INFORMATICA, 1992, Vol.3, No.1, 47–63

HYPOTHESES TESTING IN MIXTURES OF TIME SERIES I

Jiří MICHÁLEK

Institute of Information Theory and Automation
Czechoslovak Academy of Sciences
182 08 Prague 8, Pod vodárenskou věží 4, Czechoslovakia

Abstract. The paper defines the decomposition problem of a mixture of time series into homogeneous components. First part deals with a solution based on Bayesian approach in the case of independent observations, the other part is devoted to a solution of on-line decomposition for a time series consisting of weakly stationary components.

Key words: Bayesian test, change point problem, first and second kind error.

1. Introduction In practice one can meet the following statistical decision problem. Let us imagine we observe a time series where individual observations belong to different time subseries, which are stirred together in a manner unknown for the observer, i.e., we don't know which time subseries the observation actually belongs to. The task for a statistician is of course to make a decision about the last observation, which a time subseries this observation belongs to (on-line problem) or when we have all the observations at our disposal the task is to decompose the given mixture into the corresponding subseries (off-line problem). The problem described above can be illustrated by the following example. Watching the output from a survey radiolocator we meet precisely the on-line situation. We receive individual responses of airplanes detected by our radiolocator in irregular time intervals without knowing which airplane the obtained response belongs to. Thus, we are in the situation to solve the problem of assigning the last observation into one of former subseries presenting the airplanes already detected. The off-line problem we can explain using the simplest example belonging into the class of time series mixtures. Let us assume the observed time series is a mixture of two subseries only following one other. This is the problem of change detection in the behavior of a random sequence. At the beginning we observe a subseries described, e.g., by a suitable autoregressive model but the parameters of the mentioned autoregressive model have ubruptly changed at a time instant unknown for the observer. In this way the observed time series is the mixture of two components, the first is observed before the change, the other after the change. Our task is to detect the possible change. This problematics belongs to hypotheses testing and parameters estimation in nonstationary time series and is very intensively investigated during last 15 years both from the theoretical point of view and the practical one because of the direct exploration of these methods in practice, mainly in technical diagnosis. As survey papers in this problematics of change detection in time series one can recommend the following papers Kligienė, Telksnys (1983), Basseville, Benveniste (1986), Willsky (1976), Basseville (1988). The main goal of this article is to suggest some tests and methods looking for components forming an observed time series. The presented tests are mainly based on the Bayesian approach.

2. Problem formulation. Let us consider N time series $\{x_k(t)\}_{k=1}^N$, where the parameter t presents time. Let $t_1 < t_2 < t_3 < \cdots < t_n$ be a sequence of time instances, at which we observe a time series. Then the time series

$$\{x_{k_\ell}(t_j)\}_{j=1,\ell=1}^n,$$

where for every $k \in \{1, 2, \ldots, n\}$, $k_\ell \in \{1, 2, \ldots, N\}$ is named by a nixture of the time series $x_k(t)$, $k = 1, 2, \ldots, N$. The individual time series $\{x_k(t)\}$ will be called the component of the given mixture. The statistical problem explained above can be briefly characterized how to choose components from the observed mixture. The number of components need not be known to the observer in advance. Roughly speaking one must decided about every random variable $x_{k_\ell}(t_j)$ to which of components it belongs. The simplest case of a mixture we obtain when all the observation are mutually stochastically independent and all the variables forming one component have the same probability distribution function. A special attention must be paid to the case when the mixture has only two components. Then the random variable $x_{k_\ell}(t_j)$ has the distribution function P_0 or P_1. If we choose any subgroup of the mixture $\{x_{k_\ell}(t_j)\}_{j=1,\ell=1}^{n}$, let us say $\{x_{k_\ell}(t_{j_m})\}_{m=1}^{M}$, we can immediately put the question whether this part of observations forms a component or not. We have constructed in this manner a simple hypothesis H_0 that the chosen subgroup forms a component generated according to P_0 against a composed alternative hypothesis that this subgroup does not form a homogeneous component. Theoretically speaking we can decomposed this test into many simpler tests where we test that simple hypothesis against $2^M - 1$ simple alternative hypotheses. In this case the answer is given by the classical Neyman Pearson lemma suggesting the maximal likelihood test. But, in practice this approach is almost impossible owing to the large number of tests. Let us try to construct a Bayesian test in this situation. Let us assume that the distribution P_0, resp. P_1, is given by a density function f, resp., g. We shall test the hypothesis H_0 that all the observations $x_{k_\ell}(t_j) = \xi(t_j)$, $j = 1, 2, \dots, N$ have the same probability distribution P_0 against the alternative hypothesis there exists among $\{\xi(t_j)\}_{j=1}^{N}$ at least one observation with the density function g. The parametric space is composed of N-tuples from 0 and 1 if we shall consider

