Report of the UGC minor project entitled

Statistical Curvature and Second Order Efficiency

Principal Investigator: Dr. Vineetha K
Department of Mathematics
M G College
Trivandrum

Chapter 1

Geometry of Exponential Family

1 Introduction

Information geometry emerged from the geometric study of a statistical model of probability distributions. The information geometric tools are widely applied to various fields such as statistics, information theory, stochastic processes, neural networks, statistical physics, neuroscience etc. The importance of the differential geometric approach to the field of statistics was first noticed by C R Rao. On a statistical model of probability distributions he introduced a Riemannian metric defined by the Fisher information known as the Fisher information metric.

Another milestone in this area is the work of Amari. He introduced the α - geometric structures on a statistical manifold consisting of Fisher Information metric and the $\pm\alpha$ - connections. Harsha and Moosath introduced more generalized geometric structures called the (F, G) geometry on a statistical manifold which is a generalization of $\alpha-$ geometry. There are many attempts to understand the geometry of the statistical manifold and also to develop a differential geometric framework for the estimation theory.

In this chapter we will be mainly looking at the geometry of exponential family. An exponential family is an important statistical model which is attracted by many of the researchers from Physics, Mathematics and Statistics. The Exponential family contains as special cases most of the standard discrete and continuous distributions that we use for practical modelling, such as the normal, Poisson, Binomial, exponential, Gamma, multivariate normal, etc. Distributions in the Exponential family have been used in classical statistics for decades. We discuss the dually flat structure of the finite dimensional exponential family with respect to the $\pm 1-$ connections defined by Amari. Then the $q-$ exponential family relevant in the q - entropy maximum problem and its $q-$structure which is dually flat are considered. Finally a

glimpse of a more generalized notion of the exponential family the deformed exponential family and its dually flat structures are given.

2 Differential Geometry

Definition 1 *An $n-$ dimensional **topological manifold** M is a second countable, Hausdorff topological space which is locally Euclidean. That is , every point $p \in M$, there exist an open set $U \subset M$ and a homeomorphism $\phi : U \longrightarrow U'$, where U' is an open subset of $\mathbb{R}^{\kappa}$. (U, ϕ) is called a co-ordinate chart on M around p and ϕ is written as $\phi = (x^i); i = 1, ..., n$.*

If we have two charts (U, φ) and (V, ψ) on M such that $U \cap V \neq \varnothing$, the composite map $\psi \circ \varphi^{-1} : \varphi(U \cap V) \longrightarrow \psi(U \cap V)$ is called the **transition map**. The two charts (U, φ) and (V, ψ) are said to be smoothly compatible if either $U \cap V = \varnothing$ or the transition map $\psi \circ \varphi^{-1}$ is a diffeomorphism.

An **atlas** $\mathcal{A}$ for M is the collection of charts whose domain cover M and $\mathcal{A}$ is said to be a smooth atlas if any two charts in $\mathcal{A}$ are **smoothly compatible** with each other. $\mathcal{A}$ is a **maximal atlas** if any chart that is smoothly compatible with every other charts in $\mathcal{A}$ is already in $\mathcal{A}$. A **smooth structure** on any topological manifold is a maximal smooth atlas on M. A **smooth manifold** is a pair $(M, \mathcal{A})$ where M is a topological manifold and $\mathcal{A}$ is a smooth structure on M.

Definition 2 *Let M be a smooth manifold. A function $f : M \longrightarrow \mathbb{R}$ is said to be a **smooth map** if $f \circ \varphi^{-1}$ is smooth for some smooth chart (U, φ) around each point. The set of all smooth functions from M to $\mathbb{R}$ is denoted by $C^{\infty}(M)$.*

Example 1 *1. The Euclidean space $\mathbb{R}^n$ is a smooth n-dimensional manifold. $\mathcal{A} = (\mathbb{R}^n, Id_{\mathbb{R}^n})$ is a smooth atlas on $\mathbb{R}^n$.*

2. Consider the unit circle $S^1 = \{(x, y) \ / \ x^2 + y^2 = 1\}$
Let $U_i^+ = \{(x_1, x_2) \in S^1 \ / \ x_i > 0\}$ and $U_i^- = \{(x_1, x_2) \in S^1 \ / \ x_i < 0\}$ for $i = 1, 2$.
Define

$$\begin{aligned}
\phi_1^+ : (-1, 1) &\longrightarrow U_1^+ \quad ; \quad \phi_1^+(x) = (x, \sqrt{1 - x^2}) \\
\phi_{2^+} : (-1, 1) &\longrightarrow U_2^+ \quad ; \quad \phi_2^+(x) = (x, -\sqrt{1 - x^2}) \\
\phi_1^- : (-1, 1) &\longrightarrow U_1^- \quad ; \quad \phi_1^-(x) = (\sqrt{1 - x^2}, x) \\
\phi_2^- : (-1, 1) &\longrightarrow U_2^- \quad ; \quad \phi_2^-(x) = (-\sqrt{1 - x^2}, x)
\end{aligned}$$

$\mathcal{A} = \{(U_i^{\pm}, \phi^{\pm}), i = 1, 2\}$ is a smooth atlas on S^1. Hence S^1 is a 1-dimensional smooth manifold.

Definition 3 *A linear map $X : C^\infty(M) \longrightarrow \mathbb{R}$ is called a* ***derivation*** *at p if it satisfies the following*

$$X(fg) = f(p)Xg + g(p)Xf \tag{1}$$

for all $f, g \in C^\infty(M)$.

Let M be a smooth manifold and let $p \in M$. The **tangent space** to M at p, denoted by T_pM is defined as the set of all derivations of $C^\infty(M)$ at p.
Let $(U, \phi = (x^i))$ be a smooth chart on M around p. Then T_pM is a vector space of dimension n which is spanned by $\{\frac{\partial}{\partial x^i}|_p; i = 1, ..., n\}$. Each element in T_pM is called a **tangent vector** at p. Let T_p^*M denote the dual space of T_pM, i.e the set of all linear maps from T_pM to $\mathbb{R}$. T_p^*M is also an n-dimensional vector space which is spanned by $\{dx^i|_p; i = 1, ..., n\}$. ElementS of T_p^*M is called **cotangent vectors** at p.

A **tangent bundle** TM on M is the disjoint union of tangent spaces at all points of M.

$$TM = \bigcup_{p \in M} T_pM \tag{2}$$

A **cotangent bundle** T^*M on M is the disjoint union of cotangent spaces at all points of M

$$T^*M = \bigcup_{p \in M} T_p^*M \tag{3}$$

A **vector field** X on a smooth manifold M is a map $X : M \longrightarrow TM$, which associates to each point $p \in M$ a tangent vector $X_p \in T_pM$.
A **co-vector field** or a **1-form** Y on a smooth manifold M is a map $Y : M \longrightarrow T^*M$, which associates to each point $p \in M$ a cotangent vector $Y_p \in T_p^*M$.

Definition 4 *Let M be an $n-$dimensional smooth manifold. A* ***Riemannian metric*** *$g =<,>$ on M is a smooth symmetric 2-tensor field which is positive definite at each point. For every $p \in M$, $g_p =<,>_p : T_pM \times T_pM \longrightarrow \mathbb{R}$ satisfies the following conditions:*

1. $< aX+bY, Z >_p = a < X, Y >_p + b < Y, Z >_p \ \forall X, Y, Z \in T_pM, \ \forall a, b \in \mathbb{R}$
2. $< X, Y >_p =< Y, X >_p \ \forall X, Y \in T_pM$
3. *If $X \neq 0$ then, $< X, X >_p > 0$*

A **Riemannian manifold** is a manifold equipped with a Riemannian metric.

Definition 5 *Let M be an n-dimensional smooth manifold. Let $\Gamma(TM)$ denote the set of all smooth vector fields on M. A* ***linear*** *or an* ***Affine connection*** *on M is defined as a map $\nabla : \Gamma(TM) \times \Gamma(TM) \longrightarrow \Gamma(TM)$ which satisfies the following:*

1. $\nabla_X(Y + Z) = \nabla_X Y + \nabla_X Z$
2. $\nabla_{(X+Y)} Z = \nabla_X Z + \nabla_Y Z$
3. $\nabla_X(fY) = f\nabla_X Y + (XfY$
4. $\nabla_{fX} Y = f\nabla_X Y$

for all $f \in C^\infty(M)$ and $X, Y, Z \in \Gamma(TM)$.

Let $(U, \phi = (\xi^i))$ be a smooth chart in M. Then $\{\frac{\partial}{\partial \xi^i},\ i = 1, .., n\}$ are the coordinate vector fields. For the convenience, we denote $\frac{\partial}{\partial \xi^i}$ by ∂_i. The affine connection ∇ can be determined by n^3 functions Γ^k_{ij} given by

$$\nabla_{\partial_i} \partial_j = \Gamma^k_{ij} \partial_k$$

where Γ^k_{ij} are called the **Christoffel symbols** of the affine connection ∇ with respect to the coordinates $(\xi^i); i = 1, .., n$.
If ∇ is an affine connection on a Riemannian manifold M with a Riemannian metric $g =<, >$, then we have

$$< \nabla_{\partial_i} \partial_j, \partial_m >= \Gamma^k_{ij} < \partial_k, \partial_m >= \Gamma^k_{ij} g_{km} \tag{4}$$

where $g_{km} =< \partial_k, \partial_m >$.
It is often convenient to express the Christoffel symbols of the affine connection ∇ by

$$\Gamma_{ijm} = \Gamma^k_{ij} g_{km} =< \nabla_{\partial_i} \partial_j, \partial_m > \tag{5}$$

The n^3 functions Γ_{ijm} are called the **components** of the affine connection with respect to co-ordinate (ξ^i).

