Testy adaptacyjne dla problemu k prób

Grzegorz Wyłupek

Instytut Matematyczny
Polskiej Akademii Nauk
Oddział Wrocław

Problem testowania

Załóżmy, że $X_{l1}, \ldots, X_{ln_l}$, $l = 1, \ldots, k$, są k niezależnymi próbami losowymi, gdzie X_{lj} pochodzi z populacji o ciągłej dystrybuancie F_l. Rozważmy problem testowania

$$\mathcal{H} : F_1 = \ldots = F_k$$

przeciwko globalnej alternatywie

$$\mathcal{A} : F_i \neq F_j \text{ dla pewnych } 1 \leqslant i < j \leqslant k.$$

Notacja

Niech $N = \sum_{l=1}^{k} n_l$, $p_l = n_l/N$. Dla uproszczenia, będziemy identyfikować $X_{11}, \ldots, X_{1n_1}, \ldots, X_{k1}, \ldots, X_{kn_k}$ z $X_1, \ldots, X_{n_1}, \ldots, X_{n_1+1}, \ldots, X_{n_1+n_2}, \ldots, X_{n_1+\ldots+n_{k-1}+1}, \ldots, X_N$.
Niech $R_{n_1+\ldots+n_{l-1}+1}, \ldots, R_{n_1+\ldots+n_l}$ będzie wektorem rang $X_{l1}, \ldots, X_{ln_l}$ w połączonej próbie $X_1, \ldots, X_N$.
Oznaczmy przez Z_l zbiór indeksów odpowiadających l-tej próbie $Z_l = \{ n_1 + \ldots + n_{l-1} + 1, \ldots, n_1 + \ldots + n_l \}$ i połóżmy $Z = \bigcup_{l=1}^{k} Z_l = \{ 1, \ldots, N \}$.

Dla $l = 1, \ldots, k;\ i = 1, \ldots, N$ zdefiniujmy

$$c_{Ni}(l) = \sqrt{\frac{n_l(N-n_l)}{N}} \begin{cases} n_l^{-1} & \text{gdy} \quad i \in Z_l, \\ -(N-n_l)^{-1} & \text{gdy} \quad i \in Z - Z_l, \end{cases}$$

$$W_r^{[2]}(l) = \sum_{j=1}^{r} \Big\{ \sum_{i=1}^{N} c_{Ni}(l) b_j\Big(\frac{R_i - 0.5}{N}\Big)\Big\}^2 - \text{gładka 2-próbkowa statystyka,}$$

gdzie $b_0 \equiv 1, b_1, b_2, \ldots$ jest bazą wielomianów Legendre'a na $(0,1)$.

$$W_r = \sum_{l=1}^{k} (1-p_l) W_r^{[2]}(l) = \sum_{j=1}^{r} \Big[\sum_{l=1}^{k} (1-p_l) \Big\{ \sum_{i=1}^{N} c_{Ni}(l) b_j\Big(\frac{R_i - 0.5}{N}\Big)\Big\}^2 \Big],$$

W_r - gładka k-próbkowa statystyka.

W_1 - k-próbkowa statystyka Kruskala-Wallisa (1952);

W_2 - suma k-próbkowych statystyk Kruskala-Wallisa i Mooda;

W_4 - k-próbkowa statystyka Boosa (1986).

W_r jest kombinacją statystyk $W_r^{[2]}(l)$
dla podproblemów $(Z_l, Z - Z_l)$, $l = 1, \ldots, k$.

$$W_r^{[2]}(l) = \sum_{j=1}^{r} \Big\{ \sum_{i=1}^{N} c_{Ni}(l) b_j \Big(\frac{R_i - 0.5}{N} \Big) \Big\}^2.$$

Łączny rozkład w próbach odpowiadających $(Z_l, Z - Z_l)$ oraz $(Z_{l'}, Z - Z_{l'})$, $l \neq l'$, może być różny. Dlatego będziemy szukać optymalnego r, opierając się na danych, w każdym podproblemie dwóch prób $(Z_l, Z - Z_l)$ oddzielnie.

$$S(l) = \min\{r : \; W_r^{[2]}(l) - r \log N \geqslant W_j^{[2]}(l) - j \log N, \; 1 \leqslant r, j \leqslant d(N)\},$$

$d(N)$ - liczba modeli na liście.
$S(l)$ - uproszczona reguła Schwarza (1978) BIC;
patrz Janic-Wróblewska i Ledwina (2000).