$$0 \leftrightarrow f$$

$$1 \leftrightarrow g$$

i.e., $\Omega = \{\theta : \theta = \{i_1, i_2, \dots, i_n\},\ i_j = 0 \text{ or } 1\}$. Then the hypothesis H_0 is presented by the one-element subset $\theta_0 = (0, 0, \dots, 0)$ and the alternative hypothesis H_1 is given by the complement $\Omega - \theta_0$. Let us define the loss function $\ell(\theta, H_0)$, $\ell(\theta, H_1)$

$$\ell(\theta_0, H_0) = 0 \qquad \ell(\theta_0, H_1) = 1$$

$$\ell(\theta_i, H_1) = 1 \qquad \ell(\theta_i, H_1) = 0$$

for every $\theta_i \neq \theta_0$.

Let be given a prior distribution function $\{p_0, p_1, \ldots, p_{2^n-1}\}$ on Ω where, of course

$$\sum_{j=0}^{2^N-1} p_j = 1.$$

A sought decision rule is given by the prescription $\Phi(\cdot, \mathbf{x})$ where $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and

$$\Phi(H_0 \mid \mathbf{x}) + \Phi(H_1 \mid \mathbf{x}) = 1$$

so that we reject the hypothesis H_0 with probability $1 - \Phi(H_0 \mid \mathbf{x})$ under the realization $\mathbf{x}$. The Bayesian decision rule must minimize the average loss function. For this purpose one must know the aposterior distribution of the parameter θ. By use of Bayesian formula the aposterior density function has the form

$$h(\theta_i \mid \mathbf{x}) = \frac{f_i(\mathbf{x})}{\sum_{j=0}^{2^N-1} f_j(\mathbf{x})},$$

where $f_i(\cdot)$ is the N-dimensional density function corresponding to the parameter value $\theta_i \in \Omega$. Then the conditional expected value of the loss function $\ell(\cdot, \cdot)$ is

$$E\{l(\theta, H_0) \mid \mathbf{x}\} = [1 - \Phi(H_0 \mid \mathbf{x})]h(\theta_0 \mid \mathbf{x}) + \Phi(H_0 \mid \mathbf{x})[1 - h(\theta_0 \mid \mathbf{x})]. \quad (*)$$

We look for $\Phi(\cdot \mid \mathbf{x})$ in order that $E\{\ell(\theta, H_0)\}$ may be minimal. After analyzing $(*)$ we find the following decision rule:

if $h(\theta_0 \mid \mathbf{x}) > \frac{1}{2}$ then $\Phi(H_0 \mid \mathbf{x}) = 1$

if $h(\theta_0 \mid \mathbf{x}) < \frac{1}{2}$ then $\Phi(H_0 \mid \mathbf{x}) = 0$

if $h(\theta_0 \mid \mathbf{x}) = \frac{1}{2}$ then

$$E\{\ell(\theta \mid H_0) \mid \mathbf{x}\} = \frac{1}{2} \quad \text{for every } \Phi(H_0 \mid \mathbf{x}).$$

Since

$$h(\theta_0 \mid \mathbf{x}) = \frac{p_0 f_0(\mathbf{x})}{\sum_{j=0}^{2^N-1} p_j f_j(\mathbf{x})},$$

then

$$h(\theta_0 \mid \mathbf{x}) > \frac{1}{2} \quad \text{iff} \quad p_0 f_0(\mathbf{x}) > \sum_{j=1}^{2^N-1} p_j f_j(\mathbf{x})$$

$$h(\theta_0 \mid \mathbf{x}) < \frac{1}{2} \quad \text{iff} \quad p_0 f_0(\mathbf{x}) < \sum_{j=1}^{2^N-1} p_j f_j(\mathbf{x})$$

$$h(\theta_0 \mid \mathbf{x}) = \frac{1}{2} \quad \text{iff} \quad p_0 f_0(\mathbf{x}) = \sum_{j=1}^{2^N-1} p_j f_j(\mathbf{x})$$

We obtained a nonrandomized Bayesian decision rule. As long as the prior distribution is uniform on Ω we don't reject H_0 if

$$f_0(\mathbf{x}) > \sum_{j=1}^{2^N-1} f_j(\mathbf{x}).$$

Knowing this formula one can expect that this decision rule will be very restrictive, i.e., the hypothesis H_0 will be rejected in most cases. This unpleasant property can be removed only by use of knowledge of a prior distribution on the parameter space Ω. The considered Bayesian test is simple against the enormous number of classical Neyman–Pearson tests, but this simplicity is paid by possible rejection of H_0.

