Asymptotyczna kontrola FDR dla zależnych testowań wielu hipotez statystycznych

Konrad Furmańczyk
Wydział Zastosowań Informatyki i Matematyki SGGW

PLAN REFERATU

- Sformułowanie zagadnienia testowania wielu hipotez;
- Uogólnione błędy I rodzaju dla testowania wielu hipotez: $FWER$, FDP, FDR;
- model hierarchiczny dla testowania wielu hipotez;
- uogólnienia modelu hierarchicznego dla zależnych testowań;

Klasyczne sformułowanie zagadnienia

Dane X są generowane z pewnego rozkładu prawdopodobieństwa $P \in \Omega$.

Interesują nas hipotezy dotyczące rozkładu P

$$H_i : P \in \omega_i \subset \Omega$$

dla $i = 1, 2, \ldots, m$, gdzie liczba testowanych hipotez m jest bardzo duża (tysiące, miliony hipotez w analizie danych mikromacierzowych lub analizie obrazów z funkcjonalnego rezonansu magnetycznego)

Problem: na podstawie statystyk testowych: $T_1, T_2, \ldots, T_m$ (lub odpowiednich p-wartości) podjąć decyzję, które z hipotez H_i są prawdziwe, a które fałszywe.

W ten sposób otrzymujemy 2^m możliwych decyzji do podjęcia.

Uogólnienia błędu I rodzaju

Niech R oznacza liczbę odrzuconych hipotez H_i w wyniku zastosowania pewnej procedury jednoczesnego testowania.

Przez V oznaczmy (nieobserwowaną zmienną) liczbę odrzuceń prawdziwych hipotez H_i,

- Family-wise error rate ($FWER$)

$$FWER = P(V \geq 1)$$

- Procedura testowania konrtoluje $FWER$ na poziomie α gdy
  $P(V \geq 1) \leq \alpha$ dla wszystkich $P \in \Omega$

- False discovery proportion (FDP)

$$FDP = \begin{cases} V/R & \text{gdy } R > 0 \\ 0 & \text{gdy } R = 0 \end{cases}.$$

- False discovery rate (FDR)

$$FDR = E\left(FDP\right).$$

Konstruuje się procedury, które kontrolują uogólniony błąd I rodzaju ($FWER$, FDR) , tzn. takie aby błęd ten nie przekraczał pewnego $\alpha \in (0;1)$.

W przypadku miary FDP szukamy takiej procedury aby

$P(FDP > \gamma) \leq \alpha$ dla wszystkich $P \in \Omega$ dla dowolnych α,$\gamma \in (0;1)$.

Niech P_i oznacza p-wartość dla hipotezy H_i dla $i = 1, ..., m$ oraz $P_{(1)} \leq P_{(2)} \leq ... \leq P_{(m)}$ będą uporządkowanymi p-wartościami. Dla ustalonego poziomu kontroli α oznaczmy

$$M := \max\left\{i \in \{0, 1, ..., m+1\} : P_{(i)} \leq \alpha i/m\right\},$$

gdzie $P_{(0)} := 0$, $P_{(m+1)} := 1$ oraz $T_{BH} := P_{(M)}$.

Procedura Benjamini i Hochberga (BH) odrzuca wszystkie hipotezy H_i dla których $P_i \leq T_{BH}$. Gdy $M = 0$ to wszystkie hipotezy są akceptowane.

- W przypadku niezależności statystyk testowych T_1, ... , T_m procedura BH kontroluje FDR na poziomie α.

  Dokładniej Benjamini i Hochberg (1995) otrzymali

$$FDR \leq m_0\alpha/m \leq \alpha,$$

  gdzie m_0 jest liczbą prawdziwych hipotez wśród hipotez H_i dla $i = 1, ..., m$.

- Benjamini i Yekutieli (2001) pokazali, że procedura BH kontroluję FDR bez założenia o niezależności statystyk testowych, jeśli zamiast α weźmiemy

$$\alpha / \sum_{j=1}^{m} \frac{1}{j}.$$

Model hierarchiczny

Efron, Tibshirani, Storey i Tusher (2001) wprowadzili następujący model testowania wielu hipotez

Niech $H_i = 0$ (lub 1) gdy H_i jest prawdziwa (fałszywa). Zakładamy, że (H_i) są niezależnymi zmiennymi losowymi o wspólnym rozkładzie Bernoulliego

$P(H_i = 0) = 1 - \pi$, $P(H_i = 1) = \pi$

oraz (P_i, H_i) są i.i.d. takie, że

$P(P_i \leq x \mid H_i = 0) = x$,

$P(P_i \leq x \mid H_i = 1) = F(x)$ dla $x \in [0, 1]$,

gdzie F wspólną dystrybuantą p-wartości P_i gdy hipoteza H_i jest fałszywa.