Definition 6 *Let M be Riemannian manifold with a Riemannian metric g. A connection ∇ is said to be* ***metric preserving*** *or* ***metric*** *if it satisfies the following,*

$$d(g(X, Y)) = g(\nabla X, Y) + g(X, \nabla Y) \tag{6}$$

Definition 7 *Let M be a Riemannian manifold with a Riemannian metric g. The two affine connections, ∇ and ∇^* on the tangent bundle are said to be* ***dual connections*** *if*

$$d(g(X, Y)) = g(\nabla X, Y) + g(X, \nabla^* Y) \tag{7}$$

holds for any two vector fields X, Y on M.

Let $\Gamma_{ijk} = < \nabla_{\partial_i}\partial_j, \partial_k >$, $\Gamma^*_{ijk} = < \nabla^*_{\partial_i}\partial_j, \partial_k >$ be the components of the dual affine connections. Then we have the following relation

$$\partial_i g_{jk} = \Gamma_{ijk} + \Gamma^*_{ikj} \tag{8}$$

This shows that every affine connection has a unique dual determined by

$$\Gamma^*_{ikj} = \partial_i g_{jk} - \Gamma_{ijk} \tag{9}$$

If ∇ is metric, then it is self dual.

Definition 8 *A connection is said to be* ***torsion free****, if the torsion tensor*

$$T(X,Y) = \nabla_X Y - \nabla_Y X - [X,Y] = 0 \tag{10}$$

On a Riemannian manifold, there exist a unique metric preserving, torsion free connection which is called the **Levi-Civita connection**.

3 Family of Measures as Affine Spaces

Let $(\mathcal{X}, \mathcal{B})$ be a measurable space, where $\mathcal{X}$ is a non-empty set and $\mathcal{B}$ is the σ field of subsets of $\mathcal{X}$. Then the family of measures on $(\mathcal{X}, \mathcal{B})$ under suitable regularity conditions can be made into an affine space. It can be seen that the set of probability distributions on $(\mathcal{X}, \mathcal{B})$ is a subset of this affine space and a finite dimensional statistical model or a parametrized model can be thought of as a finite-dimensional submanifold of this affine space.

Definition 9 *Let V be an n-dimensional real vector space and Ξ be a non-empty set together with a translation map $+ : V \times \Xi \longrightarrow \Xi$, $(v, p) \longmapsto v + p$ which satisfies*

1. $\forall\ v, w \in V,\ \forall\ p \in \Xi, v + (w + p) = (v + w) + p$
2. *For any two points $p, q \in \Xi$, $\exists$ a unique vector $v \in V$ such that $q = p + v$*

*Then Ξ is said to an n-****dimensional affine space*** *over the vector space V. An affine space can be thought of as a set which becomes a vector space by selecting a point to be the origin.*

A characteristic of an affine space is the presence of special co-ordinate systems called affine co-ordinate system. Let $\mathcal{O}$ be a chosen origin for Ξ and let us choose $\{v_1, .., v_n\}$ as an ordered basis for the vector space of translation V of Ξ. Then for any $v \in V$,

$$v = \theta^1(v)v_1 + \theta^2(v)v_2 + ... + \theta^n(v)v_n$$

where $\theta^i, i = 1, .., n$ are linear functionals on V. Let $\phi : \Xi \longrightarrow V$ be the bijection from Ξ to V determined by the choice of origin $\mathcal{O}$. That is, for any $p \in \Xi$,

$$\phi(p) = v, \text{ where } p = \mathcal{O} + v \text{ for } v \in V.$$

Hence we can regard $\theta^i, i = 1, .., n$ as functions on Ξ by composing with ϕ.

$$\theta^i \circ \phi : \Xi \longrightarrow \mathbb{R}$$
$$\theta^i \circ \phi(p) = \theta^i(v)$$

We denote $\theta^i \circ \phi$ by θ^i itself. Any collection of functions defined on Ξ by such a process is called an **affine co-ordinate system**. $\{\mathcal{O}; v_1, .., v_n\}$ is called an **affine frame** of Ξ with the origin $\mathcal{O}$ and $\{\theta^i; i = 1, ..., n\}$ is called the affine co-ordinate system on Ξ with respect to the affine frame.

If θ and φ are any two affine co-ordinate systems on Ξ with respect to affine frames $\{p; v_1, .., v_n\}$ and $\{q; w_1, .., w_n\}$ respectively. Then there exists a non-singular matrix X^i_j and a vector $u = [u^1, ..., u^n]$ both depending on θ and φ with

$$\theta^i(p) = \textstyle\sum_{i=1}^n X^i_j \varphi^j(p) + u^i$$

The origin q of the φ system can be written as $q = p + u^1 v_1 + ... + u^n v_n$ for $u^i \in \mathbb{R}; i = 1, ..., n$. In fact $u = [u^1, ..., u^n]$ is the θ-coordinates for the origin of the φ system. And X^i_j is the matrix given by

$$[X^i_j] = [\ [w_1]_\theta, ..., [w_n]_\theta\ ]$$

where $[w_i]_\theta$ is the column vector whose elements are the coefficient of w_i with respect to the basis $\{v_1, ..., v_n\}$. That is, X^i_j is the change of basis matrix from the basis $\{v_1, ..., v_n\}$ to $\{w_1, ..., w_n\}$ and clearly it is non-singular. We say that two co-ordinates θ and φ related in this way are affinely related. And conversely, if we have a set Ξ and a collection of co-ordinates on Ξ where any two are affinely related, then Ξ is an affine space.

If V is infinite dimensional, then Ξ is an infinite dimensional affine space. In that case, we do not have a co-ordinate expression as in the finite dimensional case.

Example 2 *1. Any vector space is an affine space over itself. Translation map is just the addition operation of the vector space.*

2. Let $\Xi = (0, \infty)$, $V = \mathbb{R}$. For $x \in \Xi, v \in \mathbb{R}$, define

$$x + v = \exp(v)x \tag{11}$$

3.1 Affine structure of the family of measures

We consider the family $\mathcal{A}$ of non-negative, $\sigma-$finite measures on $(\mathcal{X}, \mathcal{B})$. Define an equivalence relation $\sim$ on $\mathcal{A}$ by two measures in $\mathcal{A}$ are equivalent if they are absolutely continuous with respect to each other. That is two measures are equivalent if they have the same sets of measure zero. Let $\mathcal{M}$ denote one of the equivalence classes of $\mathcal{A}$. So $\mathcal{M}$ is the set of all non-negative, $\sigma-$finite measures on Ω which are absolutely continuous with respect to each other.
Let $R_{\mathcal{X}}$ be the set of all measurable functions defined on $(\mathcal{X}, \mathcal{B})$. Clearly $R_{\mathcal{X}}$ is a vector space under the addition and scalar multiplication operations defined by

$$\begin{aligned} (f+g)(A) &= f(A)+g(A) && (12)\\ (cf)(A) &= cf(A) \quad \forall\, f,g \in R_{\mathcal{X}}, A \in \mathcal{B}, c \in \mathbb{R} && (13)\end{aligned}$$

Note that in general, $R_{\mathcal{X}}$ is an infinite dimensional vector space. Let us assume that for every $\nu \in \mathcal{M}$ and $f \in R_{\mathcal{X}}$, $e^f\nu$ is a $\sigma-$finite measure. Then $\mathcal{M}$ can be made into an affine space over the vector space $R_{\mathcal{X}}$ under the translation operation defined by

$$\nu + f = e^f\nu \quad \forall\, f \in R_{\mathcal{X}}, \nu \in \mathcal{M} \tag{14}$$

1. It is easy to see that for any $\mu \in \mathcal{M}$ and $f \in R_{\mathcal{X}}$, $\nu = e^f\mu$ is a non-negative, $\sigma-$finite measure on $(\mathcal{X}, \mathcal{B})$. In fact $\nu(A) = \int\limits_A e^f d\mu,\ A \in \mathcal{B}$, defined in this way has the property that $\nu(E) = 0$ whenever $\mu(E) = 0$ for $E \in \mathcal{B}$ and hence $\nu \ll \mu$ and $e^f\mu \in \mathcal{M}$.

2. For any two measures $\nu, \mu \in \mathcal{M}$, $\exists$ a unique function $f = \frac{d\nu}{d\mu} \in R_{\mathcal{X}}$ (In fact f is the Radon-Nikodym derivative) which translates μ to ν. We often call e^f as the density function with respect to the measure μ.

3. $\forall\ f, g\ \in R_{\mathcal{X}},\ \forall \mu \in \mathcal{M} \qquad (\mu + f) + g = e^f\mu + g = e^g e^f \mu = e^{f+g}\mu = \mu + (f+g)$
   (Note that the same symbol + is used for vector space addition operation and affine space translation operation.)

Hence $\mathcal{M}$ is an affine space over the vector space $R_{\mathcal{X}}$. So by choosing an origin μ, $\mathcal{M}$ can be identified with the vector space $R_{\mathcal{X}}$. It is equivalent to saying that any measure in $\mathcal{M}$ can be expressed as densities with respect to the chosen origin.