Now, we should, of course realize that both the hypothesis H_0 and the alternative hypothesis are invariant with respect to the permutation group in observations. The hypothesis H_0 is invariant with respect to every permutation and for every $\theta_i \in H_1$ and every permutation there exists the only $\theta_j \in H_1$ such that the mentioned permutation transforms the density function $f_i(\mathbf{x})$ onto $f_j(\mathbf{x})$. The chosen loss function is invariant also with respect to permutations. It means the whole decision problem (Ω, H_0, H_1, ℓ) is invariant with respect to permutation group. If τ is a prior distribution on the parameter space Ω and the decision rule δ_0 is Bayesian with respect to τ then there exists an invariant distribution τ_0 on Ω that the decision δ_0 is also Bayesian with respect to τ_0. It means, looking for Bayesian decision rules we can confine ourselves to invariant prior distributions on Ω. Let $\tau_0(\theta_0) = g_{H_0}$ be given. Then, the alternative set $\Omega-\theta_0$ can be divided into N mutually disjoint subsets $A_L,\ L=0,1,\ldots,N-1$ which are closed with respect to all the permutations of $\mathbf{x}$. Namely, $\Omega-\theta_0=\bigcup_{L=0}^{N-1}A_L$ where

$$A_0=\Big\{\theta_i\in\Omega:\ f_i(\mathbf{x})=\prod_{j=1}^{N}g(x_j)\Big\}$$
$$A_1=\Big\{\theta_i\in\Omega:\ f_i(\mathbf{x})=f(x_j)\prod_{k\neq j}g(x_k)\ j=1,2,\ldots,N\Big\}$$
$$\vdots$$
$$A_L=\Big\{\theta_i\in\Omega:\ f_i(\mathbf{x})=\prod_{l=1}^{L}f(x_{j_l})\prod_{k\neq j_l}g(x_k)\Big\}$$
$$\vdots$$
$$A_{N-1}=\Big\{\theta_i\in\Omega:\ f_i(\mathbf{x})=g(x_j)\prod_{k\neq j}f(x_k)\ j=1,2,\ldots,N\Big\}.$$

It is evident that $|A_L|=\binom{N}{L}$ and every invariant aprior distribution τ_0 on Ω is in the unique way determined by the numbers

$$q_{H_0}=\tau_0(\theta_0),\ q_0=\tau_0(A_0),\ldots,\ q_L=\tau_0(A_L),\ldots,\ q_{N-1}=\tau_0(A_{N-1}).$$

Let us name the subsets $A_L,\ L=0,1,\ldots,N-1$ as orbits. Then an invariant prior distribution on Ω is uniform on every orbit and vice versa. As long as we consider invariant prior distributions on Ω only then the Bayesian decision rule can be expressed in the form: if

$$q_{H_0}f_0(\mathbf{x})>\sum_{L=0}^{N-1}\frac{Q_L}{\binom{N}{L}}\sum_{j=1}^{\binom{N}{L}}f_{L_j}(\mathbf{x}),\qquad (**)$$

then the hypothesis H_0 is not rejected. The function $f_{L_j}(\cdot)$ is a density function consisting of L marginal densities $f(\cdot)$ and $N-L$ densities $g(\cdot)$. Every function

$$\sum_{j=1}^{\binom{N}{L}}f_{L_j}(\mathbf{x})$$

is symmetric in $\mathbf{x}$ because its value does not change under an arbitrary permutation applied to $\mathbf{x}$. The maximal invariant statistic with respect to the permutation group is the rank statistic, hence it would be suitable to find such a decision rule that can be expressed by means of the rank statistic. This demand is fulfilled by the rule (**) because the right side of (**) is a symmetric function, hence a function of the rank statistic.

