Aplikace matematiky

van Huu Nguyen

A rank decision rule for a combined problem of testing and classification

Aplikace matematiky, Vol. 19 (1974), No. 3, 152–168

Persistent URL: http://dml.cz/dmlcz/103526

A RANK DECISION RULE FOR A COMBINED PROBLEM OF TESTING AND CLASSIFICATION

NGUYEN VAN HUU

(Received September 6, 1971)

I. INTRODUCTION

The problem of testing the null hypothesis H_0 against the alternatives $K_1, K_2, \ldots \ldots, K_s$, where under H_0 the observations $X_1, \ldots, X_N$ are independent, identically distributed with the common density $f \in \mathscr{F}$ (some family of density functions), and under K_i the observations $X_1, \ldots, X_N$ have the densities

$$f_1(x) = f(x, \Delta C_{i1}), \ldots, f_N(x) = f(x, \Delta C_{iN}) \tag{1}$$

with respect to the Lebesgue measure, respectively, for $i = 1, \ldots, s$, where $f(x, 0) = f(x)$, has been considered in [7]. However, one sometimes encounters the situations where one has to decide which alternative of $K_1, \ldots, K_s$ is true when H_0 has been rejected.

Therefore, let us construct a procedure, which allows us at first to test the hypothesis H_0, then if H_0 has been rejected, to decide which of $K_1, \ldots, K_s$ is true. The problem was investigated by Pfanzagl [9]. The slipage problem – a special case of the combined problem of testing and classification was investigated by Mosteller [6], Paulson [8], Truax [11], Karlin - Truax [5], Hall - Kudo [3], Hall - Kudo - Yeh [4].

The most of procedures suggested by these authors are based directly on observations rather than on ranks. On the other hand, the procedure proposed in this work is based on ranks.

A combined decision rule for testing and classification may be characterized by a vector-valued function

$$\varphi(x) = \{\varphi_0(x), \varphi_1(x), \ldots, \varphi_s(x)\}$$

where $\varphi_0(x)$, $\varphi_i(x)$ are the probabilities of accepting H_0 and K_i, respectively, $i = 1, \ldots, s$, when x is a realization of the random vector $X = (X_1, \ldots, X_N)$. The functions $\varphi_i(x)$, $i = 0, 1, \ldots, s$, have to satisfy the following conditions:

$$\varphi_i(x) \geqq 0 \quad \text{for all } i\,, \quad \sum_{i=0}^{s} \varphi_i(x) = 1 \quad \text{for all } x\,.$$

Let $P_0, P_{1,\Delta}, \ldots, P_{s,\Delta}$; $\mathsf{E}_0, \mathsf{E}_{1,\Delta}, \ldots, \mathsf{E}_{s,\Delta}$ be the probabilities and the expectations under $H_0, K_1, \ldots, K_s$, respectively.

We say that a decision rule has the significance level $1 - \alpha$ if

$$\mathsf{E}_0\, \varphi_0(x) = \int \varphi_0(x)\, \mathrm{d}P_0(x) \geqq 1 - \alpha\,. \tag{2}$$

We shall try to find a $(1 - \alpha)$-level decision rule $\varphi(x)$ such that for some values of Δ

$$\sum_{i=1}^{s} \mathsf{E}_{i,\Delta}\, \varphi_i(X) \geqq \sum_{i=1}^{s} \mathsf{E}_{i,\Delta}\, \varphi'_i(X) \tag{3}$$

for any other $(1 - \alpha)$-level decision rule $\varphi'(x) = \{\varphi'_0(x), \varphi'_1(x), \ldots, \varphi'_s(x)\}$.

Definition. *A decision rule satisfying* (3) *is said to be optimal for the combined problem of testing and classification.*

II. A GENERALIZATION OF NEYMAN-PEARSON'S LEMMA

Assume that the probability measures $P_0, P_1, \ldots, P_s$ have the densities $f_0(x), f_1(x), \ldots, f_s(x)$ with respect to a σ-finite measure μ defined on the space $\mathcal{X}$ of the sample values $x = (x_1, \ldots, x_N)$.

Denote the expectations with respect to $P_0, P_1, \ldots, P_s$ by $\mathsf{E}_0, \mathsf{E}_1, \ldots, \mathsf{E}_s$. Consider a decision rule of the form:

$$\begin{aligned} \varphi_0(x) &= 1\,,\ \xi_0(x)\,,\ 0 \quad \text{if} \quad \max_{1 \leqq i \leqq s} f_i(x) <\,,\ =\,,\ > C f_0(x)\,, \\ \varphi_j(x) &= \quad\ \ \xi_j(x)\,,\ 0 \quad \text{if} \quad \max_{1 \leqq i \leqq s} f_i(x) \quad =\,,\ > \quad f_j(x)\,, \\ &\qquad\qquad \text{for} \quad j = 1, \ldots, s\,, \end{aligned} \tag{4}$$

where $\xi_0(x), \xi_1(x), \ldots, \xi_s(x)$ are arbitrary but subject to the condition that φ is a decision rule and C is some constant.

Let φ' be another decision rule. Then:

(i) If $\mathsf{E}_0\, \varphi'_0(X) \geqq \mathsf{E}_0\, \varphi_0(X)$ then $\sum_{i=1}^{s} \mathsf{E}_i\, \varphi'_i(X) \leqq \sum_{i=1}^{s} \mathsf{E}_i\, \varphi_i(X)\,.$

(ii) If $\mathsf{E}_0\, \varphi_0'(X) \geqq \mathsf{E}_0\, \varphi_0(X)$ and $\sum_{i=1}^{s} \mathsf{E}_i\, \varphi_i'(X) = \sum_{i=1}^{s} \mathsf{E}_i\, \varphi_i(X)$ then φ' has the form (4) a.e.. Furthermore, $\mathsf{E}_0\, \varphi_0'(X) = \mathsf{E}_0\, \varphi_0(X)$ unless $\mathsf{E}_j\, \varphi_0(X) = 0$ for all j.

(iii) For every $\alpha \in (0, 1)$ there is a decision rule of the form (4) with $\xi_0(x)$ constant, say ξ_α, such that

$$\mathsf{E}_0\, \varphi_0(X) = 1 - \alpha\,. \tag{5}$$

(See Theorem 1 in [3].)