Definition 10 *Let the map $\ell : \mathcal{M} \longrightarrow R_{\mathcal{X}}$ be given by*

$$\ell(\nu) = \ell(p\mu) = \log p$$

for any $\nu \in \mathcal{M}$ *In fact* $p = e^f$ *for* $f \in R_{\mathcal{X}}$. *Hence* $\ell(\nu) = \log p = \log e^f = f \in R_\Omega$ *is well defined.*

Let $\mathcal{P}$ be the space of all probability measures in $\mathcal{M}$. Notice that probability measures cannot form an affine subspace of $\mathcal{M}$. Because multiplying them by functions e^f will almost certainly destroy the property that their total mass is one or even total mass is finite. Rather than regarding them as points in $\mathcal{M}$, probability measures can also be regarded as non-negative finite measures up to scale because we can always divide a non-negative finite measure by its total mass to get a probability measure.

Define an equivalence relation $\sim'$ on $\mathcal{M}$ by two measures are equivalent if they are multiples of each other. In fact rescaling is one of the translation operations namely translation by a constant c (that is multiplication by e^c). Let $\mathcal{M}'$ denote the set of measures in $\mathcal{M}$ identified up to scale. That is $\mathcal{M}'$ is the set of all equivalence class of the equivalence relation $\sim'$ defined above. We can easily see that $\mathcal{M}'$ is also an affine space whose space of translation vectors are measurable functions identified up to addition of a constant. Let $R1$ be the vector space spanned by constant random variable 1. The space of measurable functions in $R_{\mathcal{X}}$ identified up to addition of a constant is therefore $R_{\mathcal{X}}/R1$.

$$R_{\mathcal{X}}/R1 = \{[f] = f + R.1 \ / f \in R_{\mathcal{X}}\} = \{f + c \ /; \ c \in \mathbb{R} \ , \ f \in R_{\mathcal{X}}\} \tag{15}$$

$R_{\mathcal{X}}/R1$ is the quotient space under the addition and scalar multiplication operations defined by

$$(f + R1) + (g + R1) = (f + g) + R1 \tag{16}$$
$$c(f + R1) = cf + R1 \tag{17}$$

Then $\mathcal{M}'$ is an affine space over the vector space $R_{\mathcal{X}}/R1$ and translation operation is defined as $[\nu] + [f] = [e^f \nu]$ where $[\nu] = \{e^c \nu / c \in \mathbb{R}\}$
The set of probability measures $\mathcal{P}$ is a subset of this affine space namely the set corresponding to finite measures up to scale.

4 Statistical Manifold

We have seen that measures can be expressed as densities with respect to some base measure. In most of the practical applications, the base measure will be Lebesgue measure on $\mathbb{R}^n$. Here we consider the sample space $\mathcal{X} \subseteq$

$\mathbb{R}^n$. Any probability measure on $\mathcal{X} \subseteq \mathbb{R}^n$ can be represented in terms of density function with respect to Lebesgue measure. We represent probability distributions on a set $\mathcal{X}$ using the density functions as

1. If $\mathcal{X}$ is a discrete set (finite or countably infinite cardinality), then by probability distribution on $\mathcal{X}$ we mean that a function $p : \mathcal{X} \longrightarrow \mathbb{R}$ which satisfies

$$p(x) \geq 0 \ (\forall \ x \in \mathcal{X}) \quad \text{and} \quad \sum_{x \in \mathcal{X}} p(x) = 1 \tag{18}$$

2. If $\mathcal{X} = \mathbb{R}^n$, then by probability distribution on $\mathcal{X}$ we mean function $p : \mathcal{X} \longrightarrow \mathbb{R}$ which satisfies

$$p(x) \geq 0 \ (\forall \ x \in \mathcal{X}) \quad \text{and} \quad \int_{\mathcal{X}} p(x)dx = 1 \tag{19}$$

(Note that if $n \geq 2$, then $\int$ denotes a multiple integral)

Definition 11 *Consider a family $\mathcal{S}$ of probability distributions on $\mathcal{X}$. Suppose each element of $\mathcal{S}$ can be parametrized using n real-valued variables $(\theta^1, ..., \theta^n)$ so that*

$$\mathcal{S} = \{p_\theta = p(x; \theta) \ / \ \theta = (\theta^1, ..., \theta^n) \in \mathbb{E}\} \tag{20}$$

*where $\mathbb{E}$ is a subset of $\mathbb{R}^n$ and the mapping $\theta \mapsto p_\theta$ is injective. We call such family $\mathcal{S}$ an n-dimensional **statistical model** or a **parametric model** or simply a **model** on θ. We often write as $\mathcal{S} = \{p_\theta\}$*

Let us now state several regularity conditions regarding the statistical model $\mathcal{S} = \{p_\theta\}$ which are required for our geometric theory.

Regularity conditions

1. We assume that $\mathbb{E}$ is an open subset of $\mathbb{R}^n$ and for each $x \in \mathcal{X}$, the function $\theta \mapsto p(x; \theta)$ is of class c^∞

2. Let $\ell(x; \theta) = \log p(x; \theta)$. For every fixed θ, n functions in x, $\{\partial_i \ell(x; \theta); i = 1, ..., n\}$ are linearly independent.

3. The order of integration and differentiation may be freely rearranged.

4. The moments of $\partial_i \ell(x; \theta)$ exists upto necessary orders.

5. For a probability distribution p on Ω, let the support of p be defined as, $\text{supp}(p) := \{x \mid p(x) > 0\}$. The case when $\text{supp}(p_\theta)$ varies with θ poses rather significant difficulties for analysis. Hence we assume that $\text{supp}(p_\theta)$ is constant with respect to θ. Then we can redefine $\mathcal{X}$ to be $\text{supp}(p_\theta)$. This is equivalent to assuming that $p(x;\theta) > 0$ holds for all $\theta \in \mathbb{E}$ and all $x \in \mathcal{X}$. This means that the model $\mathcal{S}$ is a subset of

$$\mathcal{P}(\mathcal{X}) := \{p : \mathcal{X} \longrightarrow \mathbb{R} \mid p(x) > 0 \ (\forall\, x \in \mathcal{X}); \int_{\mathcal{X}} p(x)dx = 1\} \qquad (21)$$

Definition 12 *For a model $\mathcal{S} = \{p_\theta \ / \ \xi \in \mathbb{E}\}$, the mapping $\varphi : \mathcal{S} \longrightarrow \mathbb{R}^n$ defined by $\varphi(p_\theta) = \theta$ allows us to consider $\varphi = (\theta^i)$ as a coordinate system for $\mathcal{S}$. Suppose we have a c^∞ diffeomorphism $\psi : \mathbb{E} \longrightarrow \psi(\mathbb{E})$, where $\psi(\mathbb{E})$ is an open subset of $\mathbb{R}^n$. Then if we use $\rho = \psi(\theta)$ instead of ξ as our parameter, we obtain $\mathcal{S} = \{p_{\psi^{-1}(\rho)} \mid \rho \in \psi(\mathbb{E})\}$. This expresses the same family of probability distributions $\mathcal{S} = \{p_\xi\}$. If we consider parametrizations which are c^∞ diffeomorphic to each other to be equivalent, then we may consider $\mathcal{S}$ as a c^∞ differentiable manifold and we call it as a* ***statistical manifold****.*

Example 3 *(Normal Distribution)*
$\mathcal{X} = \mathbb{R}$, $n = 2$, $\theta = (\mu, \sigma)$, $E = \{(\mu, \sigma) \ / \ -\infty < \mu < \infty, 0 < \sigma < \infty\}$

$$N(\mu, \sigma) = \{p(x;\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\} \qquad (22)$$

This is a 2-dimensional manifold which can be identified with the upper half plane.

5 The Exponential Family

The Exponential family is a practically convenient and widely used unified family of distributions on finite dimensional Euclidean spaces parametrized by a finite dimensional parameter vector. It contains as special cases most of the standard discrete and continuous distributions that we use for practical modelling, such as the normal, Poisson, Binomial, exponential, Gamma, multivariate normal, etc.

Definition 13 *The standard form of a n-dimensional exponential family of distributions $\mathcal{S} = \{p(x;\theta) \ / \ \theta \in E \subseteq \mathbb{R}^n\}$ is defined as*

$$p(x;\theta) = \exp(\sum_{i=1}^{n} \theta^i x_i - \psi(\theta)) \quad or \quad \log(p(x;\theta)) = \sum_{i=1}^{n} \theta^i x_i - \psi(\theta) \qquad (23)$$

where $x = (x_1, ..., x_n)$ is a set of random variables, $\theta = (\theta^1, .., \theta^n)$ are the canonical parameters and $\psi(\theta)$ is determined from the normalization condition.

5.1 Geometric Structures on Statistical Manifolds

Let $\mathcal{S} = \{p_\theta \mid \theta \in \mathbb{E} \subseteq \mathbb{R}^n\}$ is a statistical manifold. Let $T_\theta(\mathcal{S})$ be the tangent space to $\mathcal{S}$ at a point p_θ is given by

$$T_\theta(\mathcal{S}) = \{\alpha^i \partial_i \ / \ \alpha^i \in \mathbb{R}\} \tag{24}$$

The tangent space to a statistical manifold can be represented in a more convenient way as follows.
For the statistical manifold $\mathcal{S} = \{p(x;\theta)\}$, define $\ell(x;\theta) = \log p(x;\theta)$ and consider the partial derivatives $\partial_i \ell; i = 1, ..., n$. By our assumption, $\partial_i \ell; i = 1, ..., n$ are linearly independent functions in x. We can construct the following n-dimensional vector space spanned by n functions $\partial_i \ell; i = 1, ..., n$ in x as,

$$T^1_\theta(\mathcal{S}) = \{A(x) \ / \ A(x) = A^i \partial_i \ell\} \tag{25}$$

There is a natural isomorphism between these two vector spaces $T_\theta(\mathcal{S})$ and $T^1_\theta(\mathcal{S})$ given by

$$\partial_i \in T_\theta(\mathcal{S}) \longleftrightarrow \partial_i \ell(x;\theta) \in T^1_\theta(\mathcal{S}) \tag{26}$$

Any tangent vector $A = A^i \partial_i \in T_\theta(\mathcal{S})$ corresponds to a random variable $A(x) = A^i \partial_i \ell(x;\theta) \in T^1_\theta(\mathcal{S})$ having the same components A^i. Note that $T_\theta(\mathcal{S})$ is the differentiation operator representation of the tangent space, while $T^1_\theta(\mathcal{S})$ is the random variable representation of the same tangent space. The space $T^1_\theta(\mathcal{S})$ is called the $1-$**representation of the tangent space**.

