(p) \forall \lambda > 0$. Without loss of generality take $F'(p) = p^k$. Hence $F(p) = \frac{p^{k+1}}{k+1}$, $k \neq -1$ and $F(p) = \log p$ for $k = -1$. Take $k = -(\frac{1+\alpha}{2})$, $\alpha \in \mathbb{R}$. Then $F(p) = \frac{2}{1-\alpha}\ p^{\frac{1-\alpha}{2}}$, $\alpha \neq 1$, and $= \log p$ for $\alpha = 1$, which is Amari's α-embedding function $l_\alpha(p)$, Taking $G(p) = k_1 = 1$ and $p\frac{F''(p)}{F'(p)} = k$. We get that the (F,G)-geometric structure is simply the α-geometry.
Remarks 50: (1) The fact that (F,G)-geometry on S is non-invariant under smooth 1-1 transformation of r.v.: is very natural as there will be information loss under such transformation unless the statistics is a good one which gives substance to the concept "sufficient statistic".
(2) The (F,G)-geometries play an important role in asymptotic estimation and inference [5].
(3) Among these the original Fisher metric g on S enjoys a special place. Infact, it is a deep result that the Fisher metric on S is unique w.r.t the two properties (a) g_{ij} is invariant

under reparametrization of the sample space χ (b) g_{ij} is covariant under reparametrization of the parameter space Θ
following the works of Cencov [16] and Corcuera and Giummolé [17].
Remarks 51: (1) The geometry of exponential families, deformed q-exponential families and χ-deformed families has rich applications to statistical physics [2,10,19,33].
(2) The geometries of statistical manifolds can be studied at a more general level namely the geometry of parametrized measure models and non-parametric measure models [6],[15],[25],[26],[11] and in [3].
(3) We will discuss in a sequel the geometry of Statistical manifold arising from various divergences, the relations among them and also to generate new canonical divergences in information geometry [31],[7].

10 Acknowledgements:

The first author thanks Prof. Raj Kishore for his constant encouragement in putting down these ideas as a research article. Further this paper is dedicated (on the occasion of his 70^{th} birth year) to our beloved (late) friend and colleague at BHU, Varanasi, (learning good mathematics under our most affectionate teacher Prof. V.L.N. Sarma), Prof. Vijay Kumar Patodi in fond memory of our pleasant stay together at TIFR, Bombay in December 1974 for a week. Some parts of this paper are presented as an invited talk of Bharat Ganita Parishad annual conference held at Department of Maths. & Astronomy of Lucknow University during November 21-22, 2015.

References

1. S. Amari Theory of Information Spaces: A geometrical foundation of statistics, POSTRAAG Report 106 (1980).

2. ___: Differential geometry of Curved exponential family, Curvature and information loss, Ann. of. Stat. 10 (1982) 357-385.

3. ___ and H. Nagaoka Methods of information geometry Transl. of Math. Monographs Vol 191 (AMS) 2000.

4. C. Atkinson and A.F. Mitchell: Rao's distance measure Sankhya 43 (1981) 345-364.

5. N. Ay, J.Jost, H.V. Le and L. Schwachhöfer information geometry and sufficient statistics, Prob. Theory and Related Topics 162(2015) 327-364.

6. ___,___,___,___: parametrized measure models arXiv. 25th Oct. 2015, 29 pages.

7. ___ and S. Amari: A novel approach to canonical divergences within information geometry, entropy 17 (2015) 8111-8129.

8. ___ and W. Tuschmann: Dually flat manifolds and global information geometry: Open Systems & information Dynamics 9 (2002) 195-200.

9. O.E. Barndorff-Nielsen and P.E. Jupp: Yokes and symplectic Structures Jl. of Stat. planning and inference 63 (1997) 133-146.

10. ___,___ Approximating the exponential models, Ann. Inst of Stat. Math. 41 (1989) 247-267.

11. M. Bauer, M. Bruveris, P. Michor: Uniqueness of the Fisher-Rao metric in the space of smooth densities arXiv 1411, 5577 (2014).

12. Burbea, J. and C.R. Rao Entropy differential metric, distance and divergence measures in probablity spaces: a Unified approach, Jl. Multivariate Analysis 12 (1982) 575-596.

13. ___,___: Differential metric in probability spaces prob. and Math. Stat. 3 (1982) 115-132.

14. Burbea, J: information geometry in probability spaces Expositions of Math. 4 (1986) 347-378.

15. A. Cena and G. Pistone: Exponential Statistical Models Ann. Inst. Stat. Math 59(2007) 27-56.

16. N. Cencov: statistical decision rules and optimal inference Transl. of Math. Monographs 53 (AMS) 1981.

17. I.M. Corcuera and F. Giummole: A characterization of monotone and regular divergences, Ann. Inst. Stat. Math 50 (1998) 433-450.

18. B. Efron Defining the curvature of a statistical problem with application to second order efficiency (with discussion) Ann. of Stat. 3(1975) 1189-1242 (for A.P. Dawid, see vol 3 (1975) p 1231-34, also vol 5(1977) p 1249).

19. ___: The geometry of exponential family. Ann. of Stat. 6 (1978) 362-376.

20. S. Eguchi: Second order efficiency of minimum contrast estimators in a curved exponential family Ann. of Stat. 11(1983) 793-803.

21. N. Hicks A theorem on affine connection Ill. Jl. of Math. 3 (1959) 242-254.

22. S. Lauritzen Statistical manifolds in diffl. geometry in statistical inferences edit by S. Amari, O.E.B. Nielsen, R.E. Kass, S.L. Lauritzen and C.R. Rao: IMS lecture notes: Monograph series 10, Inst. of Math. Stat. Hayward, California 1987, 163-216.

23. T. Matumoto: Any Stat. manifoled has a contrast function on the C^3-functions taking the minimum at the diagonal of the product manifold Hiroshima Math. Jl. 23 (1993) 327-332.

24. M.K. Murrey and J.W. Rice: Differential Geometry and Statistics Chapman, 1993.

25. G. Pistone: Non parametric information geometry Geom. Science of Information, Lecture notes in computer Science 8085 (2013) springer, Heidelberg.

26. ___ amd C. Sempi: An infinite dimensional structure on the space of all the probability measures equivalent to a given one Ann. of Stat. 5(1995) 1543-1561.

27. C.R. Rao Information and accuracy attainable in the estimation of Statistical parameters Bull. Calcutta Math. Soc. 37 (1945) 81-91.

28. ___: on the distance between two populations Sankhya 9 (1949) 246-248.

29. J. Reeds Discussion on Prof. Efron's paper (1975) Ann. of Stat. 5 (1977) 1234-1238.

30. M. Sitaramayya Curvature tensors in Kahler manifolds Trans. Amer. Math. Soc. 163 (1973) 341-353.

31. ___, K.S. Subrahamanian Moosath, K.V. Harsha General geometry on Stat. Manifolds via) divergences survey article under prep. 2016.

32. K.V. Harsha, Subrahamanian Moosath, M.S.: F-geometry and Amari's α-geometry on a Statistical Manifold Entropy 16 (2014) 2472-2487.

33. ___,___: Dually flat geometries of the Deformed exponential family, Physica A: Stat. Mechanics and its applications 433 (2015) 136-147.

34. J. Zhang: Divergence function, Duality and Convex Analysis Neural Computation 16 (2004) 159-195.