III. LOCALLY OPTIMAL RANK DECISION RULE

Let $R = (R_1, \ldots, R_N)$ be the vector of the ranks of $X_1, \ldots, X_N$ and $r = (r_1, \ldots, r_N)$ be a realization of R.

Theorem 1. *Let $H_0, K_1, \ldots, K_s$ be the hypothesis and the alternatives defined by* (1). *Suppose that $f(x, \theta)$ involved in* (1) *satisfies the following conditions*:

$$\lim_{\theta \to 0} [f(x, \theta) - f(x)]/\theta = \dot{f}(x, 0) \tag{A$_1$}$$

holds and $f(x, \theta)$ is absolutely continuous on some open interval containing the point 0.

$$\lim_{\theta \to 0} \int_{-\infty}^{\infty} |\dot{f}(x, \theta)|\, dx = \int_{-\infty}^{\infty} |\dot{f}(x, 0)|\, dx < \infty\,, \tag{A$_2$}$$

where $\dot{f}(x, \theta) = \partial f(x, \theta)/\partial\theta$.

Define the decision rule $\varphi(R) = \{\varphi_0(R), \varphi_1(R), \ldots, \varphi_s(R)\}$ by

$$\begin{aligned} \varphi_0(R) &= 1,\ \xi_\alpha,\ 0 \quad \text{if} \quad \max_{1 \leqq i \leqq s} T_i(R) < \,,\, = \,,\, > C_\alpha\,, \\ \varphi_j(R) &= \xi_j(R),\ 0 \quad \text{if} \quad \max_{1 \leqq i \leqq s} T_i(R) = \,,\, > T_j(R) \quad \text{for} \quad j = 1, \ldots, s\,, \end{aligned} \tag{6}$$

where

$$T_i(r) = \sum_{k=1}^{N} C_{ik}\, \mathsf{E}_0\{\dot{f}(X^{(r_k)}, 0)/f(X^{(r_k)})\} \tag{7}$$

with $X^{(1)} < X^{(2)} < \ldots < X^{(N)}$ being the ordered observations $X_1, X_2, \ldots, X_N$, and where C_α, ξ_α, are defined so that $\mathsf{E}_0\, \varphi_0(R) = 1 - \alpha$, $\xi_1(R), \ldots, \xi_s(R)$ are arbitrary but subject to the condition that φ is a decision rule.

This decision rule is locally optimal within the class of all $(1 - \alpha)$-level rank decision rules in the sense that there exists an $\varepsilon > 0$ such that φ maximizes

$\sum_{i=1}^{s} \mathsf{E}_i\, \varphi'_i(R)$ *uniformly for all* $0 < \Delta \leqq \varepsilon$ *within the class* $\{\varphi'(R)\}$ *of all possible decision rules depending only on the ranks of* $X_1, \ldots, X_N$.

Proof. First consider the combined problem of testing and classification of H_0 against $K_1, \ldots, K_s$ with Δ fixed. Put $B_r = \{X; R = r\}$. The sub-σ-field generated by R, say $\mathscr{B}_0$, consists of all unions of such events, while the rest of the space $\mathscr{X}$ where some coordinates coincide and R is not defined may be neglected, since its probability is zero under all distributions determined by densities.

We have, under H_0, $P_0\{B_r\} = P_0\{R = r\} = 1/N!$.

Introduce on $\mathscr{B}_0$ the so-called counting measure μ which is defined by $\mu(B_r) = 1$. Then $P_{i,\Delta}\{B_r\}$, $i = 1, \ldots, s$, $P_0\{B_r\}$ may be regarded as a density of $P_{i,\Delta}$ and P_0 with respect to the counting measure μ on the sub-σ-field $\mathscr{B}_0$. Applying the above generalized Neyman - Pearson's Lemma we obtain the optimal decision rule within the class of all possible rules based on ranks for the combined problem of testing and classification with each Δ fixed. The decision rule is given by:

$$(8)\quad \varphi_0(r) = 1,\ \xi_\alpha,\ 0 \quad \text{if} \quad \max_{1 \leqq i \leqq s} P_{i,\Delta}\{R = r\} < , = , > C'_\alpha ,$$

$$\varphi_j(r) = \xi_j(r),\ 0 \quad \text{if} \quad \max_{1 \leqq i \leqq s} P_{i,\Delta}\{R = r\} = , > P_j\{R = r\} \quad \text{for} \quad j = 1, \ldots, s .$$

On the other hand, if $f(x, \theta)$ has the properties (A_1), (A_2) then, by the proof of Theorem 1 in [7], we obtain

$$P_{i,\Delta}\{R = r\} = 1/N! + (\Delta/N!) \sum_{k=1}^{N} C_{ik}\, \mathsf{E}_0[\dot{f}(X^{(r_k)}, 0)/f(X^{(r_k)})] + o(\Delta)$$

$$= 1/N! + (\Delta/N!)\, T_i(r) + o(\Delta) .$$

Consequently, there is an $\varepsilon > 0$ such that (8) is equivalent to (6) for $0 < \Delta \leqq \varepsilon$. This completes the proof.

IV. THE ASYMPTOTIC DETERMINATION OF CRITERIA

In general, the determination of the constants ξ_α and C_α, even the asymptotic determination, is very difficult. Therefore, let us try to make an asymptotic determination of the constants for some special cases, which are, however, important in practice, by examining the asymptotic behaviour of the distribution of the statistic $\max_{1 \leqq i \leqq s} T_i(R)$.

1. The s-sample problem of slippage. Let $X_1, \ldots, X_N$ be independent observations, and let $f_1(x), \ldots, f_N(x)$ be their densities, respectively, with respect to the Lebesgue measure.

Consider the hypothesis H_0 and $K_1, \ldots, K_s$ where

$$H_0 : f_j(x) = f(x) \quad \text{for all} \quad j = 1, \ldots, N\,, \tag{9}$$

$$\begin{aligned} K_i : f(x) &= f(x, \Delta) \quad \text{for all} \quad j \in I_i\,, \\ &= f(x) \quad \text{for all} \quad j \notin I_i\,, \quad i = 1, \ldots, s \end{aligned} \tag{10}$$

with

$$\begin{aligned} I_1 = \{1, 2, \ldots, n_1\}\,;\quad I_2 = \{n_1 + 1, n_1 + 2, \ldots, n_1 + n_2\}, \ldots, I_s = \\ = \{n_1 + \ldots + n_{s-1} + 1, \ldots, n_1 + \ldots + n_s\} \end{aligned}$$

where $n_1 + \ldots + n_s = N$. Thus $\operatorname{card}(I_i) = n_i$.