Define expectation with respect to the distribution $p(x;\theta)$ as

$$E_\theta(f) = \int f(x) p(x;\theta) dx \tag{27}$$

Note that $E_\theta[\partial_i \ell_{x;\theta}] = 0$ since $p(x;\theta)$ satisfies

$$\int p(x;\xi) dx = 1 \tag{28}$$

Hence for any random variable $A(x) \in T^1_\theta(\mathcal{S})$, we have $E_\theta[A(x)] = 0$.
This expectation induces an inner product on $\mathcal{S}$ in a natural way. Let A and B be two tangent vectors in $T_\theta(\mathcal{S})$ and $A(x)$ and $B(x)$ be the $1-$ representations of A and B. Then their inner product can be defined as

$$< A, B >_\theta = E_\theta[A(x)B(x)]$$

Especially the inner product of the basis vectors ∂_i and ∂_j is

$$\begin{aligned} g_{ij}(\theta) = < \partial_i, \partial_j >_\theta &= E_\theta[\partial_i \ell(x;\theta)\partial_j \ell(x;\theta)] && (29) \\ &= -E[\partial_i \partial_j \ell(x;\theta)] && (30) \\ &= \int \partial_i \ell(x;\theta)\partial_j \ell(x;\theta)p(x;\theta)dx && (31) \end{aligned}$$

It is clear that the matrix $G(\theta) = (g_{ij}(\theta))$ is symmetric (i.e $g_{ij} = g_{ji}$). For any n-dimensional vector $c = [c^1, ..., c^n]^t$

$$c^t G(\theta)c = \int \{\sum_{i=1}^{n} c^i \partial_i \ell(x;\theta)\}^2 p(x;\theta)dx \geqslant 0 \qquad (32)$$

Since $\{\partial_1 \ell(x;\theta), ..., \partial_n \ell(x;\theta)\}$ are linearly independent, G is positive definite. Hence $g =<, >$ defined in (31) is a Riemannian metric on the statistical manifold $\mathcal{S}$, called the **Fisher information metric**.

Example 4 ***Normal distribution***
For the normal family

$$\mathcal{S} = N(\mu, \sigma) = \{p(x;\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}\} \qquad (33)$$

with parameters $\theta = (\mu, \sigma)$, the log-likelihood function is given by

$$\ell(x, \theta) = -\frac{(x-\mu)^2}{2\sigma^2} - \log\sqrt{2\pi}\sigma$$

The tangent space $T^1_\theta \mathcal{S}$ is spanned by $\partial_1 = \frac{\partial}{\partial\mu}$ and $\partial_2 = \frac{\partial}{\partial\sigma}$

$$\partial_1 = \frac{(x-\mu)}{\sigma^2}\ ,\ \partial_2 = -\frac{(x-\mu)^2}{\sigma^3} - \frac{1}{\sigma}$$

Then the Fisher information matrix $G(\theta) = (g_{ij})$ is given by

$$\begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix}$$

Definition 14 *Let $\mathcal{S} = \{p(x;\theta) \ / \ \theta \in \mathbb{E}\}$ be an n-dimensional statistical manifold with the Fisher metric g. We can define n^3 functions Γ_{ijk} by*

$$\Gamma^1_{ijk} = E_\theta[(\partial_i\partial_j\ell(x;\theta))(\partial_k\ell(x;\theta))] \tag{34}$$

Γ^1_{ijk} uniquely determine an affine connection ∇ on the statistical manifold $\mathcal{S}$ by

$$\Gamma^1_{ijk} =< \nabla^1_{\partial_i}\partial_j, \partial_k > \tag{35}$$

*∇^1 is called $1-$**connection** or the **exponential connection**.*

Previously, $\ell(x;\theta)$ the logarithm of the density function $p(x;\theta)$ in a statistical model $S = \{p(x;\theta)\}$ to define the fundamental geometric structures. Amari defined one parameter family of functions called the α - embedding indexed by $\alpha \in \mathbf{R}$.

Definition 15 *$\alpha-$**representation***
Let $L_{(\alpha)}(p)$ be a one parameter family of functions defined by

$$L_{(\alpha)}(p) = \begin{cases} \frac{2}{1-\alpha}p^{\frac{1-\alpha}{2}} & \alpha \neq 1 \\ \log p & \alpha = 1 \end{cases} \tag{36}$$

and we call

$$\ell_{(\alpha)}(x;\theta) = L_{(\alpha)}(p(x;\theta)) \tag{37}$$

*the $\alpha-$**representation** of the density function $p(x;\theta)$.*

The $1-$representation $\ell_1(x;\theta)$ is the log-likelihood function $\ell(x;\theta)$ and the $(-1)-$representation $\ell_{-1}(x;\theta)$ is the density function $p(x;\theta)$ itself.

Let $T^\alpha_\theta(\mathcal{S})$ be the vector space spanned by n linearly independent functions $\partial_i\ell_\alpha(x;\theta)$ in x, $i = 1, ..., n$.

$$T^\alpha_\theta(\mathcal{S}) = \{A(x) \ / \ A(x) = A^i\partial_i\ell_\alpha(x;\theta)\} \tag{38}$$

There is a natural isomorphism between these two vector spaces $T_\theta(\mathcal{S})$ and $T^\alpha_\theta(\mathcal{S})$ given by

$$\partial_i \in T_\theta(\mathcal{S}) \longleftrightarrow \partial_i\ell_\alpha(x;\theta) \in T^\alpha_\theta(\mathcal{S}) \tag{39}$$

The vector space $T^\alpha_\theta(\mathcal{S})$ is called the $\alpha-$**representation of the tangent space** $T_\theta(\mathcal{S})$. The $\alpha-$representation of a vector $A = A^i\partial_i\ell \in T_\theta(\mathcal{S})$ is the random variable

$$A_\alpha(x) = A^i\partial_i\ell_\alpha(x;\theta) \tag{40}$$

Let us define the $\alpha-$**expectation** of a random variable f with respect to the density $p(x;\theta)$ as

$$E_\theta^\alpha(f) = \int f(x)p(x;\theta)^\alpha dx. \tag{41}$$

Let A and B be two tangent vectors in $T_\theta(\mathcal{S})$ and $A_\alpha(x)$ and $B_\alpha(x)$ be the $\alpha-$representations of A and B. Then an inner product can be defined naturally as

$$< A, B >_\theta^\alpha = E_\theta^\alpha[A_\alpha(x)B_\alpha(x)] \tag{42}$$

We have the relation

$$\partial_i \ell_\alpha(x;\theta) = p^{\frac{(1-\alpha)}{2}} \partial_i \ell(x;\theta) \tag{43}$$

Thus we have

$$\begin{aligned} < \partial_i, \partial_j >_\theta^\alpha &= \int \partial_i \ell_\alpha(x;\theta)\partial_j \ell_\alpha(x;\theta)p(x;\theta)^\alpha dx & (44)\\ &= \int \partial_i \ell \; \partial_j \ell \; p(x;\theta) \; dx & (45)\\ &= g_{ij}(\theta) & (46) \end{aligned}$$

$$(\partial_i \ell_\alpha)(\partial_j \ell_{-\alpha}) = p(x;\theta) \; \partial_i \ell \; \partial_j \ell \tag{47}$$

The inner product has the following dualistic expression for any α,

$$< A, B >_\theta = \int A_\alpha(x,\theta)B_{-\alpha}(x;\theta)dx \tag{48}$$

Then we say that the two vector spaces $T_\theta^\alpha(\mathcal{S})$ and $T_\theta^{-\alpha}(\mathcal{S})$ are **dually coupled**. That is the inner product of two vectors A and B is given by the integration of the product of their $\alpha-$ and $(-\alpha)-$representations.
We have,

$$\partial_i \partial_j \ell_\alpha = p^{\frac{(1-\alpha)}{2}} (\partial_i \partial_j \ell + \frac{1-\alpha}{2} \partial_i \ell \partial_j \ell) \tag{49}$$

Hence we can define n^3 functions Γ_{ijk}^α by

$$\Gamma_{ijk}^\alpha = \int \partial_i \partial_j \ell_\alpha(x;\theta)\partial_k \ell_{-\alpha}(x;\theta)dx \tag{50}$$

These Γ_{ijk}^α uniquely determine connections ∇^α on the statistical manifold $\mathcal{S}$ by

$$\Gamma_{ijk}^\alpha = < \nabla_{\partial_i}^\alpha \partial_j, \partial_k > \tag{51}$$

which is called α**-connection**.
Thus the one parameter family of functions $L_\alpha(p)$ defines a family of connections ∇^α, $\alpha \in \mathbb{R}$ on the statistical manifold $\mathcal{S}$.

