THE FINSLER GEOMETRY OF FAMILIES OF NON REGULAR DISTRIBUTIONS

SHUNICHI AMARI
TOKYO UNIVERSITY, FACULTY OF ENGINEERING, 1984 [1]

§1. Introduction.

In the present note we suggest that a non regular family of statistical distributions has a geometrical structure of Finsler space, and we intend to construct an asymptotical theory of statistical estimation based on the non regular model. For this we will use the fact that the stable distribution[2] of the vectorial statistical variables leads to a Minkowski metric. This note contains only preliminaries considerations, and it is not yet a definitive theory.

§2. Minkowski metrics and Finsler spaces.

Let T be an n-dimensional linear space. We will write an element $\mathbf{x}$ of the space T in the form

$$\mathbf{x} = x^i \mathbf{e}_i,$$

where we consider the Einstein's summation convention concerning the index i. Therefore we can express $\mathbf{x}$ by its components x^i. Let us suppose that the fundamental function of the linear space T is given, such that the length of the vector $\mathbf{x}$ is

$$||x|| = F(x).$$

The function F is a convex function homogeneous of degree one, i.e.

$$F(c\mathbf{x}) = cF(\mathbf{x}), \quad c > 0$$

and

$$F(\mathbf{x}) \geq 0$$

with equality only for $\mathbf{x} = 0$. Moreover, we assume $F(-\mathbf{x}) = F(\mathbf{x})$. A linear space endowed with a norm satisfying the above conditions is called a Minkowski space. A Euclidean space is a Minkowski space whose fundamental function is the quadratic form

$$F^2(\mathbf{x}) = g_{ij} x^i x^j,$$

---

[1]Translated from Japanese by S. V. Sabau, September 2002. [2] Stable distributions can be characterized in terms of their characteristic function $\varphi(z)$ in the following way: $\forall \lambda_1, \lambda_2 > 0, \quad \exists \lambda = \lambda(\lambda_1, \lambda_2) > 0$ such that $\varphi(\lambda z) = \varphi(\lambda_1 z)\varphi(\lambda_2 z)$. This can be reformulated as follows. If for two independent random variables X_1, X_2, distributed accordingly to the same distribution Φ, and if for any $\lambda_1, \lambda_2 > 0$ there exists a function $\lambda = \lambda(\lambda_1, \lambda_2) > 0$ such that $\dfrac{\lambda_1 X_1 + \lambda_2 X_2}{\lambda}$ belongs to Φ, then the distribution Φ is called *stable distribution*. In the case of an 1-dimensional stable distribution, if we put $\varphi(z) = \exp \psi(z)$, then it follows $\psi(\lambda z) = \psi(\lambda_1 z)\psi(\lambda_2 z)$, where $\psi(z) = (-c_0 + i\frac{z}{|z|}c_1)|z|^\alpha$. The number α is called the *index (exponent)* of the stable distribution.

where g_{ij} are positive definite tensors.

The set of vectors of T satisfying

$$F(\mathbf{x}) = 1$$

is called indicatrix (this is an $(n-1)$-dimensional surface including the origin). The region defined by $F(\mathbf{x}) \leq 1$ is convex. The indicatrix of a Euclidean space is a quadric surface (an ellipsoid).

Let us consider a differentiable manifold S and let us denote the local coordinates of a point $\theta \in S$ by $(\theta^1, \theta^2, ..., \theta^n)$. Let us denote the tangent space to the point θ of S by T_θ. A manifold endowed with a fundamental function $F(\mathbf{x}, \theta)$ in any tangent space T_θ, in other words all tangent spaces are Minkowski spaces, is called a Finsler space. We assume here the differentiability of F. A Riemannian space with the fundamental function given as a quadratic form in $\mathbf{x}$ is a special case of Finsler space. Like in a Riemannian manifold, the length of a curve in a Finsler space can be defined by the integral

$$\int F(\theta, \dot{\theta}(s))ds,$$

where $\dot{\theta} = \dfrac{d\theta}{ds}$. We do not discuss this here, but the fundamental function induces a connection in the tangent space.