Assume that

$$\lim_{N \to \infty} n_i/N = \lambda_i \quad (0 < \lambda_i < 1 \text{ for all } i)\,. \tag{11}$$

These alternatives are only a special case of the regression aiternatives defined by (1) with

$$\begin{aligned} C_{ij} &= 1 \quad \text{for all} \quad j \in I_i\,, \\ &= 0 \quad \text{for all} \quad j \notin I_i\,;\quad i = 1, 2, \ldots, s\,. \end{aligned} \tag{12}$$

Applying Theorem 1 we obtain the locally optimal rank decision rule for the combined of testing and classification of H_0 against $K_1, \ldots, K_s$. The decision rule is defined by

$$\begin{aligned} \varphi_0(R) &= 1,\ \xi_\alpha,\ 0 \quad \text{if} \quad \max_{1 \leqq i \leqq s} T_i(R)/\sqrt{N} <\,,\ =\,,\ > C_\alpha\,, \\ \varphi_j(R) &= \xi_j(R),\ 0 \quad \text{if} \quad \max_{1 \leqq i \leqq s} T_i(R) =\,,\ > T_j(R) \quad \text{for} \quad j = 1, \ldots, s\,, \end{aligned} \tag{13}$$

where

$$T_i(R) = \sum_{k \in I_i} \mathsf{E}_0[\dot{f}(X^{(r_k)}, 0)/f(X^{(r_k)})] = \sum_{k \in I_i} a_N(r_k, f) \tag{14}$$

with

$$\begin{aligned} a_N(k, f) &= \mathsf{E}\, \varphi(U^{(k)}, f)\,, \quad k = 1, \ldots, N\,, \\ \varphi(u, f) &= \dot{f}(F^{-1}(u), 0)/f(F^{-1}(u)) \end{aligned} \tag{15}$$

where $U^{(1)} < \ldots < U^{(N)}$ is the ordered sample from the uniform distribution on $(0, 1)$, $F(x)$ is the distribution function corresponding to
$f(x)\,,\quad F^{-1}(u) = \inf\{x;\, F(x) \geqq u\}\,.$

Lemma 1. *Assume that*

$$(16) \qquad 0 < I(f) = \int_0^1 \varphi^2(u, f)\, \mathrm{d}u < \infty\,.$$

Then $\{T_1(R)/\sqrt{N}, \ldots, T_s(R)/\sqrt{N}\}$ *is, under* H_0, *asymptotically degenerated s-variate normal with the mean vector zero and the covariance matrix* $\{\lambda_i(\delta_{ij} - \lambda_j)\, I(f)\}$ *where* $\delta_{ij} = 1, 0$ *as* $i = j$, $i \neq j$, *respectively.*

Proof. Let $Z_1, \ldots, Z_N$ be the normal random vector with the mean vector zero and the covariance matrix defined in Lemma 1. According to Theorem V.2.1 in [2] the assertion of Lemma 1 will be proved if we can show that $\sum_{i=1}^{s} \theta_i\, T_i(R)/\sqrt{N}$ converges in distribution to $\sum_{i=1}^{s} \theta_i Z_i$ for any real numbers θ_i's.

Actually, it is easy to see that

$$\mathsf{E}_0[T_i(R)] = \mathsf{E}Z_i = 0 \quad \text{for all } i\,,$$
$$\operatorname{cov}\left(T_i(R)/\sqrt{N}\,,\;\; T_j(R)/\sqrt{N}\right) \sim (n_i/N)\,(\delta_{ij} - n_j/N)\, N^{-1} \sum_{k=1}^{N} a_N^2(k, f)$$

(see Theorem II.3.1.d in [2] and note that

$$\bar{a}_N = N^{-1} \sum_{k=1}^{N} a_N(k, f) = \sum_{k=1}^{N} \mathsf{E}\{\varphi(U_1, f) \mid R_1 = k\}\, P\{R_1 = k\} =$$
$$= \mathsf{E}\,\varphi(U_1, f) = \int_0^1 \varphi(u, f)\, \mathrm{d}u = \int_{-\infty}^{\infty} f'(x, 0)\, \mathrm{d}x = 0\,,$$

by the conditions (A_1), (A_2), where $R_1, \ldots, R_N$ are the ranks of the observations $U_1, \ldots, U_N$ from the uniform distribution on (0, 1). Consequently, $\operatorname{cov}\left(T_i(R)/\sqrt{N}, T_j(R)/\sqrt{N}\right)$ converges to $\lambda_i(\delta_{ij} - \lambda_j)\, I(f) = \operatorname{cov}(Z_i, Z_j)$ as $N \to \infty$. On the other hand, we have

$$\sum_{i=1}^{s} \theta_i\, T_i(R)/\sqrt{N} = \sum_{k=1}^{N} C_k(\theta)\, a_N(R_k, f) \quad \text{where} \quad C_k(\theta) = \sum_{i=1}^{s} \theta_i (C_{ik}/\sqrt{N})$$

with C_{ik}'s given by (12). Furthermore,

$$\sum_{k=1}^{N} [(C_{ik} - \bar{C}_i)/\sqrt{N}]\, [(C_{jk} - \bar{C}_j)/\sqrt{N}] = (n_i/N)\,(\delta_{ij} - n_j/N) \to \lambda_i(\delta_{ij} - \lambda_j)\,,$$

$\max_{1 \leqq k \leqq N} (C_{ik} - \bar{C}_i)^2/N \to 0$ as $N \to \infty$, with $\bar{C}_i = \sum_{k=1}^{N} C_{ik}/N$ for all $i, j = 1, \ldots, s$. Thus the conditions of Corollary 5 in [7] are fulfilled, and by this corollary $\sum_{i=1}^{s} \theta_i\, T_i(R)/\sqrt{N}$ is asymptotically normal with the mean zero and the variance $\sum_{i,j=1}^{s} \theta_i \theta_j\, \lambda_i(\delta_{ij} - \lambda_j)\, I(f)$, which are the mean and the variance of the normally distributed random variable $\sum_{i=1}^{s} \theta_i Z_i$. Q.E.D.