Lemma 5.1 *The $\alpha-$connection ∇^{α} and the $-(\alpha)-$connection $\nabla^{-\alpha}$ are dual with respect to the Fisher information metric. In particular, the $0-$ connection is the Levi-Civita connection or the metric connection.*

Proof By the use of $\alpha-$representation, we have

$$\begin{aligned} A < B, C > &= A \int B_{\alpha}(x,\theta) C_{-\alpha}(x;\theta) dx \qquad (52) \\ &= \int (AB_{\alpha}(x,\theta)) C_{-\alpha}(x;\theta) dx + \int B_{\alpha}(x,\theta)(AC_{-\alpha}(x;\theta)) dx \\ &= < \nabla^{\alpha}_{A} B, C > \; + \; < B, \nabla^{-\alpha}_{A} C > \qquad (53) \end{aligned}$$

Now consider the exponential family $\mathcal{S} = \{p(x;\theta) \; / \; \theta \in E \subseteq \mathbb{R}^n\}$ where $p(x,\theta) = exp[\sum_{i=1}^{n} \theta^i x_i - \psi(\theta)]$.
Now $\partial_i l(x;\theta) = x_i - \partial_i \psi(\theta), \qquad \partial_i \partial_j l(x;\theta) = -\partial_i \partial_j \psi(\theta)$
Then $\Gamma^1_{ijk} = \partial_i \partial_j \psi(\theta) E_{\theta}(\partial_k l_{\theta}) = 0$
Thus we have $\nabla^1_{\partial_i} \partial_j = 0$. Then we say that the exponential family is 1 - flat. By duality we get it is -1 - flat also. Thus the exponential family is a dully flat space with respect to the ± 1 connections defined by Amari. Thus we have the

Theorem 5.2 *The exponential family is a dually flat space with respect to the ± 1 connections defined by Amari.*

A dually flat space is an important tool in the geometric study of statistical estimation. Now we have seen that the important statistical model the exponential family has a dually flat structure with respect to the $\alpha = \pm 1-$ connections.

6 Deformed Exponential Family

For any $\alpha \in \mathbb{R}$, Amari defined an $\alpha-$family of probability density functions $\mathcal{S} = \{p(x;\theta) \; / \; \theta \in E \subseteq \mathbb{R}^n\}$ as

$$L_{\alpha}(p(x;\theta)) = \sum_{i=1}^{n} \theta^i x_i - \psi(\theta) \qquad (54)$$

where $L_{\alpha}(p)$ is the $\alpha-$embedding.
When $\alpha = 1$, the $\alpha-$**family** is the exponential family. The exponential family is $1-$flat. But for $\alpha \neq 1$, $\alpha-$family is not flat with respect to the

$\alpha-$connection. So how to get dually flat connections on a $\alpha-$family? q-exponential family originated from statistical physics gave an answer to this. It is shown that a $q-$exponential family, which is an $\alpha-$family with $\alpha = 1-2q$, has a dually flat structure called q-structure. Moreover the q-geometry is the conformal flattening of $\alpha-$geometry.

Definition 16 *Two statistical manifolds (M, ∇, h) and $(M, \tilde{\nabla}, \tilde{h})$ are said to be β-conformally equivalent if there exist a positive function ϕ on M such that*

$$\tilde{h}(X,Y) = \phi\, h(X,Y) \tag{55}$$

$$\tilde{h}(\tilde{\nabla}_X Y, Z) = \phi\, h(\nabla_X Y, Z) + \frac{1-\beta}{2}\{h(Y,Z)d\phi(X) + h(X,Z)d\phi(Y)\} - \frac{1+\beta}{2}h(X,Y)d\phi(Z) \tag{56}$$

In terms of the basis vectors, we can rewrite the above expression as

$$\tilde{h}(\partial_i, \partial_j) = \tilde{h}_{ij} = \phi\, h(\partial_i, \partial_j) = \phi\, h_{ij} \tag{57}$$

$$\tilde{\Gamma}^{\beta}_{ijk} = \phi\, \Gamma_{ijk} + \frac{1-\beta}{2}\{h_{jk}\partial_i\phi + h_{ik}\partial_j\phi\} - \frac{1+\beta}{2}h_{ij}\partial_k\phi \tag{58}$$

Now let us describe the geometry of a $q-$**exponential family**.
$q-$logarithm is defined by

$$\log_q(u) = \frac{1}{1-q}(u^{1-q} - 1);\ q > 0 \tag{59}$$

and its inverse function q-exponential by

$$\exp_q(u) = \{1 + (1-q)u\}^{\frac{1}{1-q}};\ u > \frac{-1}{1-q} \tag{60}$$

in the limiting case $q \longrightarrow 1$, we get

$$\log_q(u) = \log u \tag{61}$$

$$\exp_q(u) = \exp u \tag{62}$$

Definition 17 *A statistical manifold $\mathcal{S} = \{p(x;\theta)\ /\ \theta \in E \subseteq \mathbb{R}^n\}$ is said to be a* ***q-exponential family*** *if*

$$\log_q p(x;\theta) = \sum_{i=1}^{n} \theta^i x_i - \psi_q(\theta) \tag{63}$$

where $\psi_q(\theta)$ is obtained from the normalization $\int p(x;\theta)dx = 1$.

Define a functional

$$h_q(\theta) = \int p(x;\theta)^q dx \tag{64}$$

and q-Riemannian metric g^q by

$$g^q_{ij}(\theta) = \partial_i \partial_j \psi_q(\theta) = \frac{q}{h_q(\theta)} \int (x_i - \partial_i \psi_q(\theta))\ (x_j - \partial_i \psi_q(\theta)) p(x;\theta)^{2q-1}\ dx \tag{65}$$

q-Riemannian metric can be written as

$$\begin{aligned} g^q_{ij}(\theta) &= \frac{q}{h_q(\theta)} \int (x_i - \partial_i \psi_q(\theta))\ (x_j - \partial_i \psi_q(\theta)) p(x;\theta)^{2q-1}\ dx & (66)\\ &= \frac{q}{h_q(\theta)} \int \partial_i p\ \partial_j p\ \frac{1}{p}\ dx & (67)\\ &= \frac{q}{h_q(\theta)} g_{ij}(\theta) & (68) \end{aligned}$$

where g is the Fisher information metric. When $q = 1$, $q-$Riemannian metric reduces to the Fisher information metric.
Using the convex function ψ_q, a divergence of Bregman type called $q-$divergence is defined as

$$D_q[p(x;\theta_1) : p(x;\theta_2)] = \psi_q(\theta_2) - \psi_q(\theta_1) - \nabla \psi_q(\theta_1).(\theta_2 - \theta_1) \tag{69}$$

The $q-$divergence can be written as

$$D_q[p;r] = \frac{1}{h_q(\theta)} \int (\log_q(p) - \log_q(r))\ p^q\ dx \tag{70}$$

Let

$$\tilde{D}_q[p;r] = \int (\log_q(p) - \log_q(r))\ p^q\ dx \tag{71}$$

Then $\tilde{D}_q$ is a constant multiple of well known α-divergence with $\alpha = 1 - 2q$. Thus q-divergence takes the form

$$D_q[p;r] = \frac{1}{h_q(\theta)} \tilde{D}_q[p;r] \tag{72}$$

The $q-$divergence D_q induces an affine connection ∇^{D_q} given by

$$\Gamma^{D_q}_{ijk} = \partial_i \partial_j \partial_k \psi_q(\theta) \tag{73}$$

We have

$$\begin{aligned}\partial_i\partial_j\partial_k\psi_q(\theta) &= \frac{q}{h_q(\theta)}\left(\int \partial_i\partial_j\ell\ \partial_k\ell\ p\ dx + (2-q)\int \partial_i\ell\ \partial_j\ell\ \partial_k\ell\ p\ dx\right)\\ &+ \frac{q}{h_q(\theta)}\left(\int \partial_k\partial_j\ell\ \partial_i\ell\ p\ dx + \int \partial_i\partial_k\ell\ \partial_j\ell\ p\ dx\right) \qquad (74)\end{aligned}$$

The dual D_q^* of D_q induces

$$\Gamma_{ijk}^{D_q^*} = 0 \qquad (75)$$

Thus $q-$divergence induces a dually flat structure on $\mathcal{S}$. Hence $\mathcal{S}$ is a dually flat space. Note that this new structure is different from $\alpha-$geometry. The q-geometry can be obtained as a conformal transformation of α-divergence, where $\alpha = 1-2q$ by a gauge function $\frac{1}{h_q(\theta)}$.

6.1 Dually Flat Deformed Exponential Family

Naudts introduced a generalized notion of exponential family called the deformed exponential family. For convenience we formulate the deformed exponential family using a smooth function $F : (0,\infty) \longrightarrow \mathbb{R}$ satisfying $F' > 0$ and $F'' < 0$.

Definition 18 ***F-Exponential Family***
Let $F : (0,\infty) \longrightarrow \mathbb{R}$ be any smooth increasing concave function. Let Z be the inverse function of F. Define the standard form of an n-dimensional $F-$exponential family of distributions as

$$p(x;\theta) = Z(\sum_{i=1}^{n}\theta^i x_i - \psi(\theta)) \quad or \quad F(p(x;\theta)) = \sum_{i=1}^{n}\theta^i x_i - \psi(\theta) \qquad (76)$$

where $x = (x_1,...,x_n)$ is a set of random variables, $\theta = (\theta^1,..,\theta^n)$ are the canonical parameters and $\psi(\theta)$ is determined from the normalization condition.