Let us denote by α probability of the first kind error, then evidently

$$\alpha = 1 - E_{\theta_0}\{\Phi(H_0 \mid \mathbf{x})\} = \int_K f_0(\mathbf{x})\,d\mathbf{x},$$

where $K = \{\mathbf{x} : \; q_{H_0} f_0(\mathbf{x}) \leqslant \sum_{j>0} p_j f_j(\mathbf{x})\}$.

Probability $\beta(\theta_i)$ of the second kind error equals $1 - E_{\theta_1}\Phi(H_0 \mid \mathbf{x})$. This error has, of course the form of the following integral

$$\beta(\theta_i) = \int_{K^c} f_i(\mathbf{x})\,d\mathbf{x}.$$

This fact immediately implies that the error of the second kind is constant on every orbit. In order to find out properties of the suggested test we must study the behaviour of $\beta(\cdot)$. Let us assume for simplicity that $f(\mathbf{x}) > 0$ everywhere on reals. Then

$$K^c = \left\{\mathbf{x} : \; 1 > \sum_{j=1}^{2^N-1} \frac{p_j f_i(\mathbf{x})}{q_{H_0} f_0(\mathbf{x})}\right\}.$$

We can conclude from this fact that for every $\mathbf{x} \in K^c$

$$\frac{f_i(\mathbf{x})}{f_0(\mathbf{x})} < \frac{q_{H_0}}{p_j}.$$

If $\theta_i \in A_{N-1}$ then

$$\beta(\theta_i) = \beta(A_{N-1}) = \int_{K^c} \frac{g(x_1)}{f(x_1)} f_0(\mathbf{x})\,d\mathbf{x} \leqslant N \cdot \frac{q_{H_0}}{q_{N-1}}(1-\alpha).$$

Similarly for $\theta_i \in A_{N-2}$

$$\beta(\theta_i) = \beta(A_{N-2}) = \int_{K^c} \frac{g(x_1)g(x_2)}{f(x_1)f(x_2)} f_0(\mathbf{x})\,d\mathbf{x} \leqslant \frac{q_{H_0}}{q_{N-1}} \cdot N\beta(A_{N-1}).$$

Finally, we can obtain that

$$\beta(A_0) \leqslant \frac{q_{H_0}}{q_{N-1}} \cdot N\beta(A_1).$$

From this the following conclusion follows: if $\lim_{N\to\infty} N \cdot q_{N-1} = 0$ then all the errors $\beta(A_L)$ tend to zero too with the increasing number of observations. The behaviour of the second kind errors is determined by the behaviour of $\beta(A_{N-1})$ that is quite natural because the subset A_{N-1} contains the density functions of the type $g(x_j)\prod_{k\neq j} f(x_k)$ only, which are "the most similar" to the hypothesis $f_0(\mathbf{x})$. There is no surprise that the quality of this test is given just by the behaviour of $\beta(A_{N-1})$.

Since the second kind errors $\beta(A_L)$, $L = 0, 1, \ldots, N-1$ possess very pleasant property, they are constant on their orbits we can consider the following test having a hierarchical structure. The Hypothesis H_0 is again given by the one-element subset $\{\theta_0\} \in \Omega$, but instead the composed alternative hypothesis $\Omega - \{\theta_0\}$ we shall consider N quite independent tests: the hypothesis H_0 against every orbit A_L separately. For complexity, we repeat that

$$A_L = \Big\{\theta \in \Omega : \leftrightarrow \prod_{j=1}^{L} f(x_j) \prod_{k\neq j} g(x_k)\Big\}.$$

As we know that every orbit is invariant with respect the permutation group, we must consider an invariant prior distribution on A_L only, i.e., if $\tau_0(H_0) = p_{H_0}$ then

$$\tau(\theta) = \frac{1 - p_{H_0}}{\binom{N}{L}}$$

for every $\theta \in A_L$. In other words spoken we shall test the hypothesis H_0 that all the random variables have the same density function $f(\cdot)$ against the alternative hypothesis among the observed variables one can find just L variables having the density function $g(\cdot)$. It seems to be reasonable to proceed in the following manner. First we test the hypothesis H_0 against the alternative hypothesis A_{N-1}. As long as we reject H_0 we need not continue because with high

probability among all the observations one observation is different. In case we don't reject H_0 we can in the second step test H_0 against A'_{N-2} and so on. By means of this hierarchical test we can reach the last alternative hypothesis A_0, where all the observations are distributed according to the probability density function $g(\cdot)$. It is clear that Bayesian test of the hypothesis H_0 against the alternative hypothesis A_L by $0-1$ loss function has the following form:
if

$$p_{H_0} f_0(\mathbf{x}) < \frac{1-p_{H_0}}{\binom{N}{L}} \sum_j f_j(\mathbf{x}),$$

then the hypothesis H_0 is rejected. The following lemma describes the asymptotic behaviour of the second kind error for the test given above.

Lemma 1. *The second kind error $\beta(A_L)$ of Bayesian test comparing the hypothesis H_0 against the alternative hypothesis A_L satisfies the relation:*

$$\text{if} \quad \lim_{N\to\infty} p_{H_0} = 0 \quad \text{then} \quad \lim_{N\to\infty} \beta(A_L) = 0,$$

where N is the number of observations.

Proof. Although A_L is not simple alternative hypothesis, as we know from the previous part, despite of this fact for every $\theta_i \in A_L$

$$\beta(\theta_i) = \int_{K_L^c} f_i(\mathbf{x})\,d\mathbf{x} = \beta(A_L)$$

is not depending on θ_i,

$$K_L = \Big\{\mathbf{x}: \; p_{H_0} f_0(\mathbf{x}) \leqslant \sum_{\theta_i \in A_L} \frac{1-p_{H_0}}{\binom{N}{L}} f_i(\mathbf{x})\Big\}.$$

Let α be the first kind error, i.e.,

$$\alpha = \int_{K_L} f_0(\mathbf{x})\,d\mathbf{x}.$$

Then $1-\alpha = \int_{K_L^c} f_0(\mathbf{x})\,d\mathbf{x}$ that is precisely the distribution function value of the random variables $S_L(\cdot)$ at the point $p_{H_0}/(1-p_{H_0})$ under the hypothesis H_0 if

$$S_L(\mathbf{x}) = \frac{1}{\binom{N}{L}} \frac{g(x_{i_1})\cdots g(x_{i_{N-L}})}{f(x_{i_1})\cdots f(x_{i_{N-L}})}.$$

Thanks to the left continuity of any probability distribution function one can find in every case to a chosen value $\alpha \in (0,1)$ such a number $p_{H_0}/(1-p_{H_0})$ so that

$$\int_{K_L^c} f_0(\mathbf{x}) \geqslant 1-\alpha.$$

Then the second kind error $\beta(A_L)$ equals

$$\beta(A_L) = \int_{K_L^c} f_i(\mathbf{x})\,d\mathbf{x} = F_{\theta_i}\Big(\frac{p_{H_0}}{1-p_{H_0}}\Big),$$

where $F_{\theta_i}(\cdot)$ is the distribution function of $S_L(\cdot)$ under the alternative hypothesis A_L. One can easily prove that

$$\beta(A_L) = \int_{\{\mathbf{x}: p_{H_0}/(1-p_{H_0}) > S_L(\mathbf{x})\}} S_L(\mathbf{x}) f_0(\mathbf{x})\,d\mathbf{x}.$$

Then, of course

$$\beta(A_L) \leqslant \frac{p_{H_0}}{1-p_{H_0}} \int_{K_L^c} f_0(\mathbf{x})\,d\mathbf{x} \leqslant \frac{p_{H_0}}{1-p_{H_0}}.$$

This inequality implies immediately that $\lim_{N\to\infty} \beta(A_L) = 0$ if $\lim_{N\to\infty} p_{H_0} = 0$. Q.E.D.

REMARK. In order to reach $\lim_{N\to\infty} p_{H_0} = 0$ it is sufficient to consider the uniform prior distribution. Then $\beta(A_L) \leqslant 1/N$, where N is the number of observations.

At this place it is very important to mention a special case of the hypothesis testing, which is very close to the studied questions. This problem deals with the detection of changes in the behaviour of time series. The simplest considered case presents a sequence of mutually independent random variables that are distributed according to a probability density function $f(\cdot)$ before a change and according to the density $g(\cdot)$ after a change. A special case of this statistical task was studied already by Page (1954), namely a possible jump in the mean value of Gaussian random variables. Here we have the testing of a simple hypothesis "no jump" against a composed alternative hypothesis "a jump occurred" at a time instant during the observation. Deshayes and Picard proved in Basseville, Benveniste (1986) that in this case there is no uniform best test because Neyman–Pearson lemma doesn't hold. From this reason the behaviour of the second kind error is extremely interesting for every suggested test. If we consider Bayesian test based on $0-1$ loss function then we achieve an analogous result as before; the hypothesis H_0 "no change" is rejected if

$$p_0 f_0(\mathbf{x}) < \sum_{j=1}^{N} p_j f_j(\mathbf{x}),$$

where $f_j(\mathbf{x}) = \prod_{i=1}^{j-1} f(x_i) \prod_{j=1}^{N} g(x_i)$ under a prior distribution $\{p_j\}_{j=1}^{N}$. Then the j-th second kind error is equal to

$$\beta_j = \int_{K^c} f_j(\mathbf{x})\, d\mathbf{x},$$

$K = \{\mathbf{x} : p_0 f_0(\mathbf{x}) \leqslant \sum_{j=1}^{M} p_j f_j(\mathbf{x})\}$. As it is reasonable to consider $p_0 > 0$, $f_0(\mathbf{x}) > 0$, one case rewrite the rejection rule into the form

$$1 < \sum_{j=1}^{N} \frac{p_j}{p_o} \frac{f_j(\mathbf{x})}{f_0(\mathbf{x})},$$

i.e.,

$$1 < \sum_{j=1}^{N} \frac{p_j}{p_0} \frac{g(x_j)g(x_{j+1}) \cdots g(x_N)}{f(x_j)f(x_{j+1}) \cdots f(x_N)}.$$

If the prior distribution $\{p_j\}_{j=1}^{N}$ is uniform we can write

$$1 < \frac{g(x_N)}{f(x_N)} \left(\sum_{j=1}^{N-1} \frac{g(x_j)g(x_{j+1}) \cdots g(x_N)}{f(x_j)f(x_{j+1}) \cdots f(x_N)} + 1 \right).$$

In this way we obtained a recurrent decision rule

$$1 < T_N(\mathbf{x}_n) = \frac{g(x_N)}{f(x_N)}(T_{N-1}(\mathbf{x}_{N-1}) + 1)$$

with $T_0(\mathbf{x}_0) = 0$.

We proceed in our decision making so long as long as we overstep the threshold 1 at the first time by $T_N(\mathbf{x}_N)$. In this way we defined a stopping rule change is detected if $T_N(\mathbf{x}_N) \geqslant 1$. One can prove simply that under the hypothesis H_0

$$E_{H_0}\{T_N(\mathbf{x}_N) \mid (\mathbf{x}_{N-1})\} = T_{N-1}(\mathbf{x}_{N-1}) + 1.$$

We see that $\{T_N(\mathbf{x}_N)\}_{N=0}^{\infty}$ forms a nonnegative submartingal with

$$E_{H_0}\{T_N(\mathbf{x}_N)\} = N.$$

This fact implies unfortunately that the considered test is very strong and we can expect almost in every case the rejection of H_0. We see that Bayesian global test comparing the simple hypothesis H_0 against the composed alternative hypothesis $\Omega - H_0$ does not posses suitable properties and hence we must consider a hierarchical test again, which tests H_0 versus simple alternative hypothesis gradually. Every simple alternative hypothesis consists in the assumptions a change can only occur at one of time instants among $1, 2, \ldots, N$. Then H_0 is rejected if

$$p_0 f_0(\mathbf{x}_N) < p_L f_L(\mathbf{x}_N),$$

where $p_0 + p_L = 1$, $f_0(\mathbf{x}_N) = \prod_{i=1}^{N} f(x_i)$, $f_L(\mathbf{x}_N) = \prod_{i=1}^{L-1} f(x_i) \times \prod_{j=L}^{N} g(x_j)$. Under the condition $f_0(\mathbf{x}_N) > 0$ the testing rule can be expressed as

$$1 < \frac{p_L}{p_o} \prod_{j=L}^{N} \frac{g(x_j)}{f(x_j)} = \tau_L(\mathbf{x}_N)$$

and then we have

$$E_{H_0}\{\tau_L(\mathbf{x}_N)\} = \frac{p_L}{p_0}$$

for every $L = 1, 2, \ldots, N$.

The asymptotic behaviour of the first and second kind errors is given in the following

Lemma 2. *Let α_L and β_L be the first and second kind errors of the test mentioned above in case $p_0 = p_l = \frac{1}{2}$. Then*

$$\lim_{N\to\infty}(\alpha_L + \beta_L)^{1/R_N^L(\frac{1}{2})} = \exp\{p^*\},$$

where $R_N^L(\frac{1}{2}) = -2(N-L)\ln\int_{-\infty}^{+\infty} g^{1/2}(x)f^{1/2}(x)dx$ and

$$\rho^* = \min_{a\in(0,1)}\left\{-\frac{1}{2}\frac{\ln g^a(x)f^{1-a}(x)dx}{\ln g^{1/2}(x)f^{1/2}(x)dx}\right\},$$

N is the number of observations.

Proof. Let us denote by $p_{N,L}(\mathbf{x}) = \prod_{i=1}^{L-1} f(x_i)\prod_{j=L}^{N} g(x_j)$ and $g_N(\mathbf{x}) = \prod_{j=1}^{N} f(x_j)$. Then the Hellinger integral $H_N(a)$ equals

$$H_N(a) = \int_{-\infty}^{+\infty}\left(\frac{p_{N,L}(\mathbf{x})}{q_N(\mathbf{x})}\right)^a g_N(\mathbf{x})d\mathbf{x} = H_1^{N-L}(a),$$

where

$$H_1(a) = \int_{-\infty}^{+\infty} g^a(x)f^{1-a}(x)dx \qquad \text{for } a\in\langle 0,1,\rangle.$$

This fact immediately follows from the independence of observations. Then the corresponding Ré nyi distance $R_N(a)$ satisfies the relation

$$R_N(a) = \frac{(N-L)\ln H_1(a)}{a-1}$$

for $a \in \langle 0, 1)$ and

$$R_N(1) = (N-L)\int_{-\infty}^{+\infty}\ln\frac{g(x)}{f(x)}f(x)dx$$

for $a = 1$. There is no problem to show that the given statistical model

$$\left\{\frac{p_{N,L}(\mathbf{x})}{q_N(\mathbf{x})}\right\}_{N=1}^{\infty}$$

can be understood as a martingale. Let us imagine that with the increasing number of observations N the number L presenting the observations before a change is increasing too such that as $N \to \infty$, $L \to \infty$ also but

$$\lim_{N\to\infty} \frac{L}{N} = a \in (0,1).$$

After these assumptions one can use the generalized Chernoff theorem from Vajda (1990) dealing with the behaviour of the sum $\alpha_L + \beta_L$ for Bayesian test. The statement of Lemma 2 is a simple application of the mentioned theorem. Q.E.D.

REMARK. In the other words speaking, Lemma 2 states that for sufficiently large N, L the sum $\alpha_L + \beta_L$ can be estimated below and above as follows

$$e^{(L-N)\ln H_1^2(\frac{1}{2})(\rho^*-\varepsilon)} < \alpha_L + \beta_L < e^{(L-N)\ln H_1^2(\frac{1}{2})(\rho^*+\varepsilon)}.$$

This inequality gives another possibility, namely

$$\alpha_L + \beta_L \leqslant \exp\left\{N(1-\alpha)\ln H_1\left(\frac{1}{2}\right)\right\}.$$

Since $H_1(\frac{1}{2}) = \int_{-\infty}^{+\infty} f^{\frac{1}{2}}(x) g^{\frac{1}{2}}(x)\,dx < \frac{1}{2}\int_{-\infty}^{+\infty} f(x) + g(x)\,dx = 1$, then

$$\alpha_L + \beta_L \to 0$$

exponentially as $N \to \infty$.

There is another possibility how to construct a test comparing the hypothesis H_0 against the composed alternative hypothesis consisting of all the $2^N - 1$ possible cases. Let us denote by $f_0(\mathbf{x})$ again the density function due to the hypothesis and by $f_i(\mathbf{x})$, $i = 1, 2, \ldots, 2^N - 1$ all the densities from the composed alternative hypothesis. We wish to make a decision d_j under the observation $\mathbf{x}$ that $\mathbf{x}$ was realized according to the probability density function $f_j(\mathbf{x})$. Let the loss function $\ell(\theta_i, d_j)$ be given by

$$\ell(\theta_j, \delta_i) = 1 - \delta_{ij},$$

where $\theta_j \leftrightarrow f_j(\cdot)$ and δ_{ij} is the Kronecker symbol. We look for Bayesian decision function $\{\varphi(i \mid \mathbf{x})\}_{i=1}^{2^N-1}$ satisfying

$$\sum_{i=0}^{2^N-1} \varphi(i \mid \mathbf{x}) = 1 \qquad (***)$$

and $\varphi(i \mid \mathbf{x})$ means the conditional probability accepting d_i under the condition $\mathbf{x}$. Let τ be a prior distribution on the parametric space $\Omega = \{\theta_i : i = 0, 1, \ldots, 2^N - 1\}$, i.e..

$$\tau(\theta_i) = p_i.$$

Then the conditional risk function can be expressed as

$$R(\theta_i, \varphi(\cdot \mid \mathbf{x})) = \sum_{j=0}^{2^N-1} \ell(\theta_i, d_j) E_{\theta_i}\{\varphi(j \mid \mathbf{x})\}$$
$$= 1 - E_{\theta_i}\{\varphi(i \mid \mathbf{x})\}.$$

Using this fact we can calculate the conditional average risk with respect to the prior distribution τ, namely

$$r(\tau, \varphi) = 1 - \sum_{j=0}^{2^N-1} p_i E_{\theta_i}\{\varphi(i \mid \mathbf{x})\}.$$

Now, Bayesian procedure must minimize the quantity $r(\tau, \varphi)$. The answer is given by the following decision rule

$$\Phi(\mathbf{x}) = \{\Phi(i \mid \mathbf{x}) : \; i = 0, 1, \ldots, 2^N - 1\},$$

where

$$\Phi(i \mid \mathbf{x}) = 0 \quad \text{iff} \quad p_i f_i(\mathbf{x}) < \max_{0 \leqslant j \leqslant 2^N-1} \{p_j f_j(\mathbf{x})\}.$$

In other words, if $p_{i_0} f_{i_0}(\mathbf{x}) > p_j f_j(\mathbf{x})$ for every $j \in \{0, 1, \ldots, 2^N - 1\} - \{i_0\}$, then $\Phi(i_0 \mid \mathbf{x}) = 1$ because we demand $(***)$.

The proof of the optimality for this decision rule is very simple and is based on the results from Hoel, Peterson (1949).

Next, we will apply this general approach to the problem of change detections already considered earlier, where the jump from the density function f to the density function g is to be recorded. Under a suitable choice of a prior distribution on the parameter space Ω we can admit only those parameters $\theta_j \in \Omega$, $j = 0, 1, \ldots, N$ with $\tau(\theta_j) = p_j > 0$ corresponding to possible jumps at time instants, $j = 1, 2, \ldots, N$, i.e.,

$$\theta_j \longleftrightarrow \varphi_j(\mathbf{x}) = \prod_{i=1}^{j-1} f(x_i) \prod_{i=j}^{N} g(x_i).$$

Then we obtain the decision rule stating we accept the alternative hypothesis θ_{j_0} if and only if

$$p_{j_0}\varphi_{j_0}(\mathbf{x}) > \max_{\substack{0 \le i \le N \\ i \ne j_0}} p_i\varphi_i(\mathbf{x}).$$

We shall consider two simple examples for illustration.

The first case considers a possible jump in mean value of Gaussian random variables, i.e.,

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x^2\right\}, \qquad g(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x+\delta)^2\right\},$$

where for simplicity $\delta > 0$. Then, one can easily find out that the alternative hypothesis j_0 is accepted if and only if

$$\frac{1}{j_0 - 1} \sum_{i=1}^{j_0-1} x_i < \frac{\delta}{2}$$

and at the same time for every $i = j_0 + 1, j_0 + 2, \ldots, N + 1$

$$\frac{1}{i - j_0} \sum_{i=j_0}^{i-1} x_i > \frac{\delta}{2}.$$

If we put $j_0 = N$ then we obtain the following stopping rule: a change at time instant N is recorded if

$$\max(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_{N-1}) < \frac{\delta}{2}$$

but

$$x_N > \frac{\delta}{2},$$

where

$$\bar{x}_i = \frac{1}{N - i + 1} \sum_{l=i}^{N} x_l, \qquad i = 1, 2, \ldots, N - 1.$$

The other example considers Gaussian random variable too, but with possible density functions

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \qquad g(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}.$$

Then, the alternative hypothesis j_0 will be accepted if for every $i = 1, 2, \ldots, j_0 - 1$

$$\frac{1}{j_0 - i} \sum_{l=i}^{j_0-1} x_l^2 < \frac{\ln \sigma^2}{1 - \frac{1}{\sigma^2}}$$

and simultaneously for every $i = j_0 + 1, j_0 + 2, \ldots, N + 1$

$$\frac{\ln \sigma^2}{1 - \frac{1}{\sigma^2}} < \frac{1}{i - j_0} \sum_{l=j_0}^{i-1} x_l^2.$$

REFERENCES

The list of references is given in Part II.