Alternatywna reguła wyboru modelu:

$A(I)$ - uproszczona reguła Akaike (1973) AIC

$$A(I) = \min\{r : \ W_r^{[2]}(I) - 2r \geqslant W_j^{[2]}(I) - 2j, \ 1 \leqslant r, j \leqslant d(N)\}.$$

Wybór między regułami $S(I)$ i $A(I)$, lub równoważnie pomiędzy karami, powinien być oparty na danych - Inglot i Ledwina (2006).

Mamy $W_r^{[2]}(I) = \sum_{j=1}^{r} \{\hat{B}_j(I)\}^2$,

gdzie $\hat{B}_j(I) = \sum_{i=1}^{N} c_{Ni}(I) b_j([R_i - 0.5]/N)$.

Przy ustalonym I i pod warunkiem $\mathcal{H}$,

$\hat{B}_1(I), \hat{B}_2(I), \ldots, \hat{B}_{d(N)}(I)$ są asymptotycznie nieskorelowanymi zmiennymi losowymi o rozkładzie normalnym $N(0,1)$.

Nasza decyzja o karze będzie oparta
na zmiennych $\hat{B}_j(I)$, $j = 1, \ldots, d(N)$.
Dla każdego $d(N) \to \infty$ i $c \geqslant 0$, pod warunkiem $\mathcal{A}$, mamy

$$P\Big(\max_{1 \leqslant j \leqslant d(N)} |\hat{B}_j(I)| \leqslant \sqrt{c \log N}\Big) \to 0, \text{ gdy } N \to \infty.$$

Ponadto, jeżeli $\mathcal{H}$ jest prawdziwa, to dla każdego $\gamma \in (0, 1/9]$,
$d(N) = o([N/\log N]^\gamma)$, i $c > 2\gamma$ otrzymujemy

$$P\Big(\max_{1 \leqslant j \leqslant d(N)} |\hat{B}_j(I)| \leqslant \sqrt{c \log N}\Big) \to 1, \text{ gdy } N \to \infty.$$

Przy ustalonym $c > 2\gamma$, zdefiniujmy zmienną losową

$$I_N(c, l) = \mathbf{1}\Big(\max_{1 \leqslant j \leqslant d(N)} |\hat{B}_j(l)| \leqslant \sqrt{c \log N} \Big),$$

gdzie $\mathbf{1}(\mathcal{E})$ jest indykatorem zdarzenia $\mathcal{E}$.

Na mocy powyższego, $I_N(c, l)$ może służyć jako przełącznik pomiedzy karą Schwarza i karą Akaike.

Zdefiniujmy

$$\Pi_r(c, l) = (r \log N)[I_N(c, l)] + (2r)[1 - I_N(c, l)]$$

i nową regułę wyboru modelu

$$T(l) = \min\{r : \ W_r^{[2]}(l) - \Pi_r(c, l) \geqslant W_j^{[2]}(l) - \Pi_j(c, l), \ 1 \leqslant r, j \leqslant d(N)\}.$$

