86. *Information and Statistics.* II

By Yukiyosi KAWADA
Faculty of Science, University of Tokyo

(Communicated by Shokichi IYANAGA, M. J. A., Oct. 12, 1987)

This is a continuation of Kawada [0]. We use the same notations.

II. ***L*-sets and informations.** **1.** Let $\boldsymbol{p}=(p_1,\cdots,p_m)$ and $\boldsymbol{q}=(q_1,\cdots,q_m)$ be probability distributions. We call the set

$$(9)\qquad L(\boldsymbol{p},\boldsymbol{q})=\left\{(x,y)\,\middle|\,x=\sum_{k=1}^{m}\alpha_k p_k,\ y=\sum_{k=1}^{m}\alpha_k q_k,\ 0\leqq\alpha_k\leqq 1,\ k=1,\cdots,m\right\}$$

the *Liapunov-set* (simply *L-set*) of the pair $(\boldsymbol{p},\boldsymbol{q})$. See Kudō [6], [7].

$L(\boldsymbol{p},\boldsymbol{q})$ has the following properties:

(i) $L(\boldsymbol{p},\boldsymbol{q})=\Delta$ (the diagonal segment joining $(0,0)$ and $(1,1)$) if and only if $\boldsymbol{p}=\boldsymbol{q}$.

(ii) $L(\boldsymbol{p},\boldsymbol{q})$ contains the points $(0,0)$ and $(1,1)$.

(iii) $L(\boldsymbol{p},\boldsymbol{q})$ is contained in the square $[0,1]\times[0,1]$.

(iv) $L(\boldsymbol{p},\boldsymbol{q})$ is a symmetric convex set with the center $(1/2,1/2)$.

(v) Let the indices of (p_k,q_k) be so substituted that

$$0\leqq(q_1/p_1)\leqq(q_2/p_2)\leqq\cdots\leqq(q_m/p_m)\leqq\infty$$

holds. Then

$$L(\boldsymbol{p},\boldsymbol{q})=\{(x,y)\,|\,\varphi(x)\leqq y\leqq\psi(x),\ 0\leqq x\leqq 1\}$$

where $\varphi(x)$ is a polygon with $m+1$ vertices

$$(0,0),\ (p_1,q_1),\ (p_1+p_2,q_1+q_2),\cdots,(p_1+\cdots+p_{m-1},\ q_1+\cdots+q_{m-1}),\ (1,1)$$

and $\psi(x)$ is a polygon with $m+1$ vertices

$$(0,0),\ (p_m,q_m),\ (p_m+p_{m-1},q_m+q_{m-1}),\cdots,$$
$$(p_m+p_{m-1}+\cdots+p_2,\ q_m+q_{m-1}+\cdots+q_2),\ (1,1).$$

Theorem 6. *A function $I(\boldsymbol{p},\boldsymbol{q})$ for any pair of finite probability distributions $(\boldsymbol{p},\boldsymbol{q})$ is an information if and only if*

(i) $L(\boldsymbol{p},\boldsymbol{q})=\Delta\Rightarrow I(\boldsymbol{p},\boldsymbol{q})=0$,

(ii) $L(\boldsymbol{p},\boldsymbol{q})=L(\boldsymbol{p}',\boldsymbol{q}')\Rightarrow I(\boldsymbol{p},\boldsymbol{q})=I(\boldsymbol{p}',\boldsymbol{q}')$,

(iii) $L(\boldsymbol{p},\boldsymbol{q})\supsetneqq L(\boldsymbol{p}',\boldsymbol{q}')\Rightarrow I(\boldsymbol{p},\boldsymbol{q})>I(\boldsymbol{p}',\boldsymbol{q}')$.

Namely, an information I is characterized by the property that I is a monotone functional of the family of all L-sets with $I=0$ for $L=\Delta$.

2. (i) We can characterize a fundamental information I geometrically as

$$(10)\qquad I_K(\boldsymbol{p},\boldsymbol{q})=\int_C K(d\varphi/dx)dx$$

where $K(x)$ is a non-negative differentiable function with $K(1)=K'(1)=0$, $K''(x)>0$, $\varphi(x)$ is the polygon defined as above and the integral is the curvilinear integral along the polygon $C: y=\varphi(x)$.