Let us return to the Minkowski space T. Because F is a 1-homogeneous function,

$$g_{ij}(\mathbf{x}) = \frac{1}{2}\frac{\partial^2}{\partial x^i \partial x^j}F^2(\mathbf{x})$$

is a positive definite matrix whose elements are 0-homogeneous, i.e.

$$g_{ij}(c\mathbf{x}) = g_{ij}(\mathbf{x}).$$

Moreover, we have

$$F^2(\mathbf{x}) = g_{ij}(\mathbf{x})x^i x^j.$$

Example. The space L_p with the fundamental function

$$F(\mathbf{x}) = \{|x^1|^p + |x^2|^p\}^{\frac{1}{p}}.$$

Then we have for example

$$\begin{aligned} g_{11}(\mathbf{x}) &= (2-p)|x^1|^{2p-1}\{|x^1|^p + |x^2|^p\}^{\frac{2}{p}-2} + \\ &\quad + (p-1)|x^1|^{p-2}\{|x^1|^p + |x^2|^p\}^{\frac{2}{p}-1} \\ g_{12}(\mathbf{x}) &= |x^1|^{p-1}|x^2|^{p-1}(\frac{2}{p}-1)p\{|x^1|^p + |x^2|^p\}^{\frac{2}{p}-2}. \end{aligned}$$

For two arbitrary vectors $\mathbf{x}$ and $\mathbf{y}$ we say that $\mathbf{y}$ is orthogonal to $\mathbf{x}$ if

$$g_{ij}(\mathbf{x})x^i y^j = 0.$$

It is not necessary that $\mathbf{x}$ is orthogonal to $\mathbf{y}$.

Let us define the dual vector $\mathbf{x}^*$ of the vector $\mathbf{x}$ by the following formula. If $\mathbf{x} = (x^i)$, and $\mathbf{x}^* = (x_i)$, then

$$x_i = g_{ij}(\mathbf{x})x^j.$$

We have also that

$$F^2(\mathbf{x}) = x^i x_i.$$

The vectors $\mathbf{x}$ and $\mathbf{x}^*$ are transformed one into the other by the Legendre transformation. Namely, for

$$\mathbf{u} = \mathbf{x}^* \qquad\qquad (u_i = x_i)$$

there exists a function $H(\mathbf{u})$ such that

$$u_i = \frac{1}{2}\frac{\partial F^2(\mathbf{x})}{\partial x^i}, \qquad x^i = \frac{1}{2}\frac{\partial H^2(\mathbf{u})}{\partial u_i},$$

and we have the relation

$$F^2(\mathbf{x}) + H^2(\mathbf{u}) - 2x^i u_i = 0.$$

The function $H(\mathbf{u})$ is also 1-homogeneous, i.e.

$$H(c\mathbf{u}) = |c|H(\mathbf{u})$$

and

$$F(\mathbf{x}) = H(\mathbf{u}).$$

The function $H(\mathbf{u})$ defines a Minkowski norm in the dual space $T^* = \{\mathbf{u}\}$, namely

$$g^{ij}(\mathbf{u}) = \frac{1}{2}\frac{\partial^2 H^2(\mathbf{u})}{\partial u_i \partial u_j},$$

and

$$H^2(\mathbf{u}) = g^{ij}(\mathbf{u})u_i u_j,$$

where $g_{ij}(\mathbf{x})$ is the inverse of the matrix $g^{ij}(\mathbf{u})$. We have also

$$x^i = g^{ij}(\mathbf{u})u_j, \qquad\qquad u_j = g_{ij}(\mathbf{x})x^j.$$

§3. Stable distributions and Minkowski metrics.