Let us now return to the asymptotic determination of C_α involved in (13) such that $\mathsf{E}_0\, \varphi_0(R) = 1 - \alpha$. It follows from Lemma 1 that

$$(17) \qquad \lim_{N\to\infty} P\{\max_{1\leq i\leq s} T_i(R)/\sqrt{N} < C_\alpha\} = P\{\max_{1\leq i\leq s} Z_i < C_\alpha\}\,.$$

With no loss of generality we can suppose that $I(f) = 1$. The correlation coefficient between Z_i and Z_j is equal to

$$(18) \qquad \varrho_{ij} = -\lambda_i\lambda_j/[\lambda_i(1-\lambda_i)\,\lambda_j(1-\lambda_j)]^{1/2} < 0\,.$$

Hence, according to Slepian's result [10] and Bonferroni's inequality, we obtain:

$$(19) \quad 1 - \sum_{i=1}^{s} P\{Z_i/\sigma_i \geqq C_i'\} \leqq P\{Z_1 < C_1'\sigma_1, \ldots, Z_s < C_s'\sigma_s\} \leqq \prod_{i=1}^{s} P\{Z_i/\sigma_i < C_i'\}$$

with $\sigma_i^2 = \text{var}\,(Z_i) = \lambda_i(1-\lambda_i)$.

Let $\Phi(x)$ be the standardized normal distribution function. Then Z_i/σ_i has the distribution function $\Phi(x)$. Putting $C' = C/\sigma_i$ we obtain

$$(20) \qquad 1 - \sum_{i=1}^{s}[1 - \Phi(C/\sigma_i)] \leqq P\{Z_1 < C, \ldots, Z_s < C\} \leqq \prod_{i=1}^{s} \Phi(C/\sigma_i)\,.$$

Let C_α, $C_\alpha^{(1)}$, $C_\alpha^{(2)}$ be the constants defined so that

$$(21) \qquad \begin{aligned} P\{\max_{1\leq i\leq s} Z_i < C_\alpha\} &= 1 - \alpha\,, \\ 1 - \sum_{i=1}^{s}[1 - \Phi(C_\alpha^{(1)}/\sigma_i)] &= 1 - \alpha\,, \\ \prod_{i=1}^{s} \Phi(C_\alpha^{(2)}/\sigma_i) &= 1 - \alpha\,; \end{aligned}$$

then we have

$$(22) \qquad C_\alpha^{(1)} \geqq C_\alpha \geqq C_\alpha^{(2)}\,.$$

We can expect that, with α sufficiently small, $C_\alpha^{(1)}$, $C_\alpha^{(2)}$ differ from each other very little. Hence choosing, for example, $C_\alpha = C_\alpha^{(1)}$ we obtain

$$(23) \qquad P\{\max_{1\leq i\leq s} T_i(R)/\sqrt{N} < C_\alpha\} \doteq 1 - \alpha$$

if N is sufficiently large.

The following table provides a comparison of differences between $C_\alpha^{(1)}$ and $C_\alpha^{(2)}$ for the case where $n_1 = n_2 = \ldots = n_s$ (then $\sigma_i^2 = \lambda_i(1-\lambda_i) = (1 - 1/s)/s$) at the significance levels $1 - \alpha = 95\%$ and $1 - \alpha = 90\%$. In this case the constants $C_\alpha^{(1)}$, $C_\alpha^{(2)}$ are defined by the equations

$$(24) \qquad \begin{aligned} \Phi[C_\alpha^{(1)}\,.\, s/\sqrt{(s-1)}] &= 1 - \alpha/s\,, \\ \Phi[C_\alpha^{(2)}\,.\, s/\sqrt{(s-1)}] &= (1-\alpha)^{1/s}\,. \end{aligned}$$

Table of values $C_\alpha^{(1)}$, $C_\alpha^{(2)}$ with $1 - \alpha = 95\%, 90\%$, and $s = 2\ (1)\ 10$, $n_1 = n_2 = \ldots = n_s$

	$1 - \alpha = 95\%$		$1 - \alpha = 90\%$	
s	$C_\alpha^{(1)}$	$C_\alpha^{(2)}$	$C_\alpha^{(1)}$	$C_\alpha^{(2)}$
2	0·9800	0·9776	0·8224	0·8156
3	1·0073	1·0040	0·8652	0·8561
4	0·9735	0·9694	0·8487	0·8404
5	0·9305	0·9280	0·8210	0·8140
6	0·8994	0·8956	0·7972	0·7860
7	0·8628	0·8593	0·7653	0·7561
8	0·8306	0·8289	0·7422	0·7339
9	0·8040	0·8024	0·7238	0·7159
10	0·7728	0·7713	0·6979	0·6913

2. The shift problem. Consider the combined problem of testing and classification where we test the hypothesis H_0 against the alternatives $K_1, \ldots, K_{N-1}$ defined as follows:

Let $X_1, \ldots, X_N$ be independent random observations, which have the absolutely continuous distribution functions with the densities $f_1(x), \ldots, f_N(x)$, respectively. Let

$$H_0 : f_1(x) = \ldots = f_N(x) = f(x)\,,$$

and let K_i, $i = 1, \ldots, N - 1$, be the alternatives under which the shift in a parameter involved in the distribution function of X's occurs at the i-th time point. That is, the alternatives are defined by

$$(25) \qquad K_i : \; f_1(x) = \ldots = f_i(x) = f(x)\,, \quad f_{i+1}(x) = \ldots = f_N(x) = f(x, \Delta)\,,$$

where $f(x, 0) = f(x)$.