In particular, if we put

$$K(x)=\sqrt{1+x^2}-(x+1)/\sqrt{2},$$

then $I_K(p, q)=$(the length of the polygon C)$-\sqrt{2}$.

Thus we call the information (10) of the *type of arc-length.*

We can define several other types of informations geometrically.

(ii) Type of area of L-sets. Let

$$I_A(p, q)=\text{the area of } L\text{-set } L(p, q). \tag{11}$$

Then I_A is an information by Theorem 6. We can write also

$$I_A(p, q)=\left(\sum_{k=1}^{m}\sum_{l=1}^{m}|p_k q_l - p_l q_k|\right)\Big/2.$$

If we take a continuous function $f(x, y)$ defined on $0\leqq x\leqq 1$, $0\leqq y\leqq 1$ and positive for $0<x<1$, $0<y<1$, then

$$I_{A,f}(p, q)=\iint_{L(p,q)} f(x, y)dx\,dy \tag{12}$$

is also an information by Theorem 6. We call these informations $I_{A,f}$ of the *tpye of area.*

(iii) Type of breadth of L-sets. Let $B_\theta(L)$ be the breadth of the convex set $L(p, q)$ in the direction θ with the x-axis, and $f(\theta)$ $(0\leqq\theta<\pi)$ be any positive continuous function. Then

$$I_{B,f}(p, q)=\frac{1}{\pi}\int_0^\pi (B_\theta(L)-B_\theta(\Delta))f(\theta)d\theta \tag{13}$$

is an information by Theorem 6. We call these informations $I_{B,f}$ of the *type of breadth.* Notice that $d(p, q)=B_{3\pi/4}(L)$.

3. Now we introduce the concept of completeness of a family of informations after Kudō [7].

Definition 3. A family of informations $\{I_\omega(p, q)\,|\,\omega\in\Omega\}$ is called *weakly complete* if $I_\omega(p, q)=I_\omega(p', q')$ for all $\omega\in\Omega$ implies $L(p, q)=L(p', q')$, and is called *strongly complete* if $I_\omega(p, q)\geqq I_\omega(p', q')$ for all $\omega\in\Omega$ implies $L(p, q)\supset L(p', q')$.

Any strongly complete family is evidently weakly complete, but the converse does not hold in general as shown by a counter example by K. Iseki in [7].

Theorem 7. (i) *The family of informations of the type of area:*

$$\left\{I_A^{(i,j)}(p, q)=\iint_{L(p,q)} x^i y^j dx\,dy\,\middle|\, i, j=0, 1, 2, \cdots\right\}$$

is weakly complete, but not strongly complete.

(ii) *The family of informations of the type of area*

$$\left\{I_{A,f}(p, q)=\iint_{L(p,q)} f(x, y)dx\,dy\,\middle|\, f(x, y)\geqq 0 \text{ and continuous}\right\}$$

is strongly complete.

Theorem 8. (i) *The family of informations of the type of arc-length (i.e. fundamental informations)*

$$\{I^\lambda(p, q)\,|\,\alpha<\lambda<\alpha+\varepsilon\} \qquad (\alpha>0,\ \varepsilon>0)$$

and

$$\{I^{-\mu}(p, q)\,|\,\alpha-\varepsilon<\mu<\alpha\} \qquad (1/2\geqq\alpha>\varepsilon>0)$$

are both weakly complete, but not strongly complete.

(ii) *The family of informations of the type of arc-length*

$$\left\{I_K(\boldsymbol{p},\boldsymbol{q})=\sum_{k=1}^{m} p_k K(q_k/p_k)\,|\,K(1)=K'(1)=0,\ K''(x)>0 \text{ for } x>0\right\}$$

is strongly complete.