Let us consider the T^*-valued random variable $\mathbf{z} = (z_i)$, $i = 1, 2, ..., n$). Let $\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_k$ be k arbitrary random variables (independent copies of the $\mathbf{z}$), and a_k be a real number. If

$$\frac{S_k}{a_k} = \frac{1}{a_k}(\mathbf{z}_1 + \mathbf{z}_2 + ... + \mathbf{z}_k)$$

are distributed accordingly to the initial distribution $\mathbf{x}$, then this distribution is called a *stable distribution*. If we define the characteristic function $\varphi(\mathbf{t})$, $\mathbf{t} = (t^i)$, $i = 1, 2, ..., n$ of a distribution by

$$\varphi(\mathbf{t}) = E[\exp\{i\mathbf{t} \cdot \mathbf{z}\}], \qquad \mathbf{t} \cdot \mathbf{z} = t^j z_j,$$

then using the characteristic function we can define the stable distribution by

$$[\varphi(\mathbf{t})]^k = \varphi(a_k\mathbf{t}).$$

It is known that $a_k = k^{\frac{1}{\alpha}}$, $1 \leq \alpha \leq 2$. A distribution with $a_k = k^{\frac{1}{\alpha}}$ is called *stable distribution with index* α. If we use the *cumulant function* $\phi(\mathbf{t})$ given by

$$\phi(\mathbf{t}) = \log\varphi(\mathbf{t}), \qquad \varphi(\mathbf{t}) = e^{\phi(\mathbf{t})}$$

we obtain

$$\phi(t) = -A(\mathbf{t})\{1 + iG(\mathbf{t})\}$$

(this is a generalization of the Lévy-Khinchin representation). Here, $A(\mathbf{t})$ is a positive α-homogeneous function, i.e.

$$A(c\mathbf{t}) = |c^\alpha|A(\mathbf{t})$$

and $G(\mathbf{t})$ is a 0-homogeneous function and

$$G(-\mathbf{t}) = -G(\mathbf{t}).$$

For simplicity from now on we assume that the distributions $\mathbf{x}$ and $-\mathbf{x}$ are equal. It follows

$$\phi(\mathbf{t}) = -A(\mathbf{t}), \qquad \varphi(\mathbf{t}) = \exp\{-A(\mathbf{t})\}.$$

Now we can introduce the Minkowski metric

$$F(\mathbf{t}) = \{A(\mathbf{t})\}^{\frac{1}{\alpha}}.$$

Hence the metric tensor is given by

$$g_{ij}(\mathbf{t}) = \frac{1}{2}\frac{\partial^2 F^2}{\partial t^i \partial t^j}.$$

In the special case $\alpha = 2$, the stable distribution becomes the Normal distribution and since $g_{ij}(\mathbf{t})$ does not depend on $\mathbf{t}$ we can define one metric using the quadratic form g_{ij}. Obviously, $\mathbf{z}_i$ is distributed as the normal distribution $N(0, g_{ij})$.

In the general case we have

$$\varphi(\mathbf{t}) = \exp\{-[g_{ij}(\mathbf{t})t^i t^j]^{\frac{\alpha}{2}}\}.$$

On the dual, we can introduce on the space of random variable $\mathbf{z} = (z_i)$ a Minkowski norm by

$$||\mathbf{z}|| = H(\mathbf{z}).$$

The function H is obtained from F by the means of Legendre transformation, namely

$$g^{ij}(\mathbf{z}) = \frac{1}{2}\frac{\partial^2 H}{\partial z_i \partial z_j}$$

and

$$||\mathbf{z}||^2 = \frac{1}{2}g^{ij}(\mathbf{z})z_i z_j.$$

If we relate $\mathbf{z}$ and $\mathbf{t}$ by

$$t^i = g^{ij}(\mathbf{z})z_j, \qquad z_i = g_{ij}(\mathbf{t})t^j,$$

we can regard $\mathbf{t}$ as random variable and we can introduce a probabilistic measure on the space T. We have to study the distribution of $\mathbf{t}$.

Now, let us assume that we have two stable distributions $\mathbf{z}_1$ and $\mathbf{z}_2$ with the same index α, and let us consider their direct product (direct sum) $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2)$. We assume that $\mathbf{z}_1$ and $\mathbf{z}_2$ are independent random variables. Then, for its corresponding dual vector $\mathbf{t} = (\mathbf{t}_1, \mathbf{t}_2)$, the characteristic function of $\mathbf{z}$ is

$$\varphi(\mathbf{t}) = E[e^{i\mathbf{z}_1\cdot\mathbf{t}_1}]E[e^{i\mathbf{z}_2\cdot\mathbf{t}_2}].$$