Nowy test adaptacyjny odrzuca $\mathcal{H}$ dla dużych wartości statystyki

$$W_T = \sum_{l=1}^{k} (1 - p_l) W^{[2]}_{T(l)}(l).$$

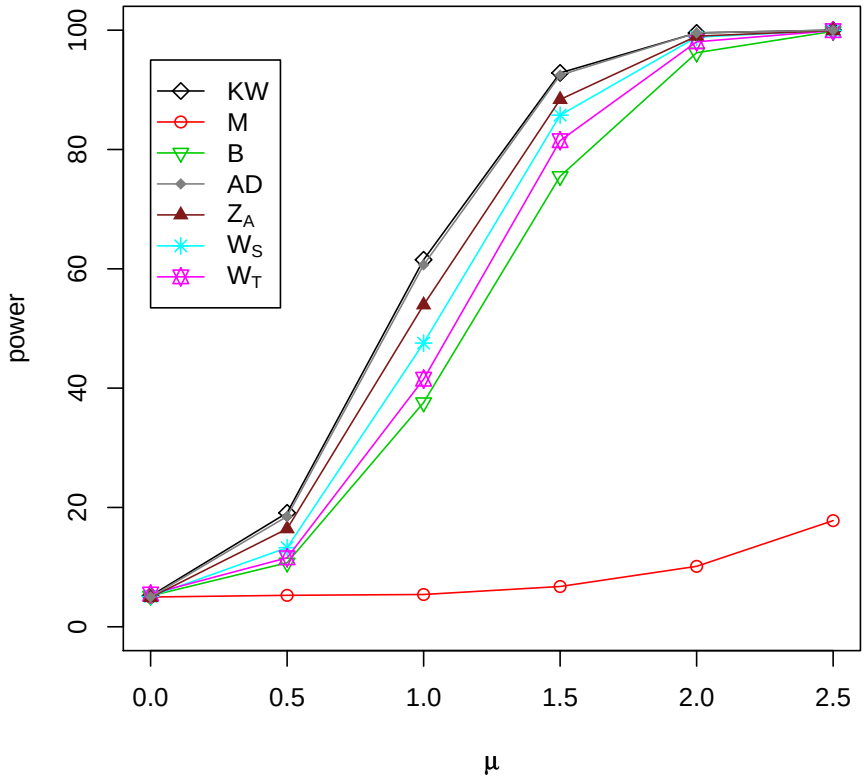
Powyższa formuła definiuje całą klasę statystyk indeksowanych parametrem $c \in [0, +\infty]$.
Wybór c w nowej karze $\Pi_r(c, l)$ i odpowiadającej statystyce W_T ma decydujący wpływ na zachowanie mocy testu adaptacyjnego. Pozwala to na konstrukcję testów typu omnibus jak i wyspecjalizowanych wariantów. Wyniki asymptotyczne zachodzą dla szerokiego zakresu c i $d(N)$.

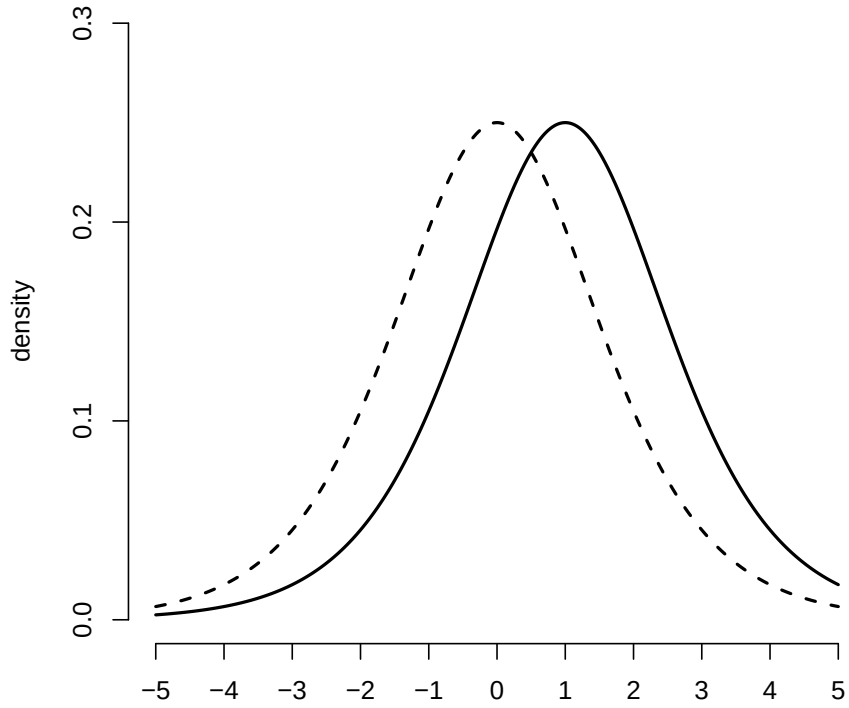
Niech χ^2_{k-1} będzie zmienną losową o rozkładzie chi-kwadrat z $k-1$ stopniami swobody. Ponadto niech $\xrightarrow{\mathcal{P}}$ i $\xrightarrow{\mathcal{D}}$ oznaczają odpowiednio zbieżność według prawdopodobieństwa i według rozkładu.