Suppose that $f(x, \theta)$ satisfies the conditions (A_1), (A_2) of Theorem 1. The alternatives have the form of the regression alternatives considered in Theorem 1 with the regression constants defined by

$$(26) \quad C_{ij} = 0\,, \quad 1 \text{ if } j \leqq i\,, \quad j \geqq i + 1\,, \quad \text{respectively,} \quad i = 1, \ldots, N - 1\,.$$

Applying Theorem 1 we obtain the locally optimal decision rule at the significance level $1 - \alpha$. The decision rule is given by

$$(27) \quad \varphi_0(R) = 1,\ \xi_\alpha,\ 0 \quad \text{if} \quad A_N^{-1} \max_i T_i(R) < , = , > C_\alpha\,,$$

$$\varphi_j(R) = \xi_j(R),\ 0 \quad \text{if} \quad \max_i T_i(R) = , > T_j(R) \quad \text{for} \quad j = 1, \ldots, N - 1\,,$$

where $T_i(R)$ are defined by

$$T_i(R) = \sum_{k=i+1}^{N} a_N(R_k, f) \tag{28}$$

with $a_N(j, f)$ defined by (15) and where

$$A_N^2 = \sum_{j=1}^{N} a_N^2(j, f)\,. \tag{29}$$

In order to determine asymptotically C_α so that $\mathsf{E}_0\, \varphi_0(R) = 1 - \alpha$, let us prove the following lemma:

Consider the stochastic process

$$T_{N,t}(R) = \sum_{j=1}^{N} C_N(j, t)\, a_N(R_j, f) \tag{30}$$

where $0 \leqq t \leqq 1$ and

$$\begin{aligned} C_N(j, t) &= 0 && \text{if } j \leqq tN\,, \\ &= j - tN && \text{if } tN \leqq j < tN + 1\,, \\ &= 1 && \text{if } tN + 1 \leqq j\,. \end{aligned} \tag{31}$$

Then $T_{N,t}(R)$ determines a probability distribution on the space $(Z, \mathscr{C})$ where Z is the space of all continuous functions on $[0, 1]$ with the usual metric

$$\|z_1 - z_2\| = \max_{0 \leqq t \leqq 1} |z_1(t) - z_2(t)|\,, \quad z_1, z_2 \in Z$$

and $\mathscr{C}$ denotes the sigma-field of Borel subsets of Z, i.e. the smallest sigma-field containing all open subsets (see Sections V.3.1 and V.3.5 in [2]).

Lemma 2. *Assume that the function $\varphi(u, f)$ given by (15) is square integrable on $(0, 1)$, non-constant,*

$$\int_0^1 \varphi^2(u, f)\, \mathrm{d}u < \infty\,,$$

and that

$$N^{-1} \max_{1 \leqq j \leqq N} a_N^2(j, f) \to 0\,. \tag{32}$$

Then the stochastic process $A_N^{-1}\, T_{N,t}(R)$ converges, under H_0, in distribution in $(Z, \mathscr{C})$ to the Brownian bridge $z_\omega(t)$.

Proof. First note that the stochastic process may be written in the form

$$T_{N,t}(R) = \sum_{j=1}^{N} C_N(D_j, t)\, a_N(j, f) \tag{33}$$

where $D_1, \ldots, D_N$ are the antiranks of $X_1, \ldots, X_N$ defined as follows:

$$D_j = k \quad \text{if and only if} \quad R_k = j\,.$$

The vector of antiranks $(D_1, \ldots, D_N)$ has, under H_0, the same distribution as $(R_1, \ldots, R_N)$, i.e. $(D_1, \ldots, D_N)$ is also uniformly distributed. We observe that the stochastic process $A_N^{-1}\, T_{N,t}(R)$ takes on the form of the stochastic process given by (2) of Section V.3.5. in [2], where our $a_N(j, f)$, $C_N(D_j, t)$ play the role of Hájek - Šidák's $C_j - \bar{C}$, $a_N(R_j, t)$, respectively, since

$$\sum_{j=1}^{N} a_N(j, f) = N \int_0^1 \varphi(u, f)\,\mathrm{d}u = N \int_{-\infty}^{\infty} f(x, 0)\,\mathrm{d}x = 0\,,$$

by the conditions (A_1), (A_2).

Note that, by (32),

$$A_N^{-2} \max_j a_N^2(j, f) \sim \max_j a_N(j, f)/N \int_0^1 \varphi^2(u, f)\,\mathrm{d}u \to 0 \tag{34}$$

(see (18) of Theorem V.1.4.b in [2]), hence the conditions of Theorem V.3.5 in [2] are satisfied and it follows from the cited theorem that the stochastic process $T_{N,t}(R)/A_N$ converges in distribution in $(Z, \mathscr{C})$ to the Brownian bridge $z_\omega(t)$. Q.E.D.

Let us now return to the problem of asymptotic determination of C_α such that $\mathsf{E}_0\, \varphi_0(R) = 1 - \alpha$. Note that

$$T_{N,0}(R) = \sum_{j=1}^{N} a_N(R_j, f) = 0\,,$$

$$T_{N,t}(R) = \sum_{j=k+1}^{N} a_N(R_j, f) + (k - tN)\, a_N(R_k, f)$$

for all $t \in [(k-1)/N, k/N]$, $k = 1, \ldots, N$; therefore

$$\max_{0 \leqq t \leqq 1} T_{N,t}(R) = \max\{0, T_1(R), \ldots, T_{N-1}(R)\}\,. \tag{35}$$

It follows from (35) that if $C_\alpha \geqq 0$ then

$$\begin{aligned} \lim_{N\to\infty} P\{\max_{1 \leqq i \leqq N-1} T_i(R)/A_N < C_\alpha\} &= \\ = \lim_{N\to\infty} P\{\max_{0 \leqq t \leqq 1} T_{N,t}(R)/A_N < C_\alpha\} &= \\ = P\{\max_{0 \leqq t \leqq 1} z_\omega(t) < C_\alpha\} = 1 - \exp(-2C_\alpha^2)& \end{aligned} \tag{36}$$

(see, for example, Doob [1]). Consequently, defining

$$C_\alpha = [(-1/2)\ln \alpha]^{1/2} \quad \text{for} \quad 0 < \alpha < 1\,, \tag{37}$$

we obtain from (36)

$$\mathsf{E}_0\, \varphi_0(R) \approx P\{ \max_{1 \leqq i \leqq N-1} T_i(R)/A_N < C_\alpha\} \approx 1 - \alpha \tag{38}$$

for N sufficiently large.

Remark. The condition (32) is always satisfied whenever $\varphi(u, f)$ is bounded. The following lemma states that the condition (32) is fulfilled under rather smooth restriction placed on $\varphi(u, f)$ which generates the scores $a_N(j, f)$.