Wtedy dystrybuanta brzegowa G zmiennej losowej P_i jest postaci

$$G(t) = (1 - \pi)t + \pi F(t) \text{ dla } t \in [0, 1].$$

False nondiscovery proportion określamy jako

$$FNP = \begin{cases} N/(m - R) & \text{gdy } R < m \\ 0 & \text{gdy } R = m \end{cases},$$

gdzie N - liczba nieodrzuconych fałszywych hipotez H_i.

Określmy następujące procesy stochastyczne (odpowiedniki FDR i FNR)

$$\Gamma_m(t) = \frac{\sum_{i=1}^{m} \mathbf{1}\left\{P_i \leq t\right\}(1-H_i)}{\sum_{i=1}^{m} \mathbf{1}\left\{P_i \leq t\right\} + \mathbf{1}\left\{P_{(1)} > t\right\}},$$

$$\Xi_m(t) = \frac{\sum_{i=1}^{m} \mathbf{1}\left\{P_i > t\right\} H_i}{\sum_{i=1}^{m} \mathbf{1}\left\{P_i > t\right\} + \mathbf{1}\left\{P_{(m)} \leq t\right\}}$$

dla $t \in [0,1]$.

Storey (2002) pokazał, że przy założeniu modelu hierarchicznego mamy dla $t > 0$,

$$E\left(\Gamma_m(t)\right) = Q(t)\left(1 - (1 - G(t))^m\right),$$

$$E\left(\Xi_m(t)\right) = \tilde{Q}(t)\left(1 - G(t)^m\right),$$

gdzie

$$Q(t) = (1 - \pi)\frac{t}{G(t)},$$

$$\tilde{Q}(t) = \pi\frac{1 - F(t)}{1 - G(t)}.$$

Zakładamy, że π jest znane.

Oznaczmy $T_{PI} := \sup\{0 \leq t \leq 1 : Q_m(t) \leq \alpha\}$, gdzie

$$Q_m(t) = (1-\pi)\frac{t}{G_m(t)},$$

$$G_m(t) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\{P_i \leq t\}.$$

Zauważmy, że

$$FDR(t) := E\left(\Gamma_m(t)\right) \approx Q_m(t),$$

więc

$$FDR(T_{PI}) \approx \alpha.$$

Dokładniej Genovese i Wasserman (2004) pokazali, że w modelu hierarchicznym

$$E\left(\Gamma_m(T_{PI})\right) = \alpha + o(1) \text{ dla } m \to \infty.$$

Zakładamy, że π jest nieznane.

Genovese i Wasserman (2004) pokazali, że następująca procedura testowania kontroluje asymptotycznie FDR.

Próg odrzucenia hipotez H_i określmy jako

$$\hat{T} = \sup\left\{0 \leq t \leq 1 : \hat{Q}_m(t) \leq \alpha\right\},$$

gdzie

$$\hat{Q}_m(t) = (1 - \hat{\pi})\frac{t}{G_m(t)}.$$

Zakładamy dodatkowo, że G jest wklęsła oraz estymator $\hat{\pi}$ jest taki, że

$$\hat{\pi} \rightarrow^P \pi_0 < \pi.$$

Wtedy

$$E\left(\Gamma_m(\hat{T})\right) = \alpha + o(1) \text{ dla } m \rightarrow \infty.$$

Farcomeni (2007) rozważył uogólnienia modelu hierarchicznego. Dopuścił pewne modele zależności ciągu p-wartości (P_i).

Np. ciąg (P_i) jest stacjonarnym ciągiem α-mieszającym takim, że

$$\sum_{k=1}^{\infty} \alpha(k) < \infty.$$

Wtedy

$$E\left(\Gamma_m(T_{PI})\right) = \alpha + o(1) \text{ dla } m \to \infty.$$

Oznaczmy

$$t_0 = Q^{-1}(\alpha) := \sup\{0 \leq t \leq 1 : Q(t) \leq \alpha\},$$

gdzie

$$Q(t) = (1-\pi)\frac{t}{G(t)}.$$

Niech Q_i dla $i = 1, ..., m_0$ odpowiadają p-wartościom dla prawdziwych hipotez H_i oraz R_i dla $i = 1, ..., m - m_0$ odpowiadają p-wartościom dla fałszywych hipotez.