And therefore, the fundamental function $F(\mathbf{t})$ of $\mathbf{t}$ will be

$$F(\mathbf{t}) = [\{F_1(\mathbf{t}_1)\}^\alpha + \{F_2(\mathbf{t}_2)\}^\alpha]^{\frac{1}{\alpha}},$$

where F_1 and F_2 are the fundamental functions in $\mathbf{t}_1$, and $\mathbf{t}_1$, respectively. This clarifies the nature of the space L_α of stable distributions of index α. In this case the dual function $H(\mathbf{z})$ corresponds to the space $L_{\alpha'}$, where $\dfrac{1}{\alpha} + \dfrac{1}{\alpha'} = 1$.

§4. Stable distributions and Finsler spaces.

Let us consider the n-dimensional space of translation family given by the distribution family $S = \{p(\mathbf{x}, \theta)\}$ with the probability density function

$$p(\mathbf{x}, \theta) = f(\mathbf{x} - \theta).$$

Here, $\theta = (\theta_1, \theta_2, ..., \theta_n)$ are the parameters defining the distribution. We assume that f is a convex function of support compact, and that on the boundary has the form

$$f(\mathbf{x}) = cd^{\alpha-1}, \qquad 1 \le \alpha \le 2.$$

In the 1-dimensional case, we can take for example

$$f(x, \theta) = c_\alpha(x - \theta)^{\alpha-1}(1 - x + \theta)^{\alpha-1},$$

here x is zero in the exterior of the interval $[\theta, 1 + \theta]$.

Let us consider the distribution family S as an n-dimensional manifold with the system of coordinates given by θ. Let us denote by T_θ the tangent space to S in the point θ. If we denote by $\{\partial_i\}$ the natural basis of T_θ we can represent ∂_i as a random variable in the following way:

$$\partial_i \sim \partial_i l(\mathbf{x}, \theta), \qquad \partial_i = \frac{\partial}{\partial \theta_i},$$
$$l(\mathbf{x}, \theta) = \log p(\mathbf{x}, \theta).$$

The vector

$$\mathbf{A} = A^i \partial_i$$

of the space T_θ can be represented as the random variable

$$A(x) = A^i \partial_i l(\mathbf{x}, \theta).$$

Let us define the degree of separation (distance) of two statistical distributions $p(\mathbf{x}, \theta)$ and $p(\mathbf{x}, \theta')$ by means of the Hellinger distance

$$D(\theta, \theta') = 4 \int (\sqrt{p(\mathbf{x}, \theta)} - \sqrt{p(\mathbf{x}, \theta')})^2 d\mathbf{x},$$

or by the distance introduced by Takeuchi

$$D(\theta, \theta') = -8 \log \int \sqrt{p(\mathbf{x}, \theta) p(\mathbf{x}, \theta')} d\mathbf{x}.$$

If we consider an infinitesimal element $d\theta$ we have

$$D(\theta, \theta + d\theta) = \begin{cases} O(|d\theta|^2), & \alpha > 2, \\ |d\theta|^2 \log |d\theta|, & \alpha = 2, \\ |d\theta|^\alpha, & 1 \leq \alpha < 2. \end{cases}$$

In order to avoid complications we will consider only the regular cases $1 < \alpha < 2$ and $\alpha > 2$.

We define the norm of the tangent vector $\mathbf{h} = h^i \partial_i$ of T_θ, $(h(\mathbf{x}) = h^i \partial_i l)$ by the fundamental function

$$F(\mathbf{h}, \theta) = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \{D(\theta, \theta + \varepsilon \mathbf{h})\}^{\frac{1}{\alpha}}.$$

The function $F(\mathbf{h}, \theta)$ is positive 1-homogeneous in $\mathbf{h}$. It follows that we can define the metric tensor (the Finsler metric)

$$g_{ij}(\mathbf{h}, \theta) = \frac{1}{2} \frac{\partial^2 F^2(\mathbf{h}, \theta)}{\partial h^i \partial h^j}$$

and therefore S becomes a Finsler space.