III. **Applications to statistics.** **1.** H. Akaike [1], [2] established the theory of AIC (Akaike information criterion), whose direct application gives a method of model selection from the standpoint of prediction. There he used as a basic tool the Kullback-Leibler information I_{KL}. Here we shall show that a similar results can be obtained if we use a regular information I, which we shall define below, instead of I_{KL}.

Definition 4. An information I is called *regular* if the following two conditions (A) and (B) hold.

(A) Let $\boldsymbol{p}=(p_1,\cdots,p_m)$, $\boldsymbol{q}=(q_1,\cdots,q_m)$ and $\boldsymbol{q}^0=(q_1^0,\cdots,q_m^0)$ be finite probability distributions, and put

$$p_k=q_k^0+u_k,\quad q_k=q_k^0+v_k\quad (k=1,\cdots,m)$$
$$u_1+\cdots+u_m=0,\qquad v_1+\cdots+v_m=0.$$

For $|u_k|<\varepsilon$, $|v_k|<\varepsilon$ $(k=1,\cdots,m)$ $I(\boldsymbol{p},\boldsymbol{q})$ is three times differentiable with respect to $(u_1,\cdots,u_{m-1},v_1,\cdots,v_{m-1})$ and

(14) $$I(\boldsymbol{p},\boldsymbol{q})=\frac{\alpha}{2}\sum_{k=1}^{m}\frac{1}{q_k^0}(u_k-v_k)^2+R,\qquad R=0(\varepsilon^3)$$

holds, where α is a positive constant. α is called the *invariant* of I.

(B) If we fix $\boldsymbol{q}$ then for any $\boldsymbol{p}$ the inequality

$$0\leqq I(\boldsymbol{p},\boldsymbol{q})\leqq c(\boldsymbol{q})$$

holds, where $c(\boldsymbol{q})$ is a certain constant. By (8) any differentiable fundamental information satisfies the condition (A), and we can easily verify that $I^\lambda(-(1/2)<\lambda<\infty)$ satisfies the condition (B).

Let $\boldsymbol{q}^0=(q_1^0,\cdots,q_m^0)$ be a probability distribution on m events $(E_1,\cdots,E_m)$. Suppose that the events $E_1,\cdots,E_m$ occur $N_1,\cdots,N_m$ times respectively in n $(n=N_1+\cdots+N_m)$ independent trials, and put

(15) $$\boldsymbol{P}=(N_1/n,\cdots,N_m/n).$$

Theorem 9. *Let I be a regular information with the invariant α. Then as $n\to\infty$ the random variable $(2n/\alpha)I(\boldsymbol{p},\boldsymbol{q}^0)$ converges in distribution to the chi-square distribution χ^2_{m-1} with $m-1$ degrees of freedom. Moreover,*

$$\lim_{n\to\infty}(2n/\alpha)E(I(\boldsymbol{P},\boldsymbol{q}^0))=m-1$$

holds, where E means the expectation of the random variable.

2. Now suppose that we are given a family of distributions $\boldsymbol{q}(\theta)=(\boldsymbol{q}_1(\theta),\cdots,\boldsymbol{q}_m(\theta))$, $\theta=(\theta_1,\cdots,\theta_r)$ $(\theta\in\Omega^{(r)})$ with r continuous parameters. We assume that the unknown true probability distribution $\boldsymbol{q}^0$ is contained in this family as $\boldsymbol{q}^0=\boldsymbol{q}(\theta^0)$, $\theta^0=(\theta_1^0,\cdots,\theta_r^0)$. We define the random probability distribution $\boldsymbol{P}$ by (15) after n independent trials. For this value $\boldsymbol{P}$, choose the value of parameters $\hat\theta=(\hat\theta_1,\cdots,\hat\theta_r)$ such that $I(\boldsymbol{P},\boldsymbol{q}(\theta))$ takes its minimum at $\theta=\hat\theta$. We can consider $\hat\theta$ also as a random vector.

Theorem 10. *As $n\to\infty$ the random vector.*
$$\sqrt{n}(\hat{\theta}_1-\theta_1^0, \cdots, \hat{\theta}_r-\theta_r^c)$$
converges in distribution to the r-dimensional normal distribution $N((0, \cdots, 0), \alpha I^{-1})$ with the mean vector $(0, \cdots, 0)$ and the variance matrix αI^{-1}, where
$$I=\left(\left(\frac{\partial^2 I}{\partial\theta_i\partial\theta_j}\right)_{\theta=\theta^0}\right)_{i,j=1,\cdots,r}=\alpha Q\cdot {}^tQ,$$
$$Q=(q_j^{(i)}/\sqrt{q_j})_{\theta=\theta^0}, \qquad q_j^{(i)}=\partial q_j/\partial\theta_i.$$

Theorem 11. *As $n\to\infty$ the random variable $(2n/\alpha)(I(\boldsymbol{P}, \boldsymbol{q}^0)-I(\boldsymbol{P}, \boldsymbol{q}(\hat{\theta})))$ converges in distribution to the chi-square distribution χ_r^2 with r degrees of freedom, and $(2n/\alpha)(I(\boldsymbol{P}, \boldsymbol{q}(\hat{\theta}))$ itself converges in distribution to the chi-square distribution χ_{m-1-r}^2 with $m-1-r$ degrees of freedom.*

3. Now let $\boldsymbol{q}^0=(q_1^0, \cdots, q_m^0)$ be unknown true probability distribution of the events $(E_1, \cdots E_m)$, and suppose that we obtain the events $E_1, \cdots, E_m$, $n_1, \cdots, n_m$ times respectively in n $(n=n_1+\cdots+n_m)$ independent trials. Put $\boldsymbol{p}^0=(n_1/n, \cdots, n_m/n)$.

Let $\Omega^{(r)}=\{\boldsymbol{q}(\theta)=(q_1(\theta), \cdots, q_m(\theta)\}$ be a model for $\boldsymbol{q}^0$ which contains $\boldsymbol{q}^0=\boldsymbol{q}(\theta^0)$. Assume that I is a regular information, and $\hat{\theta}$ is the value of θ in a neighbourhood of θ^0 such that $I(\boldsymbol{p}^0, \boldsymbol{q}(\hat{\theta}))$ is the minimum. Now define

(16) $$AIC(\Omega^{(r)})=(2n/\alpha)I(\boldsymbol{p}^0, \boldsymbol{q}(\hat{\theta}))+2r$$

after Akaike [1], [2]. Akaike's method of selection of model is as follows.

Suppose we are given several models for $\boldsymbol{q}^0$. i.e. $\Omega^{(r_1)}, \cdots, \Omega^{(r_s)}$. After n independent trials we obtain $\boldsymbol{p}^0$ as above. Compare the values $AIC(\Omega^{(r_t)})$ $(t=1, \cdots, s)$. Choose the model $\Omega^{(r_t)}$ for which $AIC(\Omega^{(r_t)})$ takes the minimum among s values.

This method depends on the following theorem in prediction theory. Namely, we repeat n^* new independent trials, for which the events $E_1, \cdots, E_m$ occur $N_1^*, \cdots, N_m^k$ times $(n^*=N_1^*+\cdots+N_m^*)$ respectively. Put
$$\boldsymbol{P}^*=(N_1^*/n^*, \cdots, N_m^*/n^*).$$
The mean value $E^*(I(\boldsymbol{P}^*, \boldsymbol{q}(\hat{\theta}))$ may be called the mean information in prediction.

Theorem 12.
$$AIC(\Omega^{(r)})=(2n/\alpha)E^*(I(\boldsymbol{P}^*, \boldsymbol{q}(\hat{\theta}))+R_1+R_2$$
where
$$R_1=(2n/\alpha)(I(\boldsymbol{p}^0, \boldsymbol{q}^0)-E^*(I(\boldsymbol{P}^*, \boldsymbol{q}^0))$$
depends only on the value $\boldsymbol{p}^0$, and R_2 is a random variable with $E(R_2)=0$.

Reference*)

[0] Y. Kawada: Information and statistics. I. Proc. Japan Acad., 63A, 281–284 (1987).

*) Other references are given in [0].