Twierdzenie 1. *Załóżmy, że $\mathcal{H}$ jest prawdziwa, $d(N) = o([N/\log N]^\gamma)$ i $0 < \gamma \leqslant 1/9$. Ponadto niech $c > 2\gamma$. Wtedy $T(l) \xrightarrow{\mathcal{P}} 1$, $l = 1, \ldots, k$, i $W_T \xrightarrow{\mathcal{D}} \chi^2_{k-1}$, gdy $N \to \infty$.*

Twierdzenie 2. *Załóżmy, że $d(N)$ jest jak w twierdzeniu 1 oraz $d(N) \to \infty$, gdy $N \to \infty$. Niech $c > 2\gamma$.*
Wtedy test adaptacyjny odrzucający $\mathcal{H}$ dla dużych wartości W_T jest zgodny dla każdej alternatywy z $\mathcal{A}$.

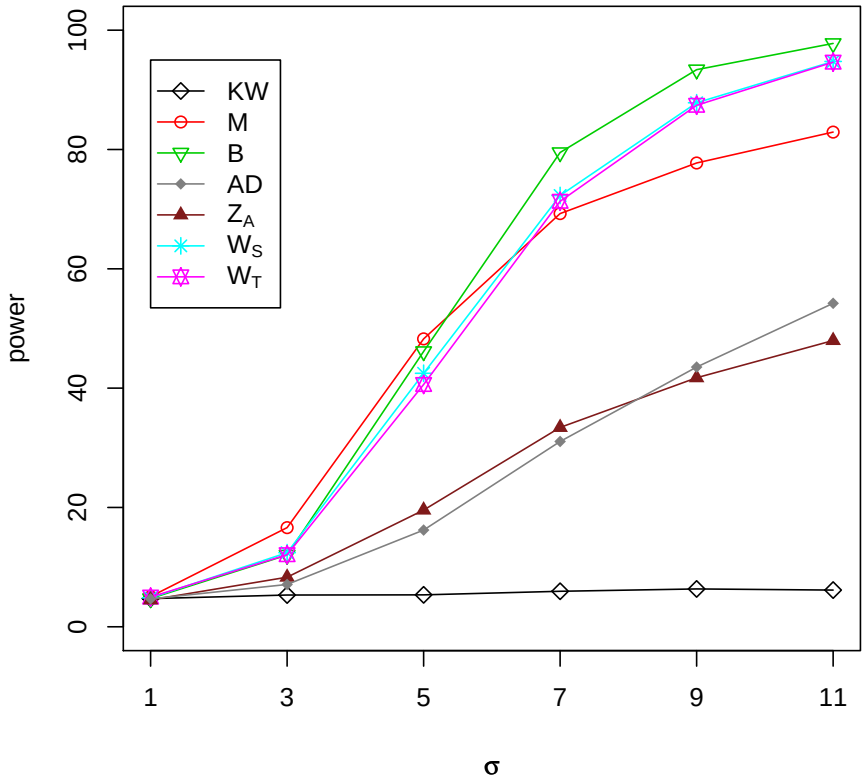
$\mathcal{H}$: *Logistic*(0)/*Logistic*(0)/*Logistic*(0) $\mathcal{A}(\mu)$: *Logistic*(0)/*Logistic*(μ)/*Logistic*(0)