Załóżmy, że dla $m \to \infty$

a)

$$\frac{m_0}{m} \to^P 1 - \pi$$

b1)

$$\frac{1}{m_0} \sum_{i=1}^{m_0} \mathbf{1}\left\{Q_i \leq t_0\right\} \to^P t_0$$

b2)

$$\frac{1}{m - m_0} \sum_{i=1}^{m-m_0} \mathbf{1}\left\{R_i \leq t_0\right\} \to^P F(t_0)$$

c) G jest ciągła.

Można pokazać (Furmańczyk(2009)), że przy warunkach **a)**, **b1)**, **b2)**, **c)** (gdy π jest znane)

$$E\left(\Gamma_m(T_{PI})\right) = \alpha + o(1) \text{ dla } m \to \infty$$

$$E\left(\Xi_m(T_{PI})\right) = \pi\frac{1-F(t_0)}{1-G(t_0)} + o(1) \text{ dla } m \to \infty.$$

Gdy π jest nieznane oraz $\hat{\pi} \to^P \pi$ to

$$E\left(\Gamma_m(\hat{T})\right) = \alpha + o(1) \text{ dla } m \to \infty$$

$$E\left(\Xi_m(\hat{T})\right) = \pi\frac{1-F(t_0)}{1-G(t_0)} + o(1) \text{ dla } m \to \infty.$$

Warunek **b1)** otrzymujemy np. gdy istnieje taki ciąg $k_{m_0} \to \infty$, $\frac{k_{m_0}}{m_0} \to^P 0$ gdy $m_0 \to^P \infty$ oraz

$$\sup_{|i-j|\geq k_{m_0}} |P(Q_i \leq t_0, Q_j \leq t_0) - P(Q_i \leq t_0)P(Q_j \leq t_0)| \to 0.$$

Podobnie otrzymujemy **b2**).

gdy istnieje taki ciąg $k_m \to \infty$, $\frac{k_m}{m} \to 0$ gdy $m \to \infty$ oraz

$$\sup_{|i-j|\geq k_m} |P(R_i \leq t_0, R_j \leq t_0) - P(R_i \leq t_0)P(R_j \leq t_0)| \to 0.$$

Estymacja π

Storey (2002) zaproponował estymator

$$\hat{\pi} = \left(\frac{G_m(s) - s}{1 - s} \right)_+$$

dla pewnego $s \in (0, 1)$. Jeśli $G(s) > s$, to

$$\hat{\pi} \to^P \frac{G(s) - s}{1 - s}$$

oraz

$$\sqrt{m} \left(\hat{\pi} - \frac{G(s) - s}{1 - s} \right) \implies \mathcal{N} \left(0, \frac{G(s)(1 - G(s))}{(1 - s)^2} \right).$$

Niech

$$P_j = \sum_{j=0}^{\infty} b_j \xi_{i-j},$$

gdzie $\left(\xi_j\right)$ i.i.d. Niech

$$B := \sum_{j=0}^{\infty} j \left|b_j\right| < \infty$$

oraz gęstość P_j jest ograniczona przez K. Oznaczmy

$$\mu := (1 + 4KB \left\|\xi_0\right\|_\infty) \sqrt{\frac{2}{m} \ln\left(\frac{2}{\alpha}\right)}.$$

Połóżmy

$$\tilde{\pi} = \max_t \frac{G_m(t) - t - \mu}{1 - t}$$

Wtedy (Furmańczyk (2009))

$$P(\tilde{\pi} \leq \pi) \geq 1 - \alpha$$

a stąd

$$P(1 - \pi \leq 1 - \tilde{\pi}) \geq 1 - \alpha.$$

Wyniki Wu (2009)

Wu (2009) wprowadził następujący wariant modelu hierarchicznego

(H_i) jest 0-1 procesem stacjonarnym oraz p-wartości (P_i) są warunkowo niezależne pod warunkiem (H_i).