Lemma 3. *Assume that the function $\varphi(u)$, $0 < u < 1$, may be expressed as the finite sum of monotone, square integrable functions. Then the following relations hold for the scores $a_N^\varphi(j) = \mathsf{E}\, \varphi(U^{(j)})$ and the so-called approximate scores $a_N(j) = \varphi(j/(N+1))$:*

$$N^{-1} \max_{1 \leqq j \leqq N} [a_N^\varphi(j)]^2 \to 0\,, \tag{39}$$

$$N^{-1} \max_{1 \leqq j \leqq N} a_N^2(j) \quad \to 0\,. \tag{40}$$

Proof. First, let us prove the relation (40) for the function φ which is supposed to be monotone, square integrable. We may also assume naturally that φ is nondecreasing and that

$$\lim_{u \to 0} \varphi(u) < 0\,, \quad \lim_{u \to 1} \varphi(u) > 0\,.$$

In such a case we have

$$\varphi(1/(N+1)) < 0\,, \quad \varphi(N/(N+1) > 0$$

for N large enough and

$$N^{-1} \max_j (a_N^2(j)) = ((N+1)/N) \max \begin{Bmatrix} \varphi^2(1/(N+1))/(N+1), \\ \varphi^2(N/(N+1))/(N+1) \end{Bmatrix} \leqq \tag{41}$$

$$\leqq ((N+1)/N) \max \left\{ \int_0^{1/(N+1)} \varphi^2(u)\, \mathrm{d}u, \int_{N/(N+1)}^1 \varphi^2(u)\, \mathrm{d}u \right\} \to 0$$

since

$$\int_0^1 \varphi^2(u)\, \mathrm{d}u < \infty\,.$$

Consider the case where

$$\varphi(u) = \sum_{i=1}^r \varphi_i(u)$$

with $\varphi_i(u)$ monotone, square integrable. Then

$$N^{-1}\max_{1\leqq j\leqq N} a_N^2(j) = N^{-1}\max_{1\leqq j\leqq N}\Big[\sum_{i=1}^{r}\varphi_i(j/(N+1))\Big]^2 \leqq r\sum_{i=1}^{r} N^{-1}\max_{1\leqq j\leqq N}\varphi_i^2(j/(N+1)) \to 0$$

by the above result. Let $[y]$ denote the entier of real number y.

In order to prove (39), note that, according to Theorem V.1.4.b and Lemma V.1.6.a in [2],

$$\int_0^1 [a_N^\varphi(1+[uN]) - a_N(1+[uN])]^2\,\mathrm{d}u \leqq 2\int_0^1 [a_N^\varphi(1+[uN]) - \varphi(u)]^2\,\mathrm{d}u\; +$$
$$+\,2\int_0^1 [a_N(1+[uN]) - \varphi(u)]^2\,\mathrm{d}u \to 0\,.$$

It follows from this that

$$N^{-1}\max_j [a_N^\varphi(j)]^2 \leqq 2N^{-1}\max_j [a_N^\varphi(j - a_N(j)]^2 + 2N^{-1}\max_j a_N^2(j) \leqq$$
$$\leqq 2N^{-1}\sum_{j=1}^{N}[a_N^\varphi(j) - a_N(j)]^2 + 2N^{-1}\max_j a_N^2(j) =$$
$$= \int_0^1 [a_N^\varphi(1+[uN]) - a_N(1+[uN])]^2\,\mathrm{d}u + 2N^{-1}\max_{1\leqq j\leqq N} a_N^2(j) \to 0\,.$$

V. LOCALLY OPTIMAL RANK DECISION RULE FOR TESTING THE HYPOTHESIS WITH SYMMETRIC DENSITY

Let us consider the combined problem of testing and classification where the density under the hypothesis is symmetric.

Suppose that the independent observations $X_1, \ldots, X_N$ have the densities $f_1(x), \ldots, f_N(x)$ with respect to the Lebesgue measure. Let H_0^* and $K_1^*, \ldots, K_s^*$ be the hypothesis and the alternatives where

$$\begin{aligned}(42)\qquad & H_0^*: f_1(x) = \ldots = f_N(x) = f(x) \quad \text{with} \quad f(x) = f(|x|)\,,\\ & K_i^*: f_1(x) = f(x, \Delta C_{i1}), \ldots, f_N(x) = f(x, \Delta C_{iN})\end{aligned}$$

with $f(x, 0) = f(x)$ and $i = 1, \ldots, s$.

Denote the probability measures and the expectations under $H_0^*, K_1^*, \ldots, K_s^*$ by P_0^*, $P_{1,\Delta}^*, \ldots, P_{s,\Delta}^*$; E_0^*, $\mathsf{E}_{1,\Delta}^*, \ldots, \mathsf{E}_{s,\Delta}^*$ and denote the ranks of $|X_1|, \ldots, |X_N|$ by $R_1^+, \ldots, R_N^+$. Let $r = (r_1, \ldots, r_N)$, $v = (v_1, \ldots, v_N)$ be a realization of the vector of ranks $R^+ = (R_1^+, \ldots, R_N^+)$ and the vector $\operatorname{sign} X = (\operatorname{sign} x_1, \ldots, \operatorname{sign} x_N)$.

Theorem 2. *Suppose that the function $f(x, \theta)$ involved in K_i^* satisfies the following conditions:*

$$\text{(B 1)}\qquad \lim_{\theta\to 0}[f(x,\theta) - f(x)]/\theta = \dot{f}(x, 0) \quad a.e.\,.$$

Further, there exist two functions $t(x)$ defined only for $x \geqq 0$ and $u(x)$ ($u(x)$ is not necessarily defined for $x \neq \pm 1$) such that $\dot{f}(x, 0)$ may be expressed in the form: $\dot{f}(x, 0) = u(\text{sign } x)\, t(|x|)$. Besides it $f(x, \theta)$ is supposed to be absolutely continuous in θ on some interval containing the point zero.

$$\text{(B 2)} \qquad \lim_{\theta \to 0} \int_{-\infty}^{\infty} |\dot{f}(x, \theta)| \, \mathrm{d}x = \int_{-\infty}^{\infty} |\dot{f}(x, 0)| \, \mathrm{d}x < \infty$$

with $\dot{f}(x, \theta) = \partial f(x, \theta)/\partial \theta$.

Define a $(1 - \alpha)$-level rank decision rule by

(43)

$$\varphi_0(R^+, \text{sign } X) = 1, \ \xi_\alpha, \qquad 0 \quad \textit{if} \quad \max_{1 \leqq i \leqq s} T_i(R^+, \text{sign } X) < , = , > C_\alpha ,$$

$$\varphi_j(R^+, \text{sign } X) = \xi_j(R^+, \text{sign } X), \ 0 \quad \textit{if} \quad \max_{1 \leqq i \leqq s} T_i(R^+, \text{sign } X) = , > T_j(R^+, \text{sign } X)$$

where

$$(44) \qquad T_j(r, \text{sign } X) = \sum_{k=1}^{N} C_{jk} \, u(\text{sign } X_k) \, \mathsf{E}_0^*[t(|X|^{(r_k)})/f(|X|^{(r_k)})$$

for $j = 1, \ldots, s$.