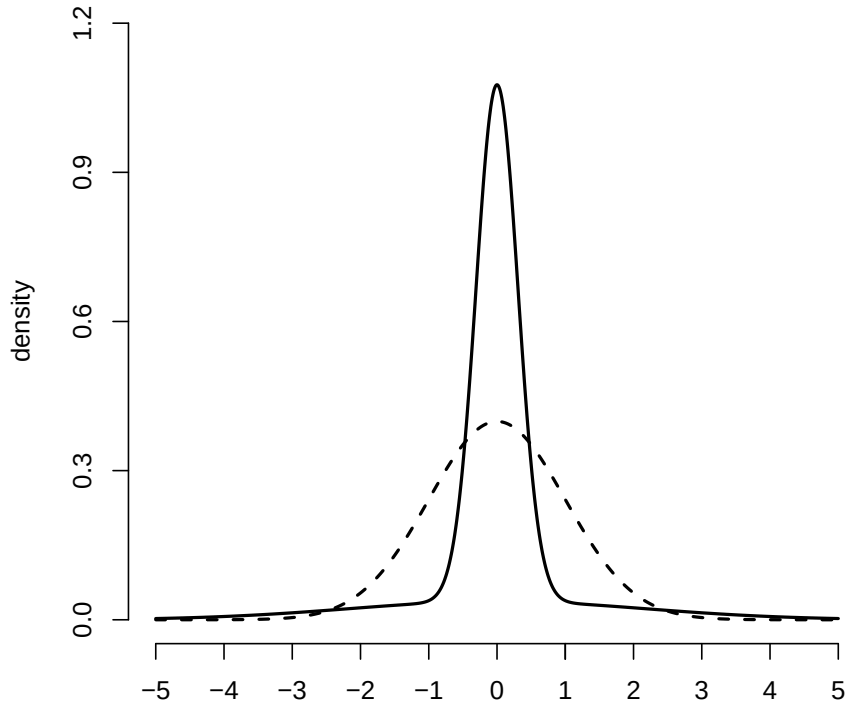




Rysunek 1. Lewy panel: porównanie mocy $n_1 = n_2 = n_3 = 30$, $\alpha = .05$, $c = 2.3$, $d(N) = 12$. Oparte na 10^4 powtórzeń MC. Moce przemnożone przez 100. Prawy panel: - - - Logistic(0), —— Logistic(1).

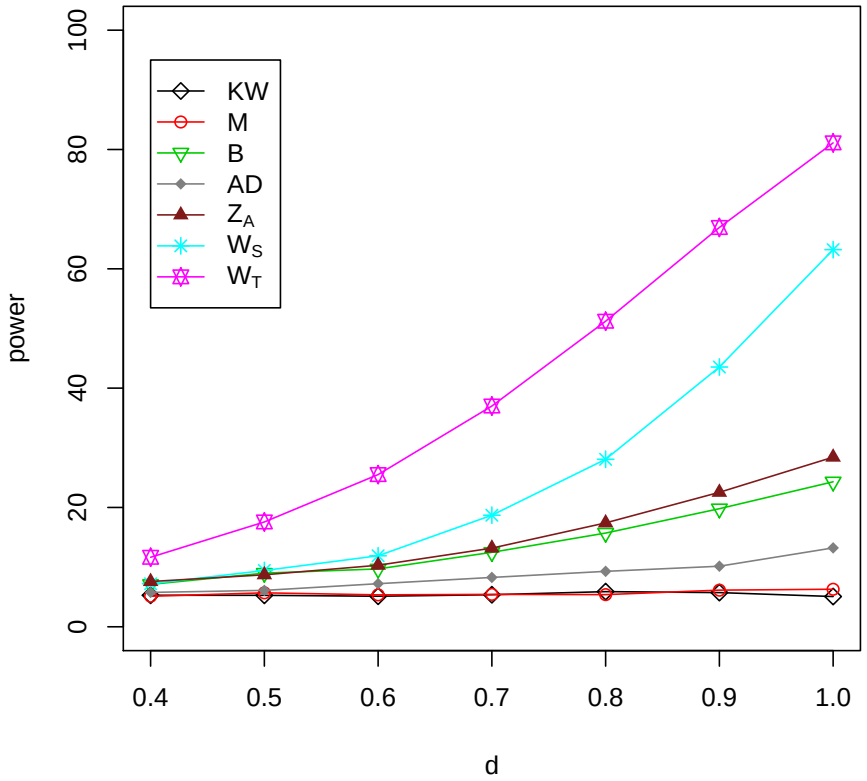
$\mathcal{H} : N(0,1)/N(0,1)/N(0,1)$ $\mathcal{A}(\sigma) : N(0,1)/SC(\sigma)/N(0,1)$