In the case we have a decomposition in a direct product, namely

$$p(\mathbf{x}, \theta) = p_1(\mathbf{x}_1, \theta_1) p_2(\mathbf{x}_2, \theta_2), \ \mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2), \ \theta = (\theta_1, \theta_2),$$

obviously we get

$$F(\mathbf{x}, \theta) = [\{F_1(\mathbf{x}_1, \theta_1)\}^\alpha + \{F_1(\mathbf{x}_1, \theta_1)\}^\alpha]^{\frac{1}{\alpha}},$$

where $\mathbf{h} = (\mathbf{h}_1, \mathbf{h}_2)$, and F_1, F_2 are fundamental functions of each component, respectively. In this case, two vectors

$$(\mathbf{h}_1, 0), \qquad (\mathbf{h}_2, 0)$$

are reciprocally orthogonal.

In the case of identical independent distributions, i.e.

$$p(\mathbf{x}_1, ..., \mathbf{x}_m; \theta) = \prod_{i=1}^{m} p(\mathbf{x}_i, \theta),$$

the fundamental function F_m obtained from m observations is

$$F_m(\mathbf{h}, \theta) = m^{\frac{1}{\alpha}} F(\mathbf{h}, \theta).$$

Let us consider next the random variable $\mathbf{z} = (\mathbf{z}_i)$, where

$$z_i = \partial l(\mathbf{x}, \theta).$$

In the following we fix the point θ and we will omit the symbol θ. Obviously,

$$E[z_i] = 0.$$

Moreover, for the independent observations $\mathbf{x}_1, ..., \mathbf{x}_k$,

$$k^{-\frac{1}{\alpha}} \sum_{i=1}^{n} \mathbf{z}_i$$

converges to a stable distribution of index α. In order to show this, in the one dimensional case ($\theta = 0$) we study for example the distribution of

$$z = l'(x) = \frac{c}{x}.$$

Let us denote by $q(z)$ the distribution of z. If $z \to \infty$, then we have

$$q(z)dz = p(x)dx = z^{-\alpha-1}dz,$$

and hence

$$1 - P(z) = \int_z^{\infty} q(z)dz = z^{-\alpha}.$$

Therefore it follows (Feller, p. 303) that l' enters the domain of attraction of the stable distribution of index α. From the same reason, the matrix $\partial_i \partial_j l(\mathbf{x}, \theta)$ enters the domain of attraction of the stable distribution of index $\frac{\alpha}{2}$.

If we denote by $\varphi(\mathbf{x}, \theta)$ the characteristic function of the stable distribution to which z converges, then we have

$$\varphi(\mathbf{x}, \theta) = E[\exp\{it^j \partial_j l(\mathbf{x}, \theta).$$

Let us denote by $\tilde{F}(\mathbf{t}, \theta)$ the Minkowski norm induced by the this stable distribution (remark that $\mathbf{z}$ itself is not a stable distribution). We have to know the functions $\tilde{F}$ and F.

We have

$$\begin{aligned}\varphi(\varepsilon\mathbf{t}) &= E[\exp\{i\varepsilon t^j \partial_j l(\mathbf{x}, \theta)\}] \\ &= E[\exp\{i\Delta_{\varepsilon\mathbf{t}} l(\mathbf{x}, \theta)\}] \\ &= E[1 - \frac{1}{2}(\Delta_{\varepsilon\mathbf{t}} l)^2] \\ &= \exp\{-\frac{1}{2}E[(\Delta_{\varepsilon\mathbf{t}} l)^2]\},\end{aligned}$$

where

$$\Delta_{\varepsilon\mathbf{t}} l(\mathbf{x}, \theta) = l(\mathbf{x}, \theta + \varepsilon\mathbf{t}) - l(\mathbf{x}, \theta).$$

On the other hand, we have

$$\begin{aligned} D(\theta, \theta + \varepsilon\mathbf{t}) &= -8 \log \int \{p + \frac{1}{2}\Delta_{\varepsilon\mathbf{t}} p - \frac{1}{8}\frac{(\Delta p)^2}{p}\} dx \\ &= E[\{\Delta_{\varepsilon\mathbf{h}} l(\mathbf{x}, \theta)\}^2] = -E\{\Delta^2_{\varepsilon\mathbf{t}} l(\mathbf{x}, \theta)] \\ &= \varepsilon^\alpha F(\mathbf{t})^{\frac{\alpha}{2}}. \end{aligned}$$

Here, when we consider the expectation of $\Delta_{\varepsilon\mathbf{t}} l$, we integrate only on the intersection of the supports of $p(\mathbf{x}, \theta)$ and $p(\mathbf{x}, \theta + \varepsilon\mathbf{t})$. This implies that F and $\tilde{F}$ have same properties.