Then there exists an $\varepsilon > 0$ such that

$$\sum_{i=1}^{s} \mathsf{E}_{i,\Delta}^* \, \varphi_i(R^+, \text{sign } X)$$

is maximum within the class of all $(1 - \alpha)$-level decision rules depending only on R^+ and sign X for all $0 < \Delta \leqq \varepsilon$.

Proof. Note that R^+ and sign X are mutually independent under H_0^* and $P_0^*\{R^+ = r, \text{sign } X = v\} = 1/(N! \, 2^N)$, hence, applying the generalized Neyman-Pearson's Lemma, we obtain the optimal decision rule within the class of all $(1 - \alpha)$-level decision rules depending only on R^+ and sign X for the combined problem of testing and classification of H_0^* against $K_1^*, \ldots, K_s^*$ with each Δ fixed. The decision rule is defined by:

$$(45) \qquad \varphi_0(r, v) = 1, \ \xi_\alpha, \quad 0 \quad \text{if} \quad \max_{1 \leqq i \leqq s} P_{i,\Delta}^*\{R^+ = r, \text{sign } X = v\} < , = , > C_\alpha ,$$

$$\varphi_j(r, v) = \xi_j(r, v), \ 0 \quad \text{if} \quad \max_{1 \leqq i \leqq s} P_{i,\Delta}^*\{R^+ = r, \text{sign } X = v\}$$

$$= , > P_{j,\Delta}^*\{R^+ = r, \text{sign } X = v\}$$

for $j = 1, \ldots, s$.

In the same way as in the proof of Theorem 2 in [7] we easily obtain

$$\lim_{\Delta \to 0} [2^N N! \, P_{i,\Delta}^*\{R^+ = r, \text{sign } X = v\} - 1]/\Delta = T_i(r, v) .$$

Consequently there exists an $\varepsilon > 0$ such that (45) is equivalent to (43) for all $0 < \Delta \leqq \varepsilon$. Q.E.D.

Corollary 1. *Put in* (42) $f(x, \theta) = f(x - \theta)$ *and suppose that*

(B 1*) $f(x)$ *is absolutely continuous and* $f(x) = f(|x|)$,

(B 2*) $\int_{-\infty}^{\infty} |f'(x)| \,\mathrm{d}x < \infty$ *where* $f'(x)$ *denotes the a.e. derivative of* $f(x)$.

Then the rank decision rule defined by (43) *with* (44) *replaced by*

$$(46) \qquad T_j(r, \operatorname{sign} X) = \sum_{k=1}^{N} C_{jk} \operatorname{sign} X_k \, \mathsf{E}_0^*[-f'(|X|^{(r_k)})/f(|X|^{(r_k)})]$$

is locally optimal within the class of all $(1 - \alpha)$*-level decision rules depending only on* R^+ *and sign* X *for the combined problem of testing and classification of* H_0^* *against* $K_1^*, \ldots, K_s^*$.

Proof. It is easy to verify that the conditions (B 1), (B 2) are satisfied with $u(x) = x$, $t(x) = f'(x)$ provided the conditions (B 1*), (B 2*) hold, since $f'(x) = (\operatorname{sign} x) \, . \, f'(|x|)$. Consequently, Corollary 1 follows from Theorem 2.

Corollary 2. *Put in* (42) $f(x, \theta) = \exp(-\theta) f(x \exp(-\theta))$ *and suppose that the condition* (B 1*) *and*

$$(\text{B } 2^{**}) \qquad \int_{-\infty}^{\infty} |x f'(x)| \,\mathrm{d}x < \infty$$

hold. Then the rank decision rule defined by (43) *with* (44) *replaced by*

$$(47) \qquad T_j(r, \operatorname{sign} X) = \sum_{k=1}^{N} C_{jk} \, \mathsf{E}_0^*[-1 - |X|^{(r_k)} f'(|X|^{(r_k)})/f(|X|^{(r_k)})]$$
$$= T_j^+(r), \quad \text{say, for } j = 1, \ldots, s,$$

is locally optimal in the sense described in Corollary 1.

Proof. Suppose that the conditions (B 1*) and (B 2**) hold, then (B 1) and (B 2) are satisfied with $u(x) = 1$, $t(x) = -1 - x f'(x)/f(x)$ since $\dot{f}(x, 0) = \dot{f}(|x|, 0) = -1 - x f'(|x|)/f(|x|)$. Consequently, the assertion of Corollary 2 follows from Theorem 2.

Remark. It is easy to verify that (46), (47) may be written in the form

$$(48) \qquad T_j(r, \operatorname{sign} X) = \sum_{k=1}^{N} C_{jk} \operatorname{sign} X_k a_{1N}^+(r_k, f),$$

$$(49) \qquad T_j^+(r) = \sum_{k=1}^{N} C_{jk} \, a_{2N}^+(r_k, f),$$

where

$$a_{1N}^{+}(k,f) = \mathsf{E}\,\varphi_1(\tfrac{1}{2}U^{(k)} + \tfrac{1}{2}, f) \quad \text{with} \quad \varphi_1(u,f) = -f'(F^{-1}(u))\,[f(F^{-1}(u))]^{-1}\,,$$

$$a_{2N}^{+}(k,f) = \mathsf{E}\,\varphi_2(\tfrac{1}{2}U^{(k)} + \tfrac{1}{2}, f) \quad \text{with} \quad \varphi_2(u,f) =$$

$$= -1 - F^{-1}(u)\,f'(F^{-1}(u))/f(F^{-1}(u))\,,$$

for $k = 1, \ldots, N$ and for $0 < u < 1$.

In order to determine asymptotically the constant C_α such that the rank decision rule given by (43) with $T_j(r, \operatorname{sign} X)$ defined by (48), (49) has the significance level $1 - \alpha$, let us note that if $K_1^*, \ldots, K_s^*$ are the s-sample slippage or shift alternatives, i.e. the regression constants C_{jk} take on the form (12) or (26), then the method of the asymptotic determination of C_α in Paragraph IV remains valid for the rank descision rule defined by (43) or (49) since R^+ is, under H_0^*, uniformly distributed.