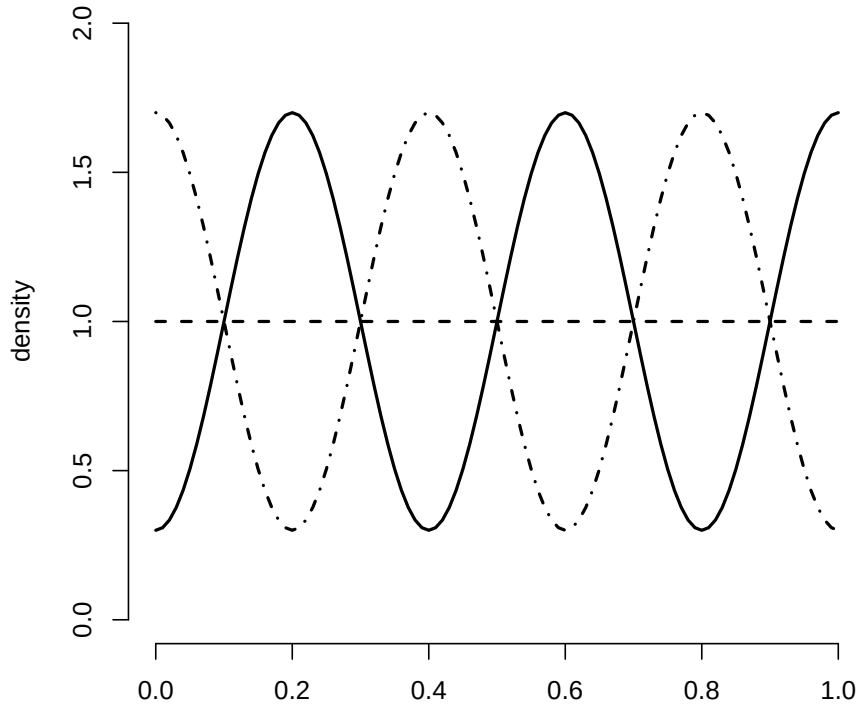




Rysunek 2. Lewy panel: porównanie mocy $n_1 = n_2 = n_3 = 30$, $\alpha = .05$, $c = 2.3$, $d(N) = 12$. Oparte na 10^4 powtórzeń MC. Moce przemnożone przez 100. Prawy panel: - - - $N(0,1)$, —— $SC(7)$.

$\mathcal{H} : U(0,1)/U(0,1)/U(0,1)$ $\mathcal{A}(d) : U(0,1)/C_5(d)/C_5(-d)$





Rysunek 3. Lewy panel: porównanie mocy $n_1 = n_2 = n_3 = 30$, $\alpha = .05$, $c = 2.3$, $d(N) = 12$. Oparte na 10^4 powtórzeń MC. Moce przemnożone przez 100. Prawy panel: - - - $U(0,1)$, —— $C_5(0.8)$, - · - · - $C_5(-0.8)$.