Remark 1 *Note that when $F(p) = logp$ the F - exponential family is the exponential family and when $F(p) = log_q p$ it is the Q - exponential family.*

The deformed exponential family is dually flat with respect to the U-geometry defined by Naudts. Also it is dually flat with respect to the Ξ - geometry defined by Amari. Both the dually flat structures are closely related to the (F,G)- geometry introduced by Harsha and Moosath.

Chapter 2

Geometric Approach to Estimation Theory

Now we discuss the consistency and efficiency of an estimator in a statistical manifold. A standard exponential family naturally has a sufficient statistics and also the dual co-ordinate has an efficient estimate. The properties of an estimator in a curved exponential family can be captured by looking at the geometric properties of the associated ancillary manifold. Consider a n-dimensional exponential family $\mathcal{S} = \{p(x;\theta) = \exp(\sum_{i=1}^{n} \theta^i x_i - \psi(\theta)) \ / \ \theta \in E \subseteq \mathbf{R^n}\}$. Let $M = \{q(x;u) \ / \ u = (u^a) \in \mathbf{R^m}\}$ be m-dimensional curved exponential family. Then we have $q(x,u) = p(x;\theta(u))$. Let $x^1, ..., x^N$ be N independent observations from $q(x;u) \in M$. Then the sufficient statistic (or the observed point) $\bar{x} = (\bar{x}_1, ..., \bar{x}_n)$ defines a distribution in $\mathcal{S}$ whose η co-ordinate is given by $\hat{\eta}_i = \bar{x}_i$. An estimator $\hat{u}_N$ for $u \in M$ is a function of the observed point $\bar{x} = \hat{\eta}$ and can be considered as a mapping from $\mathcal{S}$ to M

$$\hat{u}_N = \hat{u}_N(\bar{x}) = \hat{u}_N(\hat{\eta}) \tag{77}$$

Now an ancillary manifold (estimating submanifold) $A_N(u)$ associated with an estimator $\hat{u}_N$ is defined as the inverse image of the mapping $\hat{u}_N$ given by

$$A_N(u) = \{\eta = (\eta_i) \in \mathcal{S} \ / \ \hat{u}_N(\eta) = u\} \tag{78}$$

Now let

$$A(u) = \lim_{N \to \infty} A_N(u) \tag{79}$$

The consistency and efficiency of $\hat{u}_N$ in a curved exponential family M can be given in term of $A(u)$

- An estimator $\hat{u}_N$ for $u \in M$ is consistent if and only if every point $\eta(u) \in M \subset \mathcal{S}$ is included in the estimating submanifold $A(u)$.
- A consistent estimator $\hat{u}_N$ for $u \in M$ is first order efficient if and only if $A(u)$ is orthogonal to M at the intersecting point $\eta(u) \in M$.

In the case of a curved exponential family, it is already known that the maximum likelihood estimator is consistent and first order efficient.

7 Estimation in a Curved Exponential Family

Consider a n-dimensional exponential family $\mathcal{S} = \{p(x;\theta) \ / \ \theta \in E \subseteq \mathbb{R}^n\}$

$$p(x;\theta) = \exp(\sum_{i=1}^{n} \theta^i x_i - \psi(\theta)). \tag{80}$$

Let $x^1, ..., x^N$ be N independent observations from $p(x;\theta) \in \mathcal{S}$. Then the joint density can be written as

$$p(x^1, ..., x^N;\theta) = \prod_{j=1}^{N} \exp(\sum_{i=1}^{n} \theta^i x_i^j - \psi(\theta)) \tag{81}$$

or

$$\ell(x^1, ..., x^N;\theta) = N \left[\sum_{i=1}^{n} \theta^i \bar{x}_i - \psi(\theta) \right] \tag{82}$$

where $\bar{x} = (\bar{x}_i)$ is the arithmetic mean given by

$$\bar{x}_i = \frac{x_i^1 + ... + x_i^N}{N} \ ; \ i = 1, .., n \tag{83}$$

That is the joint density $p(x^1, ..., x^N;\theta)$ depends on the N observations $x^1, ..., x^N$ through $\bar{x}$. Thus the statistic $\bar{x}$ is a sufficient statistic for the parameter θ and we call it as the observed point.

Now let us look at the estimation problem in a curved exponential family. A (n, m)-curved exponential family $M = \{q(x;u) \ / \ u = (u^a) \in \mathbb{R}^m\}$ is a m-dimensional smooth submanifold of an n-dimensional exponential family $\mathcal{S}$.

$$q(x, u) = p(x; \theta(u)) \tag{84}$$

Let $x^1, ..., x^N$ be N independent observations from $q(x;u) \in M$. Then the sufficient statistic (or the observed point) $\bar{x} = (\bar{x}_1, ..., \bar{x}_n)$ defines a distribution in $\mathcal{S}$ whose η co-ordinate is given by $\hat{\eta}_i = \bar{x}_i$. But this point need not be in the submanifold M. An estimator $\hat{u}$ for $u \in M$ is a function of the observed point $\bar{x} = \hat{\eta}$ and can be considered as a mapping from $\mathcal{S}$ to M

$$\hat{u} = \hat{u}(\bar{x}) = \hat{u}(\hat{\eta}) \tag{85}$$

Now an ancillary manifold $A(u)$ associated with an estimator $\hat{u}$ is defined as the inverse image of the mapping $\hat{u}$ given by

$$A(u) = \{\eta = (\eta_i) \in \mathcal{S} \ / \ \hat{u}(\eta) = u\} \tag{86}$$

That is $A(u)$ is the set of all observed points η in $\mathcal{S}$ which are mapped to $u \in M$ by the estimator $\hat{u}$.

Now let us analyze the characteristics of an estimator $\hat{u}$ in a curved exponential family M using the geometric properties of the ancillary submanifold $A(u)$ placed at each $u \in M$.
For the consistency of an estimator $\hat{u}$, we have the following theorem

Theorem 7.1 *Let $M = \{q(x;u) \ / \ u = (u^a) \in \mathbb{R}^m\} \subset \mathcal{S}$ be a curved exponential family. An estimator $\hat{u}$ for $u \in M$ is consistent if and only if every point $\eta(u) \in M \subset \mathcal{S}$ is included in the associated ancillary submanifold $A(u)$ attached to the point u.*

Proof Let $x^1, ..., x^N$ be N independent observations from $q(x;u) \in M$. We have

$$E[x] = \eta(u). \tag{87}$$

Then by the law of large numbers, $\bar{x}$ converges in probability (we denote it by $\xrightarrow{p}$) to $\eta(u)$ as $N \longrightarrow \infty$. i.e

$$\bar{x} \xrightarrow{p} \eta(u) \tag{88}$$

Then we have

$$\hat{u}(\bar{x}) \xrightarrow{p} \hat{u}(\eta(u)) \tag{89}$$

Hence for the estimator $\hat{u}$ to be consistent, we must have

$$\hat{u} \xrightarrow{p} u \tag{90}$$

This implies

$$\hat{u}(\eta(u)) = u \quad \Longleftrightarrow \quad \eta(u) \in A(u). \tag{91}$$

■

Now we show that the maximum likelihood estimator for a curved exponential family is consistent.

Definition 19 ***Maximum Likelihood Estimator (MLE)***
Let $\mathcal{S} = \{p(x;\theta) \ / \ \theta \in E \subseteq \mathbb{R}^n\}$ be an $n-$dimensional statistical manifold defined on a sample space $\mathcal{X} \subseteq \mathbb{R}$. Let $\{x^1, ..., x^N\}$ be N independent observations from a probability density function $p(x;\theta) \in \mathcal{S}$. Then the ***likelihood function*** *$L(\theta)$ is given by*

$$L(\theta) = p(x^1, ..., x^N, \theta) = \prod_{i=1}^{N} p(x^i;\theta) \tag{92}$$

Since log *is a strictly increasing function, maximizing the likelihood function* $L(\theta)$ *is equivalent to maximizing the log-likelihood function* $\log(L(\theta))$.
We say that $\hat{\theta}$ *is the* ***Maximum Likelihood Estimator (MLE)*** *if*

$$\hat{\theta} = \arg\max_{\theta \in E} L(\theta) = \arg\max_{\theta \in E} \log(L(\theta)) = \arg\max_{\theta \in E} \sum_{i=1}^{N} \log(p(x^i; \theta)) \tag{93}$$

Consider an exponential family $\mathcal{S} = \{p(x;\theta) \ / \ \theta \in E \subseteq \mathbb{R}^n\}$

$$p(x;\theta) = \exp(\sum_{i=1}^{n} \theta^i x_i - \psi(\theta)) \tag{94}$$

Then for N independent observations $x^1, ..., x^N$ from $p(x;\theta) \in \mathcal{S}$, the statistic $\bar{x} = (\bar{x}_i)$ is a sufficient statistic for the parameter θ

$$\bar{x}_i = \frac{x_i^1 + ... + x_i^N}{N} \ ; \ i = 1, .., n \tag{95}$$

The log-likelihood function is given by

$$\log(L(\theta)) = \ell(x^1, ..., x^N; \theta) = N \left[\sum_{i=1}^{n} \theta^i \bar{x}_i - \psi(\theta) \right] \tag{96}$$

Then the MLE $\hat{\theta}$ for θ is determined by

$$\frac{\partial}{\partial \theta^i} \log(L(\theta)) \mid_{\theta = \hat{\theta}} = 0 \tag{97}$$

which implies

$$\partial_i \psi(\hat{\theta}) = \bar{x}_i \tag{98}$$

Since the dual co-ordinates η_i is given by $\eta_i = \partial_i \psi(\theta)$, the MLE can be directly written in terms of dual co-ordinate as

$$\hat{\eta}_i = \bar{x}_i = \partial_i \psi(\hat{\theta}) \tag{99}$$

Thus the sufficient statistic $\bar{x}$ is the MLE for the exponential family $\mathcal{S}$.

7.1 MLE for a Curved Exponential Family

Now consider a curved exponential family $M = \{q(x;u) \ / \ u = (u^a) \in \mathbb{R}^m\}$ of $\mathcal{S}$. Let $x^1, ..., x^N$ be N independent observations from $q(x;u) \in M$. Then the log-likelihood function is given by

$$\ell(\bar{x}; u) = \log p(x^1, ..., x^N; \theta(u)) = N \left[\sum_{i=1}^{n} \theta^i(u) \bar{x}_i - \psi(\theta(u)) \right] \tag{100}$$

The MLE $\hat{u}$ for M satisfies

$$\frac{\partial}{\partial u^a} \log p(x^1, ..., x^N; \theta(u)) \mid_{\hat{u}} = 0; a = 1, .., m \tag{101}$$

Thus we get

$$\sum_{i=1}^{n} \frac{\partial \theta^i}{\partial u^a}(\hat{u}) \, (\bar{x}_i - \eta_i(\hat{u})) = 0; \quad a = 1, ..., m. \tag{102}$$

Thus the ancillary submanifold $A(u)$ associated with MLE is given by

$$A(u) = \{\eta = (\eta_i) \in \mathcal{S} \; / \; \sum_{i=1}^{n} \frac{\partial \theta^i}{\partial u^a}(u) \, (\eta_i - \eta_i(u)) = 0; \quad a = 1, ..., m\} \tag{103}$$

Definition 20 *Let $\mathcal{S}$ be an n-dimensional manifold and let M be an m-dimensional submanifold of $\mathcal{S}$. Let ∇ be an affine connection on $\mathcal{S}$. Then M is said to be ∇-autoparallel if*

$$\nabla_X Y \in \mathcal{T}(M) \quad \forall \, X, Y \in \mathcal{T}(M) \tag{104}$$

where $\mathcal{T}(M)$ is the family of smooth vector fields on M.

Theorem 7.2 *(**Projection theorem**) Let $(\mathcal{S}, g, \nabla, \nabla^*)$ be a dually flat space and let M be a ∇^*-autoparallel submanifold of $\mathcal{S}$. Let D be the canonical divergence of $\mathcal{S}$. Given $p \in \mathcal{S}$, a necessary and sufficient condition for a point $q \in M$ to satisfy $D(p; q) = \min_{r \in M} D(p; r)$ is that the ∇-geodesic connecting p and q is orthogonal to M at q.*

From equation(24), it follows that $\eta(u) \in A(u)$ for all $u \in M$. Thus from theorem 1.1, the MLE $\hat{u}$ is a consistent estimator. Thus we have

Theorem 7.3 *Let $M \subset \mathcal{S}$ be a curved exponential family which is ∇^e-autoparallel. Then MLE $\hat{u}$ for $u \in M$ is a consistent estimator.*

Chapter 3

Higher order Efficiency

8 Consistency and Efficiency of Estimators

We analyze the higher-order asymptotic behaviors of a smooth estimator $\hat{u}(\bar{x})$ in an (n, m) - curved exponential family $M = \{q(x, u)\}$. It is assumed that the estimator $\hat{u}$ is a function, independently of N, of the arithmetic mean $\bar{x} = \left(\sum {}^{x}_{i}\right) \Big/ N$ of N independent observations. It is easy to extend the theory to the case when $\hat{u}(\bar{x})$ depends explicitly on N. The sufficient statistic $\bar{x}$ can be identified with the observed point $\hat{\eta} = \bar{x}$ in manifold S of the enveloping exponential family in the η - coordinate system. Hence, an estimator $\hat{u}$ defines a mapping $\hat{u} \ : \ S \to M$. The inverse image of u by the estimator $\hat{u}$ is

$$A(u) \ = \ \hat{u}^{-1}(u) = \{\eta \in S \mid \hat{u}(\eta) = u\}, \tag{105}$$

which forms an $(n-m)$ - dimensional submanifold $A(u)$ attached to the point $u \in M$. The value of the estimator $\hat{u}$ is u when the observed point $\hat{\eta} = \bar{x}$ belongs to $A(u)$. The family $A = \{A(u)\}$ of these $A(u)$'s is the ancillary family associated with the estimator $\hat{u}$. The asymptotic behaviors of an estimator $\hat{u}$ are closely related to the geometric properties of the associated ancillary family.

Definition 21 *A consistent estimator is said to be first-order or Fisher efficient, when its first-order term $g_1^{ab}(u)$ is minimal at all u among all other consistent estimators. Since g_1^{ab} is a matrix, the minimality of a matrix is defined by the order relation $h^{ab} \geq g^{ab}$ implying that $h^{ab} - g^{ab}$ is a non-negative definite matrix.*

Definition 22 *A first-order efficient estimator is said to be second-order efficient, when its second-order term $g_2^{ab}(u)$ is minimal at all u among all other first-order efficient estimator.*

Definition 23 *A first-order efficient estimator is third-order efficient, when its third-order term $g_3^{ab}(u)$ is minimal at all u among all other second-order efficient estimators.*

We first search for the geometric properties of the first-order efficient estimator. The first-order term of the distribution of $\hat{u}$ is given by integrating

with respect to $\tilde{v}$ the joint distribution of $\tilde{w} = (\tilde{u}, \tilde{v})$,

$$p(\tilde{w}; u) = n[\tilde{w}; g_{\alpha\beta}(u)] + O(N^{-1/2}).$$

By virtue of

$$\begin{aligned} g_{\alpha\beta}\tilde{w}^\alpha\tilde{w}^\beta &= g_{ab}\tilde{u}^a\tilde{u}^b + 2g_{ak}\tilde{u}^a\tilde{v}^k + g_{k\lambda}\tilde{v}^k\tilde{v}_\lambda \\ &= g_{k\lambda}(\tilde{v}^k + g^{k\mu}g_{\mu a}\tilde{u}^a)(\tilde{v}^\lambda + g^{\lambda\nu}g_{\nu b}\tilde{u}^b) + (g_{ab} - g^{\nu\mu}g_{a\mu}g_{b\nu})\tilde{u}^a\tilde{u}^b, \end{aligned}$$

We have

$$\int n(\tilde{w}\ ;\ g_{\alpha\beta})d\tilde{v} \ = \ \int c\ \exp\left\{-\frac{1}{2}g_{\alpha\beta}\tilde{w}^\alpha\tilde{w}^\beta\right\}d\tilde{v} = n(\tilde{u}\ ;\ g_{1\ ab}),$$

where c is the normalizing constant, $g^{k\lambda}$ is the inverse matrix of $g_{k\lambda}$ and

$$g_{1\ ab}(u) \ = \ g_{ab}(u) - g_{a\mu}(u)g_{b\nu}(u)g^{\mu\nu}(u)$$

is the inverse of the asymptotic variance $g_1^{ab}(u)$ of u. Since the term $g_{a\mu}g_{b\nu}g^{\mu\nu}$ is positive-semi-definite, $g_{1\ ab}$ is maximized and hence g_1^{ab} is minimized, when and only when $g_{a\mu}(u) = 0$ holds. This leads to the following theorem, because of $g_{a\mu} = \langle \partial_a, \partial_\mu \rangle$.

8.1 Theorem

The covariance of a consistent estimator $\hat{u}$ is given by

$$E[(\hat{u}^a - u^a)(\hat{u}^b - u^b)] \ = \ \frac{1}{N}g_1^{ab} + O(N^{-2}),$$

where g_1^{ab} is the inverse of

$$g_{1\ ab} \ = \ g_{ab} - g_{ak}g_{b\lambda}g^{k\lambda}. \tag{106}$$

A consistent estimator is first-order efficient, when and only when the associated ancillary family is orthogonal, i.e., $A(u)$ is orthogonal to M, $\langle \partial_a, \partial_\mu \rangle = g_{a\mu}(u) = 0$.

This is the geometrical interpretation of the well-known result. The term $g_{1\ ab}$ reduces to the Fisher information g_{ab} for an efficient estimator, and the asymptotic variance g_1^{ab} is equal to the inverse g^{ab} of g_{ba}. The first-order term of the distribution of an efficient estimator $\hat{u}$ is

$$p(\tilde{u}\ ;\ u) \ = \ n[\tilde{u}, g_{ab}(u)] + O(N^{-1/2}).$$

9 Second and third-order efficient estimator

The Edgeworth expansion of the ddistribution of the bias-corrected first-order efficient estimator $\hat{u}^*$ or $\hat{u}^{**}$ is calculated here. Due to the relation

$$\tilde{u}^{*^a} = \tilde{u}^{**^a} - \tilde{v}^k \partial_k C^a/(2N),$$

the moments of $\tilde{u}^*$ coincide with those of $\tilde{u}^{**}$ up to the terms of order N^{-1}. Hence, their distributions are the same up to the term of order N^{-1}. Therefore, in the following, we simply identify $\hat{u}^*$ and denote by $\hat{u}^*$ the estimator $\hat{u}^{**}$ which is bias-corrected at $(\hat{u}, 0)$. The bias of an estimator $\hat{u}$ is given by

$$E[\hat{u}^a - u^a] \;\; = \;\; b^a(u) + O(N^{-3/2}),$$

where

$$b^a(u) \;\; = \;\; -\frac{1}{2N} C^a = -\frac{1}{2N} C_{\alpha\beta} a_g \alpha\beta \tag{107}$$

is called the asymptotic bias of an first-order efficient estimator. By decomposing $g^{\alpha\beta}$ and $g^{k\lambda}$, we have

$$C^a \;\; = \;\; C_{cd} a_g^{cd} + C_{k\lambda} a_g^{k\lambda} = \Gamma_{cd}^{(m)} g^{cd} + H_{k\lambda}^{(m)} a_g^{k\lambda},$$

because of $g_{ak} = 0$, $C_{cd} a = \Gamma_{cd}^{(m)a}$ and $C_{k\lambda} a = H_{k\lambda}^{(m)a}$.

Hence the asymptotic bias b^a of an efficient estimator $\hat{u}^a$ is given by the sum of the two terms, one is derived from the mixture connection of M and is common to all the efficient estimators, and the other is derived from the mixture curvature of the associated A which depends on the estimator The bias-corrected estimator $(\hat{u}^{**})$ is then written as

$$\hat{u}^* \;\; = \;\; \hat{u} - b(\hat{u}).$$

The distribution of $\tilde{u}^*$ or $\tilde{u}^{**}$ is obtained by integrating ... with respect to $\tilde{v}^*$ or $\tilde{v}^{**}$ by the use of the relation $g_{ak} = 0$, giving the same result.

9.1 Theorem

The distribution of the bias corrected first-order efficient estimator $\tilde{u}^*$ is expanded as

$$p(\tilde{u}^* \;;\; u) \;\; = \;\; n[\tilde{u}^* \;;\; g_{ab}(u)] \left\{1 + A_N(\tilde{u}^* \;;\; u)\right\} + O(N^{-3/2}), \tag{108}$$

$$
\begin{aligned}
A_N(\tilde{u}^* \ ; \ u) &= \frac{1}{6\sqrt{N}} K_{abc} h^{abc} + \frac{1}{4N} C^2{}_{ab} h^{ab} + \frac{1}{24N} K_{abcd} h^{abcd} \\
&\quad + \frac{1}{72N} K_{abc} K_{def} h^{abcdef},
\end{aligned}
$$

where h^{abc} etc. are the Hermite polynomials in $\tilde{u}^*$ with respect to the metric g_{ab}. The third and fourth cumulants of $\tilde{u}^*$ are given by

$$
\begin{aligned}
K_{abc} &= -3\Gamma^{(-1/3)}_{abc}, \\
K_{abcd} &= S_{abcd} - 4D_{abcd} + 12(\Gamma^{(m)}_{eab} + \Gamma^{(e)}_{abe})\Gamma^{(m)}_{fcd} g^{ef},
\end{aligned}
$$

and they are common to all the first-order efficient estimators. The estimators differ only in the term $C^2_{ab} = C_{\alpha\beta a} C_{\gamma\delta b} g^{\alpha\gamma} g^{\beta\delta}$, which represents the geometric properties of the associated ancillary family A as

$$
C^2_{ab} = (\Gamma^m)_{ab}{}^2 + 2(H^e_M)^2_{ab} + (H^m_A)^2_{ab}, \tag{109}
$$

where

$$
(\Gamma^m)^2_{ab} = \Gamma^{(m)}_{cda} \ \Gamma^{(m)}_{efb} \ g^{ce} g^{df}, \tag{110}
$$

$$
(H^e_M)^2_{ab} = H^{(e)}_{ack} \ H^{(e)}_{bd\lambda} g^{cd} g^{k\lambda}, \tag{111}
$$

$$
(H^m_A)^2_{ab} = H^{(m)}_{k\lambda a} \ H^{(m)}_{\nu\mu b} g^{k\nu} g^{k\nu}. \tag{112}
$$

Proof.

Note that the associated ancillary family A is orthogonal.The identity

$$
\begin{aligned}
G_{akb} &= \langle \nabla^m_{\partial a}, \partial_b \rangle = \partial_a \langle \partial_k, \partial_b \rangle - \langle \partial_k, \nabla^e_{\partial a} \partial_b \rangle \\
&= \partial_a g_{kb} - H^{(e)}_{abk}
\end{aligned} \tag{113}
$$

is used in calculating C^2_{ab}, where $g_{kb} = 0$.

We define the contravariant versions of the quantities eqn(110), eqn(111 and eqn(112) by

$$
\begin{aligned}
(\Gamma^m)^{2ab} &= g^{ac} g^{bd} (\Gamma^m)^2_{cd}, \\
(H_M{}^e)^{2ab} &= g^{ac} g^{bd} (H_M{}^e)^2_{cd}, \\
(H_A{}^m)^{2ab} &= g^{ac} g^{bd} (H_A{}^m)^2_{cd},
\end{aligned}
$$

They are, respectively, the square of the mixture connection of M, the square of the exponential curvature of M, and the square of the mixture curvature of the ancillary submanifold $A(u)$. All of them are non-negative definite. The mean square error of a first-order efficient estimator is obtained by calculating

$E[\tilde{u^*}^a \tilde{u^*}^b]$ by the use of eqn(108), where the orthogonality of the Hermite polynomials guarantees

$$\int n(\tilde{u}^*)\tilde{u}^{*a}\tilde{u}^{*b}h^{c_1\cdots c_p}(\tilde{u}^*)d\tilde{u}^* \quad = \quad 0$$

except for $p = 2$, and

$$\int n(\tilde{u^*})\tilde{u}^{*a}\tilde{u}^{*b}h^{cd}d\tilde{u}^* \quad = \quad g^{ac}g^{bd} + g^{ad}g^{bc}$$

for $p = 2$.

9.2 Theorem

The mean square error of a bias corrected first-order efficient estimator is given by

$$E[\tilde{u}^{*a}\tilde{u}^{*b}] \quad = \quad g^{ab} + \frac{1}{2N}\left\{(\Gamma^m)^{2ab} + 2(H^e_M)^{2ab} + (H^m_A)^{2ab}\right\} + O(N^{-3/2}) \quad (114)$$

The first-order term g^{ab} is the inverse of the Fisher information g_{ba} of M. The second-order term, i.e., the term of order $N^{-1/2}$, vanishes for all the first order efficient estimators, so that a first-order efficient estimator is automatically second-order efficient. The third-order term is decomposed into the sum of three non-negative terms. The first is a half of the square of the components of the mixture connection. It depends on the manner of parametrization of M, but it is common to all the estimators. If we adopt the mixture normal coordinate system at a specific point u, it vanishes at this point. The second is the square of the exponential curvature of the model M. It is a tensor depending on the geometrical property of M, but not depending on the manner of parametrization nor the manner of estimation. The third is a half of the square of the mixture curvature of the ancillary submanifold $A(u)$ at $v = 0$. Only this term depends on the estimator. Hence, we have the following theorem.

9.3 Theorem

A bias-corrected first-order efficient estimator is automatically second-order efficient. It is third-order efficient when, and only when, the associated ancillary submanifold $A(u)$ has zero mixture curvature at $v = 0$.

References

[1] Amari, S. I. and Nagaoka, H. *Methods of Information Geometry, Translations of Mathematical Monographs*, Oxford University Press, Oxford, UK, **2000**.

[2] Murray, M.K.; Rice, R.W. *Differential Geometry and Statistics.* Chapman and Hall: London, UK, **1995**.

[3] Rao, C.R. Information and accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta. Math. Soc.* **1945**, *37*, 81–91.

[4] Harsha, K.V; Subrahamanian Moosath, K.S. F-geometry and Amari's α-geometry on a statistical manifold. *Entropy*, **2014**, *16(5)*, 2472-2487.

[5] Burbea, J. Informative geometry of probability spaces. *Expo.Math.*, **1986**, *4*, 347-378.

[6] Eguchi, S. Second order efficiency of minimum contrast estimators in a curved exponential family. *Annals of Statistics* **1983**, *11*, 793-803.

[7] Matsuzoe, H. and Ohara, A. Geometry for $q-$exponential families, Proceedings of the 2nd International Colloquium on Differential Geometry and its Related Fields, Veliko Tarnovo, September 6-10, **2010**.

[8] Amari, S. I., and Ohara, A. Geometry of q-Exponential Family of Probability Distributions, *Entropy*, **2011**, *13*, 1170-1185.

[9] Amari, S. I., Ohara, A. and Matsuzoe, H. : Geometry of deformed exponential families: Invariant, dually flat and conformal geometries. *Physica A:Statistical Mechanics and its Applications*, **2012**, *391*, 4308-4319.

[10] Harsha, K. V. and Subrahamanian Moosath, K. S. Geometry of F-likelihood Estimators and F-Max-Ent Theorem, Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MaxEnt 2014), Amboise, France, 21-26 September 2014. AIP Conference Proceedings, **2015**, *1641*, 263–270.

[11] Harsha, K.V and Subrahamanian Moosath, K.S. Dually Flat Geometries of the Deformed Exponential Family. *Physica A*, **2015**, *433*, 136-147.

[12] Kumon, M. and Amari, S. (1983). Geometrical Theory of Higher-Order Asymptotics of Test, Interval Estimator and Conditional Inference. In *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 1983, volume 387(1793), pp. 429-458.

[13] Amari, S. and Kumon, M. (1983). Differential geometry of Edgeworth expansions in curved exponential family. *Annals of the Institute of Statistical Mathematics*, 35(1): 1-24.

[14] Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *Annals of Statistics*, 3: 1189-1242.

[15] Vineetha K and K S S Moosath. Geometry of the Exponential Family. Presented at the International Conference on 'Advances in Applied Mathematics, Materials Science And Nanotechnology for Engineering and Industrial Applications', **January 2016**.