Wtedy też

$$P(P_i \leq x \mid H_i = 0) = x,$$
$$P(P_i \leq x \mid H_i = 1) = F(x) \text{ dla } x \in [0,1].$$

Rozważmy następujące warunki regularności na proces (H_i)

$$\left\| \sum_{i=1}^{m} H_i - m\pi \right\|_2 = \mathcal{O}(\sqrt{m})$$

$$\frac{1}{\sqrt{m}} \left(\sum_{i=1}^{m} H_i - m\pi \right) \implies \mathcal{N}(0, \sigma^2)$$

dla pewnego $\sigma^2 < \infty$.

Oznaczmy

$$\upsilon_{BH} := \sup\left\{0 \leq t \leq 1 : \frac{t}{G_m(t)} \leq \alpha\right\}$$

oraz

$$\upsilon_0 := \sup\left\{0 \leq t \leq 1 : \frac{t}{G(t)} \leq \alpha\right\}.$$

Wtedy przy pewnych warunkach regularności

$$\sqrt{m}(\Gamma_m(\upsilon_{BH}) - \alpha(1-\pi)) \implies \mathcal{N}(0, \sigma_1^2)$$

$$\sqrt{m}(\Xi_m(\upsilon_{BH}) - \pi\frac{1-F(\upsilon_0)}{1-G(\upsilon_0)}) \implies \mathcal{N}(0, \sigma_2^2)$$

dla pewnych $\sigma_1^2 < \infty$, $\sigma_2^2 < \infty$. oraz

$$\Gamma_m(\hat{T}) = \alpha + \alpha\left(\frac{\hat{\pi} - \pi}{1-\pi}\right) + \mathcal{O}_P(m^{-1/2}).$$

Przykłady procesów (H_i)

a) ucięty proces liniowy

$H_i = \mathbf{1}\{X_i \leq t\}$ dla ustalonego $t \in \mathbb{R}$

gdzie

$$X_i = \sum_{j=-\infty}^{\infty} b_j \xi_{i-j}$$

oraz (ξ_j) i.i.d. o ograniczonej gęstości, $E\left|\xi_j\right|^d < \infty$ dla pewnego $d > 0$ oraz

$$\sum_{j=-\infty}^{\infty} |b_j|^{\min(1,d)/2} < \infty.$$

b) model Isinga (pole losowe w $A \subset \mathbb{Z}^2$)

niech

$$L_s = 2H_s - 1$$

$L_s = 1$ (lub -1) gdy $H_s = 0$ (lub 1) oraz sąsiedztwo punktu $s = (j, k)$

$$\mathcal{N}_s = \{(j', k') : |j - j'| + |k - k'| = 1\}$$

$L_A = \{L_a, a \in A\}$, wtedy rozkład H_i spełnia warunki regularności

$$P(L_s = l_s | L_{\mathbb{Z}^2 \setminus s} = l_{\mathbb{Z}^2 \setminus s}) = P(L_s = l_s | L_{\mathcal{N}_s} = l_{\mathcal{N}_s})$$
$$= \frac{\exp(\beta l_s \sum_{t \in \mathcal{N}_s} l_t)}{\exp(\beta \sum_{t \in \mathcal{N}_s} l_t) + \exp(-\beta \sum_{t \in \mathcal{N}_s} l_t)}$$

dla $0 \leq \beta \leq \log(1 + \sqrt{2})/2$.

Literatura

[1] Y. Benjamini and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing,* J. Roy. Statist. Soc. Ser. B 57 (1995), 289-300.

[2] Y. Benjamini and D. Yekutieli, *The control of the false discovery rate in multiple testing under dependency,* Ann. Stat. 29 (2001), 1165-1188.

[3] B. Efron, R. Tibshirani, J. Storey and V. Tusher, *Empirical Bayes analysis of microarray experiment*, Journ. Amer. Stat. Assoc. 96 (2001), 1151-1160.

[4] A. Farcomeni, *Some results on the control of the false discovery rate under dependence,* Scan. Journ. Stat. 34 (2007), 275-297.

[5] K. Furmańczyk, *On the control of the false discovery rate under dependence, preprint* (2009).

[6] C. Genovese and L. Wasserman, *A stochastic process approach to false discovery control,* Ann. Statist. 32 (2004), 1035-1061.

[7] J. Storey, *A direct approach to false discovery rates,* J. R. Stat. Soