Kybernetika

Jiří Michálek

Maximum likelihood principle and I-divergence: discrete time observations

Kybernetika, Vol. 34 (1998), No. 3, [265]--288

Persistent URL: http://dml.cz/dmlcz/135207

KYBERNETIKA — VOLUME 34 (1998), NUMBER 3, PAGES 265 - 288

MAXIMUM LIKELIHOOD PRINCIPLE AND I-DIVERGENCE: DISCRETE TIME OBSERVATIONS[1]

JIŘÍ MICHÁLEK

The paper investigates the relation between maximum likelihood and minimum I-divergence estimates of unknown parameters and studies the asymptotic behaviour of the likelihood ratio maximum. Observations are assumed to be done in the discrete time.

INTRODUCTION

In the monograph by Kullback [4] one can find an interesting relation between the maximum likelihood estimate of an unknown parameter and the I-divergence of two probability measures where one of them is determined by the value of the MLE in question. It will be best to describe this relation by a simple example used in the monograph mentioned above.

Let $x_1, x_2, \ldots, x_n$ be a sample from the Gaussian family $N(\mu, \sigma^2)$, where the parameter (μ, σ^2) is unknown. Let $\overline{x}$ be the arithmetic mean and

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Further, let (μ_0, σ_0^2) be an arbitrary element of the parametric space $\Theta = R_1 \times R_1^+$. At this moment it is necessary to mention the notion of I-divergence. Let P, Q be two probability measures defined on a measurable space, let f_P, f_Q be their corresponding Radon–Nikodym derivatives with respect to a dominating σ-finite measure μ. Then the I-divergence between P and Q is defined as

$$I(P:Q) = I(f_P : f_Q) = \int f_P(x) \ln \frac{f_P(x)}{f_Q(x)}\, \mu(\mathrm{d}x).$$

Let now f_Q be fixed and let us look for the minimum of I-divergence over f_P under the constrain

$$\theta = \int T(x)\, f_P(x)\, \mu(\mathrm{d}x),$$

[1] This work was supported by the Grant Agency of the Czech Republic under Grant 201/96/415.

where $T(x)$ is a statistic, θ is a fixed value from the parametric space Θ. In our case $T(x_1, \ldots, x_n) = (\overline{x}, s^2)$ and $\theta = (\mu, \sigma^2)$. The optimal density function $f^*(x)$ is of the exponential type, namely

$$f^*(x) = \frac{e^{\tau(\theta)\,T(x)}\, f_Q(x)}{M(\tau(\theta))}$$

where $M(\tau(\theta)) = \int e^{\tau(\theta)\,T(x)}\, f_Q(x)\,\mathrm{d}x$ is assumed to be finite. As we do not know the true value of θ, it is reasonable to substitute it by its maximal likelihood estimate and in this way to obtain an estimate $\hat{I}(f^*, f_Q)$ of the minimum of I-divergence $I(f^* : f_Q)$. In our example with $f_Q \sim N(\mu_0, \sigma_0^2)$

$$\hat{I}(f^* : f_Q) = \frac{n(\overline{x} - \mu_0)}{2\sigma_0^2} + \frac{n}{2}\left(\frac{s^2}{\sigma_0^2} - \ln\frac{s^2}{\sigma_0^2} - 1\right).$$

But this estimate is nothing else but the I-divergence between two Gaussian measures, namely

$$I(N(\overline{x}, s^2)) : N(\mu_0, \sigma_0^2))$$

multiplied by n. Here we can see a very close connection between the MLE and the I-divergence. Kullback shows also in Chapter 5 of his monograph that the estimate for the minimum of I-divergence can be expressed as

$$\hat{I}(f^* : f_Q) = \hat{\theta}\,\tau(\hat{\theta}) - \log M(\tau(\hat{\theta})) = \ln\frac{\sup_\theta f^*(x)}{f_Q(x)}$$

where $\hat{\theta}$ is the MLE for the parameter θ and $f^*(\cdot)$ is the density of the exponential type derived from the underlying density function $f_Q(\cdot)$. Kullback uses this relation in the case of in discrete time observations only. Motivating by this interesting example we will state the aim of the paper. The main goal of this paper is to study this relation between the MLE and the I-divergence in the general case of dependent observations. Some results discussed in the paper are not principially new and they are presented in order to see the relation between the maximum likelihood method and minimal I-divergence method.

1. EXPONENTIAL FAMILIES

In this part we will restrict ourselves to the case of i.i.d. random variables generated by the exponential family, i.e. with the density function

$$f(x, \theta) = C(\theta)\, h(x)\, e^{\tau(\theta)\,T(x)},$$

where θ is a parameter from Θ. We assume, of course,

$$\int_{-\infty}^{\infty} C(\theta)\, h(x)\, e^{\tau(\theta)\,T(x)}\,\mathrm{d}x = 1$$

for each $\theta \in \Theta$. We will consider the case $\Theta \subset R_1$ only.

If $x_1, x_2, \ldots, x_n$ is a sequence of observations coming from this family then the logarithm of the likelihood ratio equals

$$\ln \prod_{j=1}^{n} f(x_j, \theta) = n \ln C(\theta) + \sum_{j=1}^{n} h(x_j) + \tau(\theta) \sum_{j=1}^{n} T(x_j).$$

The MLE $\hat{\theta}_n$ must satisfy the following condition

$$\ln \prod_{j=1}^{n} f(x_j, \hat{\theta}_n) = \max_{\theta \in \Theta} \ln \prod_{j=1}^{n} f(x_j, \theta).$$

Under the existence of appropriate derivatives, the MLE $\hat{\theta}_n$ is given by

$$\frac{1}{n} \sum_{j=1}^{n} T(x_j) = \frac{1}{C(\hat{\theta}_n)} C'(\hat{\theta}_n) \frac{1}{\tau'(\hat{\theta}_n)}.$$

This relation says that $\hat{\theta}_n$ exists if and only if

$$\frac{1}{n} \sum_{1}^{n} T(x_j) \in \operatorname*{Range}_{\theta \in \Theta} \pi(\theta),$$

where $\pi(\theta) = [\ln C(\theta)]' [\tau'(\theta)]^{-1}$. Therefore it is necessary to assume $\tau'(\theta) \neq 0$ for each $\theta \in \Theta$. The law of large numbers immediately gives

$$\frac{1}{n} \sum_{1}^{n} T(x_j) \underset{n \to \infty}{\longrightarrow} \mathsf{E}_{\theta^*} \{T(x)\} \quad \text{a.s.}$$

where θ^* is a true parameter value. As for each $\theta \in \Theta$

$$\mathsf{E}_{\theta} \{T(x)\} = \pi(\theta),$$

then the strong consistency of $\left\{\frac{1}{n} \sum_1^n T(x_j)\right\}_{n=1}^{\infty}$ implies

$$\mathsf{P} \left\{ \omega : \frac{1}{n} \sum_{j=1}^{n} T(x_j(\omega)) \in \operatorname*{Range}_{\theta \in \Theta} \pi(\theta) \right\} \underset{n \to \infty}{\longrightarrow} 1.$$

Now we mention a close connection with the I-divergence. Let $\theta_0 \in \Theta$. Then

$$I(\theta : \theta_0) = \mathsf{E}_{\theta} \left\{ \ln \frac{f(x, \theta)}{f(x, \theta_0)} \right\} = \ln \frac{C(\theta)}{C(\theta_0)} + (\tau(\theta) - \tau(\theta_0)) \frac{1}{\tau'(\theta)} [\ln C(\theta)]'.$$

We can continue and prove the following relation

$$\frac{1}{n} \sum_{1}^{n} \ln f(x_j, \theta) \quad = \quad - \ln \frac{C(\hat{\theta}_n)}{C(\theta)} + \frac{C'(\hat{\theta}_n)}{C(\hat{\theta}_n)} \frac{(\tau(\theta) - \tau(\hat{\theta}_n))}{\tau'(\hat{\theta}_n)}$$

$$
\begin{aligned}
&+\ln C(\hat{\theta}_n)+\frac{1}{n}\sum_1^n h(x_j)+\tau(\hat{\theta}_n)\frac{1}{n}\sum_{j=1}^n T(x_j)\\
&= -I(\hat{\theta}_n:\theta)+\ln C(\hat{\theta}_n)+\frac{1}{n}\sum_1^n \ln h(x_j)+\tau(\hat{\theta}_n)\frac{1}{n}\sum_1^n T(x_j)\\
&= -I(\hat{\theta}_n:\theta)-H(\hat{\theta}_n),
\end{aligned}
$$

where $H(\theta)=-\mathrm{E}_\theta\{\ln f(x,\theta)\}$. In the case of $\theta=\hat{\theta}_n$ the corresponding mean value is calculated with respect to the empirical distribution function $\hat{F}(\cdot)$, i.e.

$$
\begin{aligned}
E_{\hat{\theta}_n}\{\ln f(x,\hat{\theta}_n)\} &= \int_{-\infty}^{\infty}\Big(\ln C(\hat{\theta}_n)+\ln h(x)+\tau(\hat{\theta}_n)T(x)\Big)\,\mathrm{d}\hat{F}(x)\\
&= \ln C(\hat{\theta})+\frac{1}{n}\sum_1^n T(x_j).
\end{aligned}
$$

Using these facts we see that

$$
\max_{\theta\in\Theta}\frac{1}{n}\sum_{j=1}^n \ln f(x_j,\theta)=-\min_{\theta\in\Theta} I(\hat{\theta}_n:\theta)-H(\hat{\theta}_n)=-H(\hat{\theta}_n).
$$

if $\hat{\theta}_n\in\Theta$. In this way we have proved that the MLE $\hat{\theta}_n$ equals the estimate of unknown parameter θ obtained by minimizing the corresponding I-divergence. Further, we can assert

$$
\begin{aligned}
\lim_{n\to\infty}\frac{1}{n}\max_{\theta\in\Theta}\sum_{j=1}^n \ln f(x_j,\theta) &= -\lim_{n\to\infty} H(\hat{\theta}_n)\\
&= \int_{-\infty}^{\infty}\ln f(x,\theta_*)\,f(x,\theta_*)\,\mathrm{d}x=-H(\theta_*)\quad \text{a.s.},
\end{aligned}
$$

if θ_* is a true value and $H(\theta_*)$ exists. When we evaluate the likelihood ratio we can write

$$
\sum_{j=1}^n \ln\frac{f(x_j,\theta)}{f(x_j,\theta_0)}=-n\,I(\hat{\theta}_n:\theta)+n\,I(\hat{\theta}_n:\theta_0).
$$

This relation immediately gives

$$
\max_{\theta\in\Theta}\frac{1}{n}\sum_{j=1}^n \ln\frac{f(x_j,\theta)}{f(x_j,\theta_0)}=I(\hat{\theta}_n:\theta_0).
$$

Using the properties of MLE we can similarly state that

$$
\lim_{n\to\infty}\max_{\theta\in\Theta}\frac{1}{n}\sum_{j=1}^n \ln\frac{f(x_j,\theta)}{f(x_j,\theta_0)}=I(\theta_*:\theta_0)\quad \text{a.s.}
$$

Now, let us write $C(\theta) = e^{A(\theta)}$, then

$$\sup_{\theta\in\Theta} \ln \sum_{j=1}^{n} \frac{f(x_j,\theta)}{f(x_j,\theta_0)} = n\left(A(\hat\theta_n) - A(\theta_0) - A'(\hat\theta_n)\cdot\frac{\tau(\hat\theta_n)-\tau(\theta_0)}{\tau'(\hat\theta_n)}\right).$$

Using Taylor's expansion we can express

$$\begin{aligned} A(\theta_0) - A(\hat\theta_n) &= A'(\hat\theta_n)(\theta_0-\hat\theta_n) + \frac{A''(\hat\theta_n)}{2}(\theta_0-\hat\theta_n)^2 + o((\theta_0-\hat\theta_n)^2) \\ \tau(\theta_0) - \tau(\hat\theta_n) &= \tau'(\hat\theta_n)(\theta_0-\hat\theta_n) + o(|\theta_0-\hat\theta_n|). \end{aligned}$$

Then

$$\sup_{\theta\in\Theta} \ln \sum_{j=1}^{n} \frac{f(x_j,\theta)}{f(x_j,\theta_0)} = n\left\{\frac{(\hat\theta_n-\theta_0)^2}{2}A''(\hat\theta_n) + o(|\theta_0-\hat\theta_n|)\right\}.$$

When $\hat\theta_n \to \theta_*$ a.s. then we obtain the following behaviour of the likelihood ratio maximum

$$\lim_{n\to\infty}\frac{1}{n}\sup_{\theta\in\Theta}\ln\sum_{j=1}^{n}\frac{f(x_j,\theta)}{f(x_j,\theta_0)} = \frac{(\theta_*-\theta_0)^2}{2}A''(\theta_*) + o(|\theta_*-\theta_0|).$$

This result implies that the rate of the loglikelihood ratio maximum behaves like the Euclidean distance and if the MLE $\hat\theta_n$ is asymptotically normal then this rate will be asymptotically a noncentral χ^2-distribution with noncentrality depending on the difference $(\theta_*-\theta_0)^2$.

Hence the statistic $I(\hat\theta_n : \theta_0)$ can be used for testing the null hypothesis $\theta = \theta_0$ against the alternative hypothesis $\theta \neq \theta_0$. If the value of $I(\hat\theta_n : \theta_0)$ is far from zero then the null hypothesis is rejected. But it is necessary to have in mind that our decision is strongly dependent on the distance $I(\theta_* : \theta_0)$. About the asymptotic behaviour of $I(\hat\theta_n : \theta_*)$ in the i.i.d. case the reader can be referred to Kupperman [5] where under some regularity conditions

$$2n\, I(\hat\theta_n : \theta_*)$$

behaves asymptotically with $\chi^2(1)$ distribution if the parameter θ is one only and the null hypothesis is true. Extension of this result to multivariate parameters, and also to divergences different from the I-divergences, can be found in Morales et al [9].

In the case where the observations $x_1, x_2, \ldots, x_n$ are from a discrete distribution function we can say even more about the relation between the MLE and I-divergence. Suppose that $f(x,\theta)$ is a density on a finite set $\mathcal{X}$, i.e.

$$P_\theta\{X = x\} = f(x,\theta).$$

For simplicity let $f(x,\theta) > 0$ for each $x \in \mathcal{X}$ and each $\theta \in \Theta$. Then the joint distribution function

$$f(x_1,\ldots,x_n,\theta) = \prod_{j=1}^{n} f(x_j,\theta) = \exp\left\{n\sum_{a\in\mathcal{X}}\frac{N_x(a)}{n}\ln f(a,\theta)\right\}$$

$$= \exp\left\{-n\left(-\sum_{a\in\mathcal{X}} R_x(a)\ln R_x(a)\right)\right\}\exp\left\{-n\sum_{a\in\mathcal{X}} R_x(a)\ln\frac{R_x(a)}{f(a,\theta)}\right\},$$

where $R_x(\cdot)$, $x=(x_1,x_2,\ldots,x_n)$ is the empirical probability mass function defined by the relative frequencies

$$R_x(a)=\frac{N_x(a)}{n},\quad a\in\mathcal{X}$$

where $N_x(a)$ counts the number of occurrences of $a\in\mathcal{X}$ in the sequence $x_1,x_2,\ldots,x_n$.

Then we get to the expression

$$\frac{1}{n}\sum_{j=1}^{n}\ln f(x_j,\theta)=-I(R_x(\cdot)|\theta)-H(R_x(\cdot)),$$

where $I(R_x(\cdot)|\theta)$ is the I-divergence information between the empirical mass function $R_x(\cdot)$ and the theoretical $f(\cdot|\theta)$. The quantity $H(R_x(\cdot))$ is the Shannon entropy of $R_x(\cdot)$. By continuity, we set in the above formulas $\ln\frac{0}{n}=0$, $n\ln\frac{n}{0}=\infty$, $0\ln 0=0$.

The above expression gives immediately that the MLE of θ is nothing else but the estimate minimizing the distance $I(R_x(\cdot)|\theta)$. We see that in a discrete case of an exponential family we can write

$$\frac{1}{n}\sum_{j=1}^{n}\ln f(x_j,\theta)=-I(R_x(\cdot)|\theta)-H(R_x(\cdot))=-I(\hat\theta_n|\theta)-H(\hat\theta_n).$$

Hence

$$\max_{\theta\in\Theta}\frac{1}{n}\sum_{j=1}^{n}\ln f(x_j,\theta)=-H(\hat\theta_n)=-\min_{\theta\in\Theta}I(R_x(\cdot)|\theta)-H(R_x(\cdot)).$$

As $I(R_x(\cdot)|\theta)\geq 0$, we have

$$H(\hat R_{\boldsymbol{x}}(\cdot))\leq H(\hat\theta_n).$$

If there exists $\theta(\boldsymbol{x})\in\Theta$ satisfying

$$f(\cdot,\theta(\boldsymbol{x}))=\hat R_{\boldsymbol{x}}(\cdot),$$

then $\min_{\theta\in\Theta} I(\hat R_{\boldsymbol{x}}(\cdot)|\theta)=0$ and

$$H(\hat\theta_n)=H(\hat R_{\boldsymbol{x}}(\cdot)).\qquad(*)$$

This property leads us to the following definition.

We say that an exponential family $\{f(x,\theta),\theta\in\Theta\}$ is complete if there exists $\theta\in\Theta$, $\theta=\theta(\boldsymbol{x})$ such that

$$f(\boldsymbol{x},\theta)=R(\boldsymbol{x})\quad\text{a.s.}$$

Under this property the equality $(*)$ holds. Unfortunately, this relation cannot be extended to continuous models because of the absence of analogy with the empirical mass function in the discrete case.

The following example shows that the relation connecting MLE and I-divergence need not be valid in general. As a counter-example we can take the Cauchy distribution function with density

$$p(x) = \frac{1}{\pi} \frac{\delta}{\delta^2 + (x-\mu)^2}$$

where $\delta > 0$ and $\mu \in R_1$. For the sake of simplicity we put $\mu = 0$. Then the MLE $\hat{\delta}_n$ must satisfy the equation

$$\sum_{j=1}^{n} \frac{x_j^2 - \delta^2}{x_j^2 + \delta^2} = 0.$$

Further, we need the corresponding I-divergence

$$I(p(\cdot,\delta) : p(\cdot,1)) = \ln\left\{\frac{(1+\delta)^2}{4\delta}\right\}.$$

It is sufficient to show that the considered relation is not valid for $n = 2$. Then the MLE $\hat{\delta}_2$ satisfies the equation

$$\frac{x_1^2 - \delta^2}{x_1^2 + \delta^2} + \frac{x_2^2 - \delta^2}{x_1^2 + \delta^2} = 0$$

which gives $x_1^2 x_2^2 = \delta^4$, i.e. $\hat{\delta}_2 = \sqrt{|x_1 x_2|}$. If we substitute this expression into the formula of likelihood ratio we obtain

$$\max_{\delta>0} \sum_{j=1}^{2} \ln \frac{p(x_j,\delta)}{p(x_j,1)} = \ln\left\{\frac{(\hat{\delta}_2)^2 (1+x_1^2)(1+x_2^2)}{\left((\hat{\delta}_2)^2 + x_1^2\right)\left((\hat{\delta}_2)^2 + x_2^2\right)}\right\}.$$

If we use $\hat{\delta}_2$ for $I(\delta : 1)$, we get

$$I(p(\cdot,\hat{\delta}_2) : p(\cdot,1)) = \ln\left(\frac{(1+\hat{\delta}_2)^2}{4\hat{\delta}_2}\right).$$

At the first sight we see that these expressions are not proportional each other.

When we drop the independence among observations we obtain the family of random sequences and processes with the exponential type Radon–Nikodym derivatives. We say that a random process $\{x(t), t \geq 0\}$ has an exponential family of distributions if there exists a dominating measure P such that for each $t > 0$

$$\frac{\mathrm{d}P_t(\theta)}{\mathrm{d}P_t}(\omega) = a(t,\theta)\, q(t,\omega) \exp\left\{\sum_{j=1}^{M} \tau_j(t,\theta)\, B_j(t,\omega)\right\},$$

where θ is k-dimensional parameter, random processes $q(t,\omega)$ and $B_j(t,\omega)$, $j = 1,2,\ldots,M$ are $\mathcal{F}_t$-measurable, $\{\mathcal{F}_t\}$ is a system of nondecreasing σ-algebras. The measures $P(\theta)$ and P are defined on $\sigma\left(\bigcup_{t>0} \mathcal{F}_t\right)$ and P_t is the projection on $\mathcal{F}_t$. For

more details, see Küchler and Sorensen [6]. We will again consider for simplicity the one-dimensional case $k = 1$. We start with

$$\ln \frac{\mathrm{d}P_t(\theta)}{\mathrm{d}P_t} = \ln a(t,\theta) + \ln q(t,\omega) + \sum_{j=1}^{M} \tau_j(t,\theta)\, B_j(t,\omega).$$

If the appropriate derivatives exist, then the MLE $\hat{\theta}_t$ must satisfy the equation

$$\frac{a'(t,\hat{\theta}_t)}{a(t,\hat{\theta}_t)} + \sum_{j=1}^{M} \tau_j'(t,\hat{\theta}_t)\, B_j(t,\omega) = 0$$

because for each $\theta \in \Theta$

$$\mathsf{E}_\theta \left\{ \sum_{j=1}^{M} \tau_j'(t,\theta)\, B_j(t,\omega) \right\} = -(\ln a(t,\theta))'.$$

The entropy $H_t(\theta) = \mathsf{E}_\theta \left\{ \ln \frac{\mathrm{d}P_t(\theta)}{\mathrm{d}P_t} \right\}$, if exists, has the following form

$$H_t(\theta) = \ln a(t,\theta) + \mathsf{E}_\theta\{q(t,\omega)\} + \sum_{j=1}^{M} \tau_j(t,\hat{\theta}_t)\, \mathsf{E}_\theta\{B_j(t,\omega)\}.$$

For $\theta = \hat{\theta}_t$ we have

$$H_t(\hat{\theta}_t) = \ln a(t,\hat{\theta}_t) + \ln q(t,\omega) + \sum_{j=1}^{M} \tau_j(t,\hat{\theta}_j)\, B_j(t,\omega).$$

It is easy to show that

$$\ln \frac{\mathrm{d}P_t(\theta)}{\mathrm{d}P_t} = -I(\hat{\theta}_t : \theta) + H(\hat{\theta}_t)$$

which is a similar relation as in the i.i.d. case.

This relation immediately yields that the MLE $\hat{\theta}_t$ is also an estimate minimizing the corresponding I-divergence and

$$\max_{\theta \in \Theta} \ln \frac{\mathrm{d}P_t(\theta)}{\mathrm{d}P_t} = H(\hat{\theta}_t)$$

because $I(\hat{\theta}_t : \hat{\theta}_t) = 0$.

The most important case occurs if the parametric functions are not depending on time, i.e. $\tau_j(t,\theta) = \tau_j(\theta)$, $j = 1, 2, \ldots, M$ only. Then we will speak about the time-homogeneous exponential family and we can say more about the corresponding I-divergence. In many cases under *time-homogeneity* the logarithm of Radon–Nikodym derivative can be expressed as

$$\ln \frac{\mathrm{d}P_t(\theta)}{\mathrm{d}P_t(\theta_0)} = K(\theta) - K(\theta_0) + t(Q(\theta) - Q(\theta_0)) + \sum_{j=1}^{M} (\gamma_j(\theta) - \gamma_j(\theta_0))\, B_j(t) + o(t).$$

If for each $j = 1, 2, \ldots, M$ there exist limits

$$\lim_{t\to\infty} \frac{1}{t} B_j(t) = m_j(\theta) \quad \text{a.s. } [P(\theta)],$$

then there exists

$$\lim_{t\to\infty} \frac{1}{t} \ln \frac{\mathrm{d}P_t(\theta)}{\mathrm{d}P_t(\theta_0)} = (Q(\theta) - Q(\theta_0)) + \sum_{j=1}^{M} (\gamma_j(\theta) - \gamma_j(\theta_0))\, m_j(\theta).$$

This limit is called the asymptotic I-divergence rate $\overline{I}(P(\theta) : P(\theta_0))$ and then we can write

$$\frac{1}{t} \ln \frac{\mathrm{d}P_t(\theta)}{\mathrm{d}P_t(\theta_0)} = -\overline{I}(P(\theta) : P(\theta_0)) + o(1).$$

The estimates obtained by minimizing $\overline{I}(P(\theta) : P(\theta_0))$ are called asymptotic MLE's and in many cases they have similar asymptotic behaviour as the original MLE's. The main advantage of them is computational simplicity very often. The random processes $\left\{\frac{1}{t} B_j(t),\ j = 1, 2, \ldots, M\right\}$ then form an "asymptotic" sufficient statistic.

2. THE CHANGE POINT PROBLEM AND I-DIVERGENCE

Let us start with the simplest case of change detection in mean value of a normal population with constant and known dispersion. This problem was firstly studied by Page [10] and its solution is known as the Page–Hinkley test.

Hence, we have observations $y_1, y_2, \ldots, y_n$ mutually independent and satisfying the relation

$$y_j = \mu_j + e_j, \quad e_j \sim N(0, \sigma^2), \quad j = 1, 2, \ldots, n$$

where

$$\mu_j = \mu_0 \quad \text{for} \quad j = 1, 2, \ldots, r-1$$

$$\mu_j = \mu_1 \quad \text{for} \quad j = r, r+1, \ldots, n.$$

In the simplest case we assume the knowledge of μ_0, μ_1, σ^2 and the task is to estimate an unknown moment r of a change. When we accept the maximum likelihood principle then

$$\hat{r}_n = \operatorname*{argmax}_{1\le r\le n} \left\{ \prod_{j=1}^{r-1} p_0(y_j) \prod_{j=r}^{n} p_1(y_j) \right\},$$

where $p_i(\cdot)$ is the density function of $N(\mu_i, \sigma^2)$, $i = 1, 2$. This approach gives the well-known detector of a change

$$g_n = \max_{1\le r\le n} S_r^n(\mu_0, \mu_1) \gtrless \lambda,$$

where $S_r^n(\mu_0, \mu_1)$ are the so called cumulative sums

$$S_r^n(\mu_0, \mu_1) = (\mu_1 - \mu_0) \sum_{j=r}^{n} \left(y_j - \frac{\mu_1 + \mu_0}{2} \right)$$

with $\sigma^2 = 1$ for simplicity. A more realistic case occurs if we know a value μ_0 before a change but we do not know μ_1 after the change. This case can be solved in two ways. The first one means to substitute the unknown jump $\mu_1 - \mu_0$ in the Page–Hinkley test by a maximal acceptable jump, the second one means to use a maximum likelihood estimate $\hat{\mu}_1$ of μ_1. In this case the change detector has the form

$$\max_{1\le r\le n}\left(\max_{\mu_1} S_r^n(\mu_0,\mu_1)\right) \gtrless \lambda.$$

It is easy to prove that

$$\max_{\mu_1} S_r^n(\mu_0,\mu_1) = \frac{n-r+1}{2}(\overline{Y}_r^n - \mu_0)^2,$$

with

$$\overline{Y}_r^n = \frac{1}{n-r+1}\sum_{j=r}^{n} y_j,$$

which is the arithmetic mean of observations after the change. But, this maximum can be rewritten in a more interesting form, namely

$$\max_{\mu_1} S_r^n(\mu_0,\mu_1) = \frac{n-r+1}{2} I(p(\cdot,\overline{Y}_r^n) : p(\cdot,\mu_0)),$$

where $p(\cdot,\theta)$ is the density function of $N(\theta,1)$. Then the proposed detector of a change works in such a way that a change is detected with a high probability in that instant $\hat{r}$ where the largest discrepancy exists between probability density functions $p(\cdot,\mu_0)$ and $p(\cdot,\overline{Y}_r^n)$ measured via I-divergence.

Similarly for illustration we can proceed in the case when the observations $y_1, y_2, \ldots$ $\ldots, y_{r-1}$ before a change are from the population $N(\mu_0,\sigma_0^2)$ (we assume the knowledge of $\theta_0 = (\mu_0,\sigma_0^2)$) and the observations $y_r, y_{r+1}, \ldots, y_n$ are from $N(\mu,\sigma^2)$ after a change, but we know neither μ nor σ^2. A test of a change will be based on the statistics

$$\max_{(\mu,\sigma^2)} S_r^n((\mu,\sigma^2),(\mu_0,\sigma_0^2)), \quad r = 1,2,\ldots,n,$$

where $S_r^n(\cdot,\cdot)$ is a cumulative sum derived from the maximum likelihood principle, hence

$$S_r^n((\mu,\sigma^2),(\mu_0,\sigma_0^2)) = \frac{1}{n}(n+1-r)\ln\frac{\sigma_0^2}{\sigma^2} + \sum_{j=r}^{n}\left\{\frac{(y_j-\mu_0)^2}{\sigma_0^2} - \frac{(y_j-\mu)^2}{\sigma^2}\right\}.$$

As familiarly known the MLE's of (μ,σ^2) are for a fixed r

$$\hat{\mu}_r = \frac{1}{n+1-r}\sum_{j=r}^{n} y_j, \qquad \hat{\sigma}_r^2 = \frac{1}{n+1-r}\sum_{j=r}^{n}(y_j-\hat{\mu}_r)^2.$$

When we substitute these values into $S_r^n(\cdot,\cdot)$ instead of unknown (μ,σ^2) we obtain

$$\max_{(\mu,\sigma^2)} S_r^n((\mu,\sigma^2),(\mu_0,\sigma_0^2)) = \frac{n-r+1}{2}\left(\frac{(\hat{\mu}_n-\mu_0)^2}{\sigma_0^2} + \frac{\hat{\sigma}_r^2}{\sigma_0^2} - \ln\frac{\hat{\sigma}_r^2}{\sigma_0^2} - 1\right).$$

This result is very interesting because the maximum of the cumulative sum $S_r^n(\cdot,\cdot)$ after the change at the instant r can be expressed using I-divergence again. If we summarize, the test of change detection is given in the form

$$g_n = \max_{1\le r\le n} \frac{n+r-1}{2} I(p(\cdot,\hat{\mu}_r,\hat{\sigma}_r^2) : p(\cdot,\mu_0,\sigma_0^2)),$$

where $(\hat{\mu}_r,\hat{\sigma}_r^0)$ are MLE's based on the observations $y_r, y_{r+1},\dots,y_{n-1},y_n$ after a possible change.

Especially, a very interesting result arises when we consider a change in variability only, i.e. $\mu_0 = \mu = 0$ for simplicity. In this case the detection of a change is based on the statistic of the form

$$\frac{\hat{\sigma}_r^2}{\sigma^2} - \ln\frac{\hat{\sigma}_r^2}{\sigma^2} - 1.$$

Then the estimate $\hat{r}$ of the time of a change is given by

$$\hat{r} = \operatorname*{argmax}_{1\le r\le n} \left\{ \frac{n-r+1}{2}\left(\frac{\hat{\sigma}_r^2}{\sigma^2} - \ln\frac{\hat{\sigma}_r^2}{\sigma^2} - 1\right)\right\}.$$

A maximum likelihood test for the inspection of the behaviour of variability in the case of i.i.d. Gaussian random variables can be found in Krishnaiah and Miao [3], but without any remark about its close relation to the I-divergence.

On the basis of these particular cases one can suggest a test inspecting changes in unknown parameters based on the maximum likelihood principle as follows. Let $p(\cdot,\theta)$ be a probability density function depending on a parameter $\theta \in \Theta$. Let $x_1, x_2,\dots,x_{r-1}$ be observations from the population generated by $P(\cdot,\theta_0)$, where θ_0 is known; further, let $x_r, x_{r+1},\dots,x_{n-1},x_n$ be observations of the distribution $p(\cdot,\theta_1)$, where $\theta_1 \in \Theta$, but unknown. We will assume that $p(\cdot,\theta)$ is of the exponential type family, then on the basis of the previous results

$$\sup_{\theta\in\Theta} \sum_{j=r}^{n} \ln\frac{p(x_j,\theta)}{p(x_j,\theta_0)} = \frac{n+1-r}{r} I(p(\cdot,\hat{\theta}_r) : p(\cdot,\theta_0))$$

for each $r = 1,2,\dots,n$. Hence, the maximum likelihood detector has the form

$$g_n = \max_{1\le r\le n} \left\{ \frac{n-r+1}{2} I(p(\cdot,\hat{\theta}_r) : p(\cdot,\theta_0))\right\},$$

where $\hat{\theta}_r = \hat{\theta}_r(x_r, x_{r+1},\dots,x_n)$ is a maximum likelihood estimate of θ after a possible change at the time r.

At this moment we can, of course, proceed further by considering stochastically dependent variables in general. Let $x_1, x_2,\dots,x_n$ create a random sequence whose evolution can be described by conditional probability densities

$$p(x_{j+1} \mid x_j, x_{j-1},\dots,x_1, \theta),$$

$\theta \in \Theta$. Then the detection of a change in the parameter θ at the time $r \in \{1, 2, \ldots, n\}$ is based of conditional cumulative sums

$$S_r^n(\theta, \theta_0) = \sum_{j=r}^{n} \ln \frac{p(x_j \mid x_{j-1}, \ldots, x_1, \theta)}{p(x_j \mid x_{j-1}, \ldots, x_1, \theta_0)}$$

and the corresponding generalized MLE has the form

$$\hat{r}_n = \operatorname*{argmax}_{1 \le r \le n} \left\{ \max_{\theta \in \Theta} S_r^n(\theta, \theta_0) \right\},$$

(for more details, cf. Basseville and Benveniste [2]). If the collection of conditional density functions satisfies the relation between maximum likelihood estimates and I-divergence then the estimate $\hat{r}_n$ can be expressed using I-divergences. In the case of Gaussian stationary sequences the detection of a change can be based on I-divergence asymptotic rate as used in Michálek [8]. But, this case must be investigated more carefully because this is one of the possibilities how to come closer to dependent variables cases. A further possibility how to utilize the relation between MLE and I-divergence is a construction of a change detector based on I-divergences although the corresponding likelihood ratio does not satisfy the mentioned relation. Hence we can consider the detector

$$\hat{r}_n = \operatorname*{argmax}_{1 \le r \le n} \left\{ \frac{n - r + 1}{2} I(p(\cdot, \hat{\theta}_r) : p(\cdot, \theta_0)) \right\}$$

without any reference to the likelihood ratio.

Next we will use this approach showing a simple test concerning a change at a time instant r_2 against the alternative hypothesis that the change occurred at r_1. For simplicity let $r_1 < r_2$. We will investigate the simplest case when observations before the change are generated according to $N(0, 1)$ and after the change according to $N(\mu, 1)$, when μ is not known beforehand. Let us denote by

$$f_j(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{r_j - 1} \frac{1}{\sqrt{2\pi}} e^{-\frac{x_i^2}{2}} \prod_{i=r_j}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i - \mu)^2}{2}}, \quad j = 1, 2,$$

the joint density functions. Then $I(f_1 : f_2) = \frac{\mu^2}{2}(r_2 - r_1)$, as we have put $r_2 > r_1$, otherwise

$$I(f_1 : f_2) = \frac{\mu^2}{2} |r_2 - r_1|.$$

The proposed test is based on the estimates $\hat{I}(* : f_1)$ and $\hat{I}(* : f_2)$, respectively, where r_2, r_1, respectively is substituted by the maximum likelihood estimate of a change $\hat{r}$. The hypothesis H_2 (i.e. the change at r_2) is rejected when

$$\hat{I}(* : f_2) - \hat{I}(* : f_1) \ge C,$$

i.e.

$$\frac{\mu^2}{2}|r_2 - \hat{r}| - \frac{\mu^2}{2}|r_1 - \hat{r}| \geq C,$$

or, equivalently,

$$|r_2 - \hat{r}| - |r_1 - \hat{r}| \geq \frac{2C}{\mu^2}$$

where the constant C is chosen in such a way that the first kind error would be smaller or equal to a prescribed level α that is

$$P\{\hat{I}(* : f_2) - \hat{I}(* : f_1) \geq C\} \leq \alpha$$

under H_2.

In general, we have three possibilities, namely

$$\text{a) } \hat{r} \geq r_2 > r_1, \qquad \text{b) } r_2 > \hat{r} - r_1, \qquad \text{c) } r_2 > r_1 \geq \hat{r}.$$

In the case a) then $|r_2 - \hat{r}| = \hat{r} - r_2$ and $|r_1 - \hat{r}| = \hat{r} - r_1$ and the hypothesis H_2 cannot be rejected. In the case b) the hypothesis H_2 is not rejected if $\hat{r} \geq \frac{r_1+r_2}{2}$. Otherwise, all the remaining possibilities depend on the value of C. The case c) means that the left hand side of the inequality equals $r_2 - r_1$ and the test depends on $r_2 - r_1$, μ and C whether the hypothesis H_2 is rejected or not. For a better illustration the chart in Figure 1 presents all three possibilities.

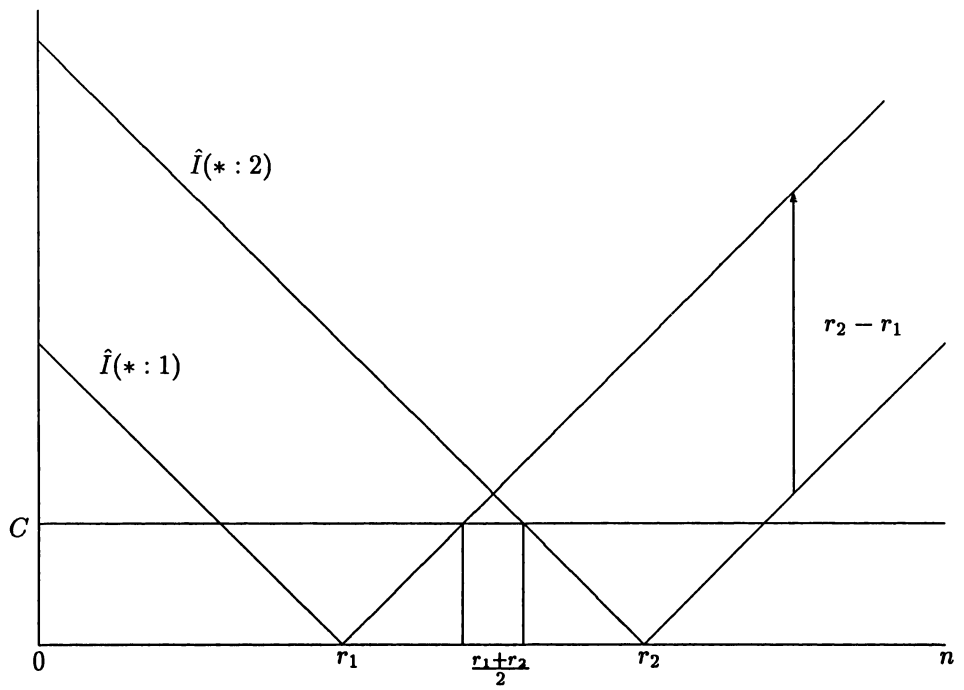

Fig. 1. Testing changes in mean using I-divergences.

Now, let us investigate the behavior of the first kind error, i.e.

$$\begin{aligned} & \mathsf{P}\{\hat{I}(*:f_2)-\hat{I}(*:f_1)\geq C/f_2\}= \\ = \ & \mathsf{P}\left\{|\hat{r}-r_2|-|\hat{r}-r_1|\geq \frac{2C}{\mu^2}\Big/ f_2\right\}= \\ = \ & \mathsf{P}\left\{r_1-r_2\geq\frac{2C}{\mu^2}\right\}+\mathsf{P}\left\{\frac{r_1+r_2}{2}-\hat{r}\geq\frac{C}{2\mu^2}\right\}+ \\ & +\mathsf{P}\left\{r_2-r_1\geq\frac{2C}{\mu^2}\right\}. \end{aligned}$$

As $r_2-r_1>0$, it holds $\mathsf{P}\left\{r_1-r_2\geq\frac{2C}{\mu^2}\right\}=0$. But, in the case $r_2-r_1\geq\frac{2C}{\mu^2}$, the third term would be 1 which is impossible because we demand the first kind error to be less or equal to α. Hence, it remains the case $r_2-r_1<\frac{2C}{\mu^2}$. Then the first kind error equals

$$\mathsf{P}\left\{\frac{r_1+r_2}{2}-\hat{r}\geq\frac{C}{\mu^2}\Big/ f_2\right\}=\mathsf{P}\left\{\hat{r}\leq\frac{r_1+r_2}{2}-\frac{C}{\mu^2}\Big/ f_2\right\}.$$

This result immediately implies that the following inequalities must hold simultaneously

$$\frac{r_2-r_1}{2}<\frac{C}{\mu^2},\qquad \frac{r_2+r_1}{2}>\frac{C}{\mu^2}.$$

If $\mathsf{P}\{\hat{r}\leq r_1/f_2\}\leq\alpha$ then the first kind error will be less than α.

At this moment can be answered the question why we have chosen the statistic $\hat{I}(*:f_2)-\hat{I}(*:f_1)$. This is closely connected with the test based on the likelihood ratio (cf. the well known Neyman–Pearson lemma). The logarithm of likelihood ratio has the form

$$\mu\sum_{i=r_1}^{r_2-1}x_i-\frac{\mu^2}{2}(r_2-r_1).$$

This implies that the Neyman–Pearson lemma suggests the test based on the statistic

$$\mu(r_2-r_1)\,(\overline{x}_{r_1}^{r_2}-\mu)$$

with the condition

$$\mathsf{P}\left\{\mu(r_2-r_1)\,(\overline{x}_{r_1}^{r_2}-\mu)\geq C/f_2\right\}\leq\alpha,$$

i.e.

$$\mathsf{P}\left\{\overline{x}_{r_1}^{r_2}\geq\frac{C}{\mu(r_2-r_1)}+\mu/f_2\right\}\leq\alpha,$$

which can be easily satisfied because $\overline{x}_{r_1}^{r_2}$ has the distribution $N\left(0,\frac{1}{\sqrt{r_2-r_1}}\right)$ under the null hypothesis with $\mu>0$. In case $\mu<0$ the situation is quite analogous. When the density functions satisfy the relation between the likelihood ratio maximum and

the I-divergence using the MLE's then the statistic $\hat{I}(* : 2) - \hat{I}(* : 1)$ is nothing else but the likelihood ratio because

$$\sup_{\theta\in\Theta} \ln \frac{f(\boldsymbol{x},\theta)}{f(\boldsymbol{x},\theta_2)} = \sup_{\theta\in\Theta} \ln \frac{f(\boldsymbol{x},\theta)}{f(\boldsymbol{x},\theta_1)} + \ln \frac{f(\boldsymbol{x},\theta_1)}{f(\boldsymbol{x},\theta_2)}$$

so that

$$\hat{I}(* : 2) - \hat{I}(* : 1) = \ln \frac{f(\boldsymbol{x},\theta_1)}{f(\boldsymbol{x},\theta_2)}.$$

This proves that the Neyman–Pearson test can be expressed as the difference of the corresponding I-divergences. But if the density functions do not satisfy the above mentioned relation then both the tests can be quite different. This is the case with the detection of a change time. Namely,

$$\begin{aligned}
&\sup_{r\in\{1,2,\ldots,n\}} \ln \frac{f(\boldsymbol{x},r)}{f(\boldsymbol{x},r_2)} \\
&= \sup_{r\in\{1,2,\ldots,n\}} \left\{ \mu \sum_{r}^{n} x_i - \frac{1}{2}\mu^2(n-r+1) - \mu \sum_{r_2}^{n} x_i + \frac{1}{2}\mu^2(n-r_2+1) \right\} \\
&= \sup_{r\in\{1,2,\ldots,n\}} \left\{ \mu^* \sum_{\min(r,r_2)}^{\max(r,r_2)-1} x_i + \frac{1}{2}\mu^2(r-r_2) \right\} \\
&= \mu^* \sum_{\min(\hat{r},r_2)}^{\max(\hat{r},r_2)-1} x_i + \frac{1}{2}\mu^2(\hat{r}-r_2),
\end{aligned}$$

where $\mu^* = \mu \operatorname{sign}(r_2 - \hat{r})$.

At the first sight we see that the final result contains

$$\hat{I}(* : r_2) = \frac{1}{2}\mu^2(\hat{r} - r_2),$$

but, in addition, also the statistic of the type

$$\mu^* \sum_{r_2}^{\hat{r}-1} x_i$$

is present.

The reason why it is useful to use the statistic $\hat{I}(* : 2) - \hat{I}(* : 1)$ is evident in testing composed hypotheses. Let us consider the simple hypothesis that a change occurred at the time r_2 but the alternative hypothesis is composed a change occurred after the time r_2. Then

$$\hat{I}(* : 2) = \frac{\mu^2}{2}|r_2 - \hat{r}|$$

but $\hat{I}(* : 1)$ is substituted by $\inf_{r>r_2} \frac{\mu^2}{2}|r - \hat{r}|$.

It is evident that $\hat{I}(* : 1) = 0$ if $\hat{r} \geq r_2$ and $\hat{I}(* : 1) = \frac{\mu^2}{2}|r_2 - \hat{r}|$ otherwise, i.e. if $\hat{r} < r_2$.

Then the test statistic equals

$$\hat{I}(*:2) - \inf_{r>r^2} \hat{I}(*:1) = \frac{\mu^2}{2}(\hat{r} - r_2) \quad \text{for} \quad \hat{r} \geq r_2$$
$$= 0 \quad \text{for} \quad \hat{r} < r_2.$$

The simple hypotheses is rejected if $\frac{\mu^2}{2}(\hat{r} - r_2) \geq C$, which gives

$$\hat{r} \geq r_2 + \frac{2C}{\mu^2}$$

where the constant C is determined so that

$$\mathsf{P}\left\{\frac{\mu^2}{2}|r_2 - \hat{r}| \geq C/H_2\right\} \leq \alpha.$$

Now, we can use this approach in a general scheme. Let $x_1, x_2, \ldots, x_n$ be an n-tuple of observations of mutually independent random variables which are generated by a density function $f(\cdot)$ before a change and by a density $g(\cdot)$ after a change at an instant $r \in \{1, 2, \ldots, n\}$. Let $r_1 < r_2$ and let us calculate I-divergence of the corresponding density functions. It is easy to prove that

$$I_n(r_1 : r_2) = (r_2 - r_1)\, I_1(f : g).$$

Let $\hat{r}$ be the MLE of the change time, hence

$$\hat{r} = \operatorname*{arg\,max}_{\{1,2,\ldots,n\}} \sum_{r}^{n} \ln \frac{g(x_i)}{f(x_i)},$$

i.e. for each $r \neq \hat{r}$

$$\sum_{\hat{r}}^{n} \ln \frac{g(x_i)}{f(x_i)} > \sum_{r}^{n} \ln \frac{g(x_i)}{f(x_i)}.$$

Then

$$\max_{r \in \{1,2,\ldots,n\}} \ln \frac{f_r(x_1, \ldots, x_n)}{f_{n+1}(x_1, \ldots, x_n)} = (n + 1 - \hat{r})\, M_{\hat{r}}^n(g : f),$$

where

$$M_{\hat{r}}^n(g : f) = \frac{1}{n + 1 - \hat{r}} \sum_{\hat{r}}^{n} \frac{g(x_i)}{f(x_i)}.$$

3. LIKELIHOOD RATIO TEST AND I-DIVERGENCE

The classical likelihood ratio test is based on the statistic

$$T(x_1, x_2, \ldots, x_n) = \frac{\sup_{\theta \in \Theta_0} \prod_{j=1}^{n} p(x_j, \theta)}{\sup_{\theta \in \Theta} \prod_{j=1}^{n} p(x_j, \theta)},$$

where Θ is a parametric space and Θ_0 is a subset of Θ corresponding to the hypothesis H_0. Thanks to properties of the function $\ln(\cdot)$ one can write

$$\begin{aligned}\ln(T(x_1, x_2, \ldots, x_n)) &= \ln\left(\sup_{\theta\in\Theta_0}\prod_{j=1}^{n} p(x_j,\theta)\right) - \ln\left(\sup_{\theta\in\Theta}\prod_{j=1}^{n} p(x_j,\theta)\right)\\ &= \sup_{\theta\in\Theta_0}\ln\prod_{j=1}^{n} p(x_j,\theta) - \sup_{\theta\in\Theta}\ln\prod_{j=1}^{n} p(x_j,\theta)\\ &= \sup_{\theta\in\Theta_0}\sum_{j=1}^{n}\ln\frac{p(x_j,\theta)}{p(x_j,\theta_0)} - \sup_{\theta\in\Theta}\sum_{j=1}^{n}\ln\frac{p(x_j,\theta)}{p(x_j,\theta_0)},\end{aligned}$$

where θ_0 is an arbitrary element of Θ. Now, let us assume that the density functions $\{p(\cdot,\theta),\ \theta\in\Theta\}$ satisfy the basic relation between the likelihood ratio maximum and I-divergence (e. g. $p(\cdot,\theta)$ belongs to the exponential family of densities). Then we can express the test statistic $T(x_1, x_2, \ldots, x_n)$ in the form

$$T(x_1, x_2, \ldots, x_n) = \exp\left\{-\frac{n}{2}(I(p(\cdot,\hat\theta) : p(\cdot,\theta_0)) - I(p(\cdot,\hat\theta_0) : p(\cdot,\theta_0))\right\}$$

where $\hat\theta$ is a global MLE over all the space Θ and $\hat\theta_0$ is a local MLE over the hypothesis domain Θ_0 only. At this moment it is evident that there are many important questions about the existence of these estimates.

The following examples serve as an illustration of the situation.

Example 1. Let $x_1, x_2, \ldots, x_n$ be an n-tuple of mutually independent observations coming from $N(\mu, 1)$, where the parameter μ is unknown and set up the hypothesis $\mu\in\langle a, b\rangle$ against the alternative hypothesis $\mu\notin\langle a, b\rangle$. The test statistic $T(x_1, x_2, \ldots, x_n)$ has in this case the form

$$\begin{aligned}\ln T(x_1, x_2, \ldots, x_n) &= \sup_{\mu\in\langle a,b\rangle}\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right\} - \sup_{\mu\in R_1}\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}\\ &= 0 \qquad \text{if} \quad \overline{x} = \frac{1}{n}\sum_{1}^{n} x_i \in \langle a, b\rangle\\ &= -\frac{1}{2}\sum_{i=1}^{n}(x_i-a)^2 + \frac{1}{2}\sum_{i=1}^{n}(x_i-\overline{x})^2 \quad \text{if} \quad \overline{x} < a\\ &= -\frac{1}{2}\sum_{i=1}^{n}(x_i-b)^2 + \frac{1}{2}\sum_{i=1}^{n}(x_i-\overline{x})^2 \quad \text{if} \quad \overline{x} > b.\end{aligned}$$

Using these relation we can state that

$$\begin{aligned}T(x_1, x_2, \ldots, x_n) &= 1 && \text{for} \quad \overline{x}\in\langle a, b\rangle\\ T(x_1, x_2, \ldots, x_n) &= \exp\left\{-\tfrac{n}{2}(\overline{x}-a)^2\right\} && \text{for} \quad \overline{x} < a\\ T(x_1, x_2, \ldots, x_n) &= \exp\left\{-\tfrac{n}{2}(\overline{x}-b)^2\right\} && \text{for} \quad \overline{x} > b.\end{aligned}$$

Let us denote by

$$\begin{aligned} \overline{x}_0 &= \overline{x} \quad \text{for} \quad \overline{x} \in \langle a, b\rangle \\ \overline{x}_0 &= a \quad \text{for} \quad \overline{x} < a \\ \overline{x}_0 &= b \quad \text{for} \quad \overline{x} > b, \end{aligned}$$

then

$$T(x_1, x_2, \ldots, x_n) = \exp\left\{-\frac{n}{2}(\overline{x} - \overline{x}_0)^2\right\},$$

which in the sense of I-divergence means

$$T(x_1, x_2, \ldots, x_n) = \exp\left\{-\frac{n}{2} I(p(\cdot, \overline{x}_0) : p(\cdot, \overline{x}))\right\}.$$

There is no problem to show that

$$\inf_{\mu_0 \in \langle a,b\rangle} \left\{ \sup_{\mu \in R_1} \sum_{j=1}^{n} \frac{p(x_j, \mu)}{p(x_j, \mu_0)} \right\} = \frac{n}{2} I(p(\cdot, \overline{x}_0) : p(\cdot, \overline{x})).$$

From here we see again a close relation between MLE and I-divergence. It is quite natural to ask how this relation can be generalized. We can, of course, write immediately

$$T(x_1, x_2, \ldots, x_n) = \exp\left\{ - \inf_{\mu_0 \in \langle a,b\rangle} \left\{ \sup_{\mu \in R_1} \sum_{j=1}^{n} \ln \frac{p(x_j, \mu)}{p(x_j, \mu_0)} \right\} \right\}.$$

This follows from the relation

$$\sup_{\mu \in R_1} \sum_{j=1}^{n} \ln \frac{p(x_j, \mu)}{p(x_j, \mu_0)} = \frac{n}{2} I(p(\cdot, \overline{x}) : p(\cdot, \mu_0)) = \frac{n}{2}(\overline{x} - \mu_0)^2.$$

Now, $\inf_{\mu_0 \in \langle a,b\rangle}\{(\overline{x} - \mu_0)^2\}$ either equals 0, if $\overline{x} \in \langle a, b\rangle$ or equals a if $\overline{x} < a$ and equals b if $\overline{x} > b$. These facts establish the formula for $T(x_1, x_2, \ldots, x_n)$.

Example 2. Let us have observations $x_1, x_2, \ldots, x_n$ from the population $N(0, \sigma^2)$ and test the hypothesis $H_0 : 0 < \sigma^2 < K$ against the alternative hypothesis $H_1 : \sigma^2 > K$.

In this case we can easily calculate that the likelihood ratio has a form

$$\ln \frac{p(x, \sigma^2)}{p(x, \sigma_0^2)} = \frac{1}{2} \ln \frac{\sigma_0^2}{\sigma^2} + \frac{1}{2}\left(\frac{x}{\sigma_0^2} - \frac{x}{\sigma_1^2}\right).$$

This gives

$$\sum_{j=1}^{n} \ln \frac{p(x_j, \sigma^2)}{p(x_j, \sigma_0^2)} = \frac{n}{2} \ln \frac{\sigma_0^2}{\sigma^2} + \frac{1}{2} \frac{\sigma^2 - \sigma_0^2}{\sigma^2 \sigma_0^2} \sum_{j=1}^{n} x_j^2.$$

Let us use the relation

$$\sup_{\sigma^2>0} \sum_{j=1}^{n} \ln \frac{p(x_j,\sigma^2)}{p(x_j,\sigma_0^2)} = \frac{n}{2}\left(\frac{s^2}{\sigma_0^2} - \ln\frac{s^2}{\sigma_0^2} - 1\right),$$

where $s^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2$.

Next, let us calculate

$$\inf_{0<\sigma_0^2<K}\left\{\frac{n}{2}\left(\frac{s^2}{\sigma_0^2} - \ln\frac{s^2}{\sigma_0^2} - 1\right)\right\}.$$

Using properties of the function $x - \ln x - 1$ we can find out that this infimum equals s^2 in the case $0 < s^2 \leq K$ and equals K if $s^2 > K$, because

$$\lim_{x \searrow 0+}\left(\frac{s^2}{x} - \ln\frac{s^2}{x} - 1\right) = +\infty.$$

In this way we have proved that

$$\inf_{0<\sigma_0^2<K}\left\{\sup_{\sigma^2>0}\sum_{j=1}^{n}\ln\frac{p(x_j,\sigma^2)}{p(x_j,\sigma_0^2)}\right\} \begin{array}{ll} = 0 & \text{for } 0 < s^2 \leq K \\ = \frac{n}{2}\left(\frac{s^2}{K} - \ln\frac{s^2}{K} - 1\right) & \text{for } s^2 > K. \end{array}$$

This result is also very closely connected with the I-divergence, as seen at the first sight. Similarly, we can proceed in the case of the hypothesis $H_0 : K_0 \leq \sigma^2 \leq K_1$ against the alternative hypothesis H_1: $\sigma^2 \notin \langle K_0, K_1\rangle$. Then the resulting test is based on the following statistic

$$\begin{aligned} s_0^2(x_1,x_2,\ldots,x_n) &= s^2 \quad \text{for} \quad s^2 \in \langle K_0, K_1\rangle \\ s_0^2(x_1,x_2,\ldots,x_n) &= K_0 \quad \text{for} \quad s^2 < K_0 \\ s_0^2(x_1,x_2,\ldots,x_n) &= K_1 \quad \text{for} \quad s^2 > K_1. \end{aligned}$$

Here, we can see again an interesting relation, namely

$$\inf_{\sigma^2\in\langle K_0,K_1\rangle}\left\{\sup_{\sigma^2\in R_1}\sum_{j=1}^{n}\ln\frac{p(x_j,\sigma^2)}{p(x_j,\sigma_0^2)}\right\} = \frac{n}{2}\, I(p(\cdot,s^2) : p(\cdot,s_0^2)).$$

Then the test statistic $T(x_1,x_2,\ldots,x_n)$ equals

$$T(x_1,x_2,\ldots,x_n) = \exp\left\{-\frac{n}{2}\left(\frac{s^2}{s_0^2} - \ln\frac{s^2}{s_0^2} - 1\right)\right\}.$$

Example 3. Let us consider an n-tuple $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ of observations from a two-dimensional population $N(\mathbf{0}, I\!R)$, i.e. $\boldsymbol{x}_j = (x_{j1}, x_{j2})$, with the density function

$$p(\boldsymbol{x}, I\!R) = \frac{1}{2\pi\,|I\!R|^{1/2}}\exp\left\{-\frac{1}{2}\,\boldsymbol{x}\, I\!R^{-1}\,\boldsymbol{x}\right\}.$$

Here the parameter space Θ contains all the positive definite symmetric matrices $\mathbb{R}$, i.e.

$$\Theta = \left\{ \mathbb{R} : \mathbb{R} = \begin{pmatrix} R_1 & R_3 \\ R_3 & R_2 \end{pmatrix},\ \mathbb{R} > 0 \right\}.$$

The hypothesis is given by $\Theta_0 : \left\{ \mathbb{R}_0 : \mathbb{R}_0 = \begin{pmatrix} R_1 & 0 \\ 0 & R_2 \end{pmatrix},\ \mathbb{R}_0 > 0 \right\}$.

This means that the subset Θ_0 characterizes the stochastic independence of the coordinates of observations. If we calculate the MLE over all Θ we obtain the following estimates

$$\hat{R}_1 = \frac{1}{n}\sum_{j=1}^{n} x_{j1}^2, \quad \hat{R}_2 = \frac{1}{n}\sum_{j=1}^{n} x_{j2}^2, \quad \hat{R}_3 = \frac{1}{n}\sum_{j=1}^{n} x_{j1}\, x_{j2}.$$

By a similar calculation we can find out that the MLE over the hypothesis Θ_0 equals

$$\hat{R}_1 = \frac{1}{n}\sum_{j=1}^{n} x_{j1}^2, \quad \hat{R}_2 = \frac{1}{n}\sum_{j=1}^{n} x_{j2}^2.$$

In both the cases the MLE's belong to the parameter space Θ, Θ_0, respectively with probability 1.

The corresponding I-divergence between two-dimensional Gaussian population with vanishing mean and the covariance matrices $\mathbb{R}_0$, $\mathbb{R}_1$ equals

$$\begin{aligned} & I(p(\cdot, \mathbb{R}_1) : p(\cdot, \mathbb{R}_0)) = \\ = \ & \frac{1}{2}\left(\operatorname{tr} \mathbb{R}_1 \mathbb{R}_0^{-1} - \ln\det\{\mathbb{R}_1 \mathbb{R}_0^{-1}\} - 2\right). \end{aligned}$$

If we choose for simplicity $\mathbb{R}_0 = \mathbb{E} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ then

$$I(\mathbb{R} : \mathbb{E}) = \frac{1}{2}\left(\operatorname{tr} \mathbb{R} - \ln\det \mathbb{R} - 2\right).$$

Using this we can write

$$T(x_1, x_2, \ldots, x_n) = \exp\left\{-\frac{n}{2}(\operatorname{tr} \hat{\mathbb{R}}_1 - \ln\det \hat{\mathbb{R}}_1 - 2 + \operatorname{tr} \hat{\mathbb{R}}_0 + \ln\det \hat{\mathbb{R}}_0 + 2\right),$$

where

$$\hat{\mathbb{R}}_1 = \begin{pmatrix} \hat{R}_1, & \hat{R}_3 \\ \hat{R}_3, & \hat{R}_2 \end{pmatrix}, \quad \hat{\mathbb{R}}_0 = \begin{pmatrix} \hat{R}_1, & 0 \\ 0, & \hat{R}_2 \end{pmatrix}.$$

Then

$$\begin{aligned} & T(x_1, x_2, \ldots, x_n) = \exp\left\{-\frac{n}{2}\ln\det(\hat{\mathbb{R}}_0 \hat{\mathbb{R}}_1^{-1})\right\} \\ = \ & \exp\left\{\frac{n}{2}\ln\frac{\det\hat{\mathbb{R}}_1}{\det\hat{\mathbb{R}}_0}\right\} = \left(\frac{\det\hat{\mathbb{R}}_1}{\det\hat{\mathbb{R}}_0}\right)^{n/2} = [1 - r^2(x_1, x_2, \ldots, x_n)]^{n/2}, \end{aligned}$$

where

$$r(x_1, x_2, \ldots, x_n) = \frac{\hat{R}_3}{\sqrt{\hat{R}_1 \hat{R}_2}}.$$

In this way we have got that the test statistic $T(x_1, x_2, \ldots, x_n)$ equals

$$T(x_1, x_2, \ldots, x_n) = \left(\sqrt{1-r^2}\right)^n,$$

where r is the sample correlation coefficient between $x_{j,1}$ and $x_{j,2}$, $j = 1, 2, \ldots, n$.

4. THE CASE OF DEPENDENT VARIABLES - AUTOREGRESSION

First, we will investigate in detail the most elementary case, i.e. the autoregressive sequence of the first order. Let us have observations

$$x_{j+1} + a\, x_j = e_{j+1}, \quad j = 0, 1, \ldots, n-1,$$

$$x_{j+1} + b\, x_j = e_{j+1}, \quad j = 0, 1, \ldots, n-1, \quad \text{respectively,}$$

where for simplicity $e_{j+1} \sim N(0, \sigma_0^2)$, $N(0, \sigma_1^2)$ respectively. We will look at the behavior of

$$\max_{b \in R_1} \frac{f(x_0, x_1, \ldots, x_n, b, \sigma_1^2)}{f(x_0, x_1, \ldots, x_n, a, \sigma_0^2)}.$$

It is easy to prove that

$$\ln \frac{f(x_0, x_1, \ldots, x_n, b, \sigma_1^2)}{f(x_0, x_1, \ldots, x_n, a, \sigma_0^2)} = -\frac{n}{2} \ln \frac{\sigma_1^2}{\sigma_0^2} - \frac{1}{2\sigma_1^2} \sum_0^{n-1} (x_{j+1} + b\, x_j)^2 + \frac{1}{2\sigma_0^2} \sum_0^{n-1} (x_{j+1} + a\, x_j)^2,$$

if for simplicity $x_0 \sim N(0, 1)$. Then the MLE of the parameter b equals

$$\hat{b} = -\frac{\sum_0^{n-1} x_{j+1}\, x_j}{\sum_0^{n-1} x_j^2}.$$

When we substitute $\hat{b}$ for b in the likelihood ratio, we obtain

$$\max_{b \in R_1} \ln \frac{f(x_0, x_1, \ldots, x_n, b, \sigma_1^2)}{f(x_0, x_1, \ldots, x_n, a, \sigma_0^2)} = -\frac{n}{2} \ln \frac{\sigma_1^2}{\sigma_0^2} + \frac{n}{2\sigma_0^2} \boldsymbol{a}^T \hat{\mathbb{R}}\, \boldsymbol{a} - \frac{\hat{n}}{2\sigma_1^2} \hat{\boldsymbol{b}}^T \hat{\mathbb{R}}\, \hat{\boldsymbol{b}},$$

where $\hat{R}_{00} = \frac{1}{n} \sum_0^{n-1} x_j^2$, $\hat{R}_{01} = \hat{R}_{10} = \frac{1}{n} \sum_0^{n-1} x_j\, x_{j+1}$, $\hat{R}_{11} = \frac{1}{n} \sum_0^{n-1} x_{j+1}^2$ and

$$\hat{\mathbb{R}} = \begin{pmatrix} \hat{R}_{00}, & \hat{R}_{01} \\ \hat{R}_{10}, & \hat{R}_{11} \end{pmatrix}, \qquad \boldsymbol{a}^T = (1, a), \qquad \hat{\boldsymbol{b}}^T = (1, \hat{b}).$$

The form of the likelihood ratio maximum which shows a close connection with I-divergence can be expressed as follows

$$\begin{aligned} \max_{b \in R_1} \ln \frac{f(x_0, x_1, \ldots, x_n, b, \sigma_1^2)}{f(x_0, x_1, \ldots, x_n, a, \sigma_0^2)} &= \frac{n}{2} \left(\frac{\boldsymbol{a}^T \hat{\mathbb{R}}\, \boldsymbol{a}}{\sigma_0^2} - \ln \frac{\sigma_1^2}{\sigma_0^2} - 1 \right) \\ &\quad - \frac{n}{2} \left(\frac{\hat{\boldsymbol{b}}^T \hat{\mathbb{R}}\, \hat{\boldsymbol{b}}}{\sigma_1^2} - \ln \frac{\sigma_1^2}{\sigma_1^2} - 1 \right). \end{aligned}$$

This result strongly reminds the asymptotic I-divergence rate between two Gaussian autoregressive sequences, see Michálek [7].

A quite similar situation occurs in the case of a general autoregressive sequence of the pth order with parameters $(a_1, a_2, \ldots, a_p, \sigma_0^2)$ and $(b_1, b_2, \ldots, b_p, \sigma_1^2)$, i.e.

$$x_{i+1} + \sum_{k=1}^{p} a_k x_{i-k} = \sigma_0 e_{i+1},$$

$$x_{i+1} + \sum_{k=1}^{p} b_k x_{i-k} = \sigma_1 e_{i+1}, \quad \text{respectively.}$$

Then the logarithm of the likelihood ratio is given by

$$\ln \frac{f(x_0, x_1, \ldots, x_n, \boldsymbol{b}, \sigma_1^2)}{f(x_0, x_1, \ldots, x_n, \boldsymbol{a}, \sigma_0^2)} = \frac{n}{2} \ln \frac{\sigma_1^2}{\sigma_0^2} + \frac{1}{2} \sum_{i=0}^{n-1} \frac{(x_{i+1} + a_1 x_2 + \cdots + a_p x_{i-p})^2}{\sigma_0^2}$$
$$- \frac{1}{2} \sum_{i=0}^{n-1} \frac{(x_{i+1} + b_1 x_1 + \cdots + b_p x_{i-p})^2}{\sigma_1^2}.$$

Now, we must find the MLE of $\boldsymbol{b}^T = (b_1, b_2, \ldots, b_p)$. We obtain a system of linear equations $k = 1, 2, \ldots, p$

$$0 = \frac{\partial}{\partial b_k} = \sum_{1}^{n} x_j x_{j-k} + b_1 \sum_{1}^{n} x_{j-1} x_{j-k} + \cdots + b_p \sum_{j=1}^{n} x_{j-p} x_{j-k}.$$

The matrix of the system is positive definite with probability 1 and the matrix of second partial derivatives is negative definite, which means there is the only solution of this system and this solution gives the maximum of likelihood ratio. If we substitute the maximum likelihood estimate $\hat{\boldsymbol{b}}$ into the logarithm of the ratio we get

$$\max_{\boldsymbol{b}} \ln \frac{f(\boldsymbol{x}, \boldsymbol{b}, \sigma_1^2)}{f(\boldsymbol{x}, \boldsymbol{a}, \sigma_0^2)} = \frac{n}{2} \ln \frac{\sigma_1^2}{\sigma_0^2} + \frac{n}{2} \frac{\boldsymbol{a}^T \hat{I\!R} \boldsymbol{a}}{\sigma_0^2} - \frac{n}{2} \frac{\hat{\boldsymbol{b}}^T \hat{I\!R} \hat{\boldsymbol{b}}}{\sigma_1^2},$$

where

$$\hat{I\!R} = \{\hat{R}_{ij}\}_{i,j=1}^{p}, \qquad \hat{R}_{ij} = \frac{1}{n} \sum_{k=1}^{n} x_{k-i} x_{k-j}.$$

More interesting situation arises if we consider the whole parameter, i.e. including dispersion σ_1^2. Then the maximum likelihood estimate of σ_1^2 equals

$$\hat{\sigma}_1^2 = \frac{1}{n} \sum_{j=1}^{n} (x_j + \hat{b}_1 x_{j-1} + \cdots + \hat{b}_p x_{j-p})^2,$$

i.e.

$$\hat{\sigma}_1^2 = \hat{\boldsymbol{b}}^T \hat{I\!R} \hat{\boldsymbol{b}}$$

and hence, the expression of the likelihood ratio maximum over (b, σ_1^2) is given as

$$\max_{(b,\sigma_1^2)} \ln \frac{f(x, b, \sigma_1^2)}{f(x, a, \sigma_0^2)} = \frac{n}{2} \left(\frac{a^T \hat{R} a}{\sigma_0^2} - ln \frac{\hat{\sigma_1}^2}{\sigma_0^2} - 1 \right).$$

Now, the expression within parentheses would be identical with the asymptotic I-divergence rate between two Gaussian autoregression sequences if the Toeplitz matrix $\hat{T} = \{\hat{T}_{ij}\}_{i,j=1}^n$ with

$$\hat{T}_{ij} = \frac{1}{n} \sum_{k=1}^{n} x_k \, x_{k+|i-j|}$$

were used instead of the matrix $\hat{R}$.

We can state the following theorem dealing with the asymptotic behaviour of the likelihood ratio maximum in the case of Gaussian autoregressive sequences.

Theorem. Let $\{x_i\}_{i=1}^{\infty}$ be an stationary autoregressive Gaussian sequence of the pth order with parameters $(b_1, b_2, \ldots, b_p, \sigma_1^2)$. Let $(a_1, a_2, \ldots, a_p, \sigma_0^2)$ be parameters of another Gaussian autoregressive stationary sequence. Then the likelihood ratio maximum satisfies

$$\lim_{n\to\infty} \frac{1}{n} \max_{b_1,b_2,\ldots,b_p,\sigma_1^2} \frac{p(x, b, \sigma_1^2)}{p(x, a, \sigma_0^2)} =$$

$$= \frac{1}{2} \overline{I}((b^*, \sigma^*) : (a, \sigma_0)),$$

where (b^*, σ^*) are true parameters and

$$\overline{I}((b, \sigma_1^2) : (a, \sigma_0^2)) = \frac{1}{4\pi} \int_{-\pi}^{\pi} \left(\frac{\varphi_{(b,\sigma_1)}(\lambda)}{\varphi_{(a,\sigma_0)}(\lambda)} - \ln \frac{\varphi_{(b,\sigma_1)}, (\lambda)}{\varphi_{(a,\sigma_0)}(\lambda)} - 1 \right) \mathrm{d}\lambda,$$

and $\varphi_{(b,\sigma_1)}(\cdot)$, $\varphi_{(a,\sigma_0)}(\cdot)$ are the corresponding spectral density functions.

Proof. It follows almost immediately from the strong consistency of maximum likelihood estimates in a stationary case, e. g. see Anderson [1]. □

(Received November 7, 1996.)

REFERENCES

[1] T. W. Anderson: The Statistical Analysis of Time Series. Wiley, New York 1971.

[2] M. Basseville and A. Benveniste: Detection of Abrupt Changes in Signals and Dynamical Systems. Springer–Verlag, Berlin 1986.

[3] P. R. Krishnaiah and B. Q. Miao: Review about estimation of change points. In: Handbook of Statistics (P. R. Krishnaiah and C. R. Rao, eds.), Elsevier Sci. Publishers, Amsterdam 1988, Vol. 7, pp. 375–402.

[4] S. Kullback: Information Theory and Statistics (in Russian). Nauka, Moscow 1967. Translated from the English original.

[5] M. Kupperman: Further application of information theory to multivariate analysis and statistical inference. Ann. Math. Statist. *27* (1956), 1184.
[6] V. Kűchler and M. Sorensen: Exponential families of stochastic processes: A unifying semimartingale approach. Internat. Statist. Rev. (1989), 123–144.
[7] J. Michálek: Yule–Walker estimates and asymptotic I-divergence rate. Problems Control Inform. Theory *19* (1990), 5–6, 387–398.
[8] J. Michálek: A method of detecting changes in the behaviour of locally stationary sequences. Kybernetika *31* (1995), 1, 17–29.
[9] D. Morales, L. Pardo and I. Vajda: About classical and some new statistics for testing hypothesis in parametric models. J. Multivariate Anal. (to appear).
[10] E. Page: Continuous inspection schemes. Biometrika *41* (1954), 100–115.

RNDr. Jiří Michálek, CSc., Institute of Information Theory and Automation – Academy of Sciences of the Czech Republic, Pod vodárenskou věží 4, 182 08 Praha 8. Czech Republic.
e-mail: michalek@utia.cas.cz