Let us consider here the m.l.e. (maximum likelihood estimator) $\hat{\theta}$. If we put

$$\hat{\theta} = \theta_0 + d\theta,$$

where θ_0 is the real value, for N measurements $\mathbf{x}_1, \ldots, \mathbf{x}_N$, we have

$$\sum_{k=1}^{N} \partial_i l(\mathbf{x}_k, \theta_0 + d\theta) = 0.$$

It follows

$$\begin{aligned} & N^{\frac{-1}{\alpha}} \sum \partial_i l(\mathbf{x}_k, \theta_0) = -N^{\frac{-2}{\alpha}} \sum \partial_i \partial_j l(\mathbf{x}_k, \theta_0) d\tilde{\theta}^j \\ & d\tilde{\theta}^j = N^{\frac{1}{\alpha}} d\theta^j. \end{aligned}$$

Since the LHS converges to the random variable z_i, and the coefficients of the RHS to w_{ij}, we get

$$d\tilde{\theta}^j = w^{ij} z_i$$

and $d\theta$ converges of order $N^{\frac{-1}{\alpha}}$. We have to study the properties of the random variable w_{ij}. (In the regular case, this converges to the matrix of constants g_{ij}).

We have to remark the fact that $d\theta$ is a dual vector, and it can be represented as a random variable. Asymptotically, we may have

$$E[F(d\tilde{\theta})] \geq n, \quad \text{or} \quad \{E[F(d\theta)^\alpha]\}^{\frac{1}{\alpha}} \geq n,$$

the equality may hold in the case of an efficient estimator. This can be regarded as the Cramer-Rao theorem in the non regular case. The equality may not be obtained by m.l.e., but from some L_α, $L_{\alpha'}$ norms asymptotically.

References

1. H. Rund, *The Differential Geometry of Finsler Spaces*, Springer-Verlag (1959).

2. K. Takeuchi, *The geometry of nonregular statistical models*, Technical report (1983).

Translator's Notes.

1. The paper explain that a stable distribution induces a Minkowski norm $F(\mathbf{t}) = \{A(\mathbf{t})\}^{\frac{1}{\alpha}}$.

Q1.1. What is the explicit formula for A?

Q1.2. Can we calculate the concrete form of the induced Minkowski norm for Normal, Cauchy and Levy distributions? (In the case of the Normal distribution we should obtain a Euclidean norm).

2. For a nonregular distribution it is possible to find a Finsler metric.

" Let f(x, y) be a two dimensional distribution density function whose support is concentrated on the unit sphere, that is, f(x,y) is non-zero inside the unit sphere of the x-y plane, and is zero outside. We assume that f is continuous.

Let (u, v) be the parameter which specifies the probability density function by p(x, y; u, v) = f(x - u, y - v).

That is, the family p(x, y; u,v) is a shift family, whose location is specified by u and v. When f is not differentiable at the support, the Fisher information diverges to infinity in general.

The set of these distributions is a smooth two-dimensional manifold whose admissible coordinates are (u, v). We can define the tangent space. But we cannot define the Riemannian metric by the Fisher information matrix.

We can think about the Hellinger distance such that between the two distributions, p(x,y; u,v) and p(x, y; u +du, v+dv). We cannot expand this in the quadratic form of (du, dv). However, rescaling the length, we can define a Finsler metric or the indicatrix, which gives a Finsler structure to the manifold" (From a letter of the author to Z. Shen, July 2002).

Q2.1. How to construct concrete examples of Finsler metrics for concrete nonregular distributions?

Q2.2. How are these metrics related with the Minkowski norms induced by stable distributions?