As to the signed rank decision rule defined by (43) and (48), it is difficult to determine C_α for the regression constants of the form (26). On the contrary, it is easy to determine C_α for C_{jk} of the form (12).

Actually, suppose that C_{jk} take on the form (12), then (48) reduces to

$$T_j(r, \operatorname{sign} X) = \sum_{k \in I_j} \operatorname{sign} X_k \, a_{1N}^{+}(r_k, f)\,. \tag{50}$$

We have, for $i \neq j$

$$\operatorname{cov}\,(T_i(R^+, \operatorname{sign} X),\, T_j(R^+, \operatorname{sign} X)) =$$

$$= \sum_{m \in I_i} \sum_{k \in I_j} \mathsf{E}_0^*\{\operatorname{sign} X_m \operatorname{sign} X_k a_{1N}^{+}(R_m^+, f)\, a_{1N}^{+}(R_k^+, f)\} = 0$$

since $\operatorname{sign} X_m$, $\operatorname{sign} X_k$, R^+ are mutually independent for $m \neq k$ and $\mathsf{E}_0^* \operatorname{sign} X_k = 0$ because $f(x)$ is symmetric about zero. Furthermore it is easy to see that the joint distribution of the vector $(T_1(R^+, \operatorname{sign} X)/\sqrt{N}, \ldots, T_s(R^+, \operatorname{sign} X)/\sqrt{N})$ converges to the s-variate normal distribution with the mean zero and the covariance matrix $\{\sigma_{ij}\}$ where $\sigma_{ij} = 0$ for $i \neq j$ and $\sigma_{ii} = \lambda_i I(f)$ for all $i, j = 1, \ldots, s$.

Consequently,

$$\lim_{N \to \infty} P\{\max_{1 \leq j \leq s} T_j(R^+, \operatorname{sign} X)/\sqrt{N} < C_\alpha\} = \tag{51}$$

$$= P\{\max_{1 \leq j \leq s} Z_j' < C_\alpha\} = \prod_{j=1}^{s} P\{Z_j' < C_\alpha\} = \prod_{j=1}^{s} \Phi(C_\alpha/\sqrt{(\lambda_j I(f))})$$

where $Z_1', \ldots, Z_s'$ denote the normally distributed independent random variables with the mean zero and the variances $\lambda_1 I(f), \ldots, \lambda_s I(f)$, respectively. It follows from (51) that C_α may be defined so that

$$\mathsf{E}_0^* \varphi_0(R^+, \operatorname{sign} X) \approx P\{\max_{1 \leqq j \leqq s} T_j(R^+, \operatorname{sign} X) < C_\alpha\} \approx 1 - \alpha$$

for N large enough.

Remark. Omitting the classification of $K_1, \ldots, K_s$ and of $K_1^*, \ldots, K_s^*$ we obtain rank tests from the locally optimal decision rules given by (6), (7) and (43), (44). The rank tests are defined as follows:

$$(52) \qquad \psi(R) = 1 - \varphi_0(R) = 1,\ \xi_\alpha,\ 0 \quad \text{if} \quad \max_{1 \leqq i \leqq s} T_i(R) > , = , < C_\alpha ,$$

$$(53) \qquad \psi(R^+, \operatorname{sign} X) = 1 - \varphi_0(R^+, \operatorname{sign} X) = 1,\ \xi_\alpha,\ 0 \quad \text{if} \quad \max_{1 \leqq i \leqq s} T_i(R^+, \operatorname{sign} X) > , = , < C_\alpha .$$

We expect that the rank tests for testing H_0, H_0^* against $K_1, \ldots, K_s$ and $K_1^*, \ldots, K_s^*$, respectively, will have some good properties. However, the investigation of the properties of the above rank tests is out of the framework of this article.

Acknowledgment. The author wants to express his wholehearted gratitude to Dr. Z. Šidák, under whose direction this work, which is a part of author's thesis, has been written.

References

[1] *Doob, J. L.:* Heuristic approach to the Kolmogorov - Smirnov theorem. Annals of Math. Stat. 20 (1949), 393—403.

[2] *Hájek, J.* and *Šidák, Z.:* Theory of Rank Tests. Academia, Publishing House of the Czechoslovak Academy of Sciences. Praha 1967.

[3] *Hall, I. J.* and *Kudo, A.:* On slippage tests I. A generalization of Neyman-Pearson's Lemma. Annals of Math. Stat. 39 (1968), 1693—1699.

[4] *Hall, I. J., Kudo, A.* and *Yeh, N.C.:* On slippage tests II. Similar slippage tests. Annals of Math. Stat. 39 (1968), 2029—2037.

[5] *Karlin, S.* and *Truax, D. R.:* Slippage problems. Annals of Math. Stat. 31 (1960), 296—324.

[6] *Mosteller, F.:* A *k*-sample slippage test for an extreme population. Annals of Math. Stat. 19 (1948), 53—65.

[7] *Nguyen van Huu:* Rank test of hypothesis of randomness against a group of regression alternatives. Aplikace Matematiky 17 (1972), 422—447.

[8] *Paulson, E.:* An optimal solution to *k*-sample slippage problem for the normal distribution. Annals of Math. Stat. 23 (1952), 610—616.

[9] *Pfanzagl, J.:* Ein kombiniertes Test & Klassifikations-Problem. Metrika 2 (1959), 11—45.

[10] *Slepian, D.:* The one-sided barrier problem for Gaussian noise. Bell System Techn. J. 41 (1962), 463—501.

[11] *Truax, D. R.:* An optimum slippage test for the variance of k normal distributions. Annals of Math. Stat. 24 (1953) 669—674.

Souhrn

POŘADOVÁ ROZHODOVACÍ PROCEDURA PRO KOMBINOVANÝ PROBLÉM TESTOVÁNÍ A KLASIFIKACE

NGUYEN VAN HUU

Článek se týká problému testování hypotézy náhodnosti proti skupině regresních alternativ, kombinovaného s následujícím rozhodnutím, která z alternativ platí. Jsou navrženy pořadové rozhodovací procedury pro tento problém, které jsou lokálně optimální. V některých speciálních případech jsou též studována asymptotická rozložení testovacích statistik.

Author's address: Dr. *Nguyen van Huu*, CSc., Mathematical Faculty of Hanoi University, Hanoi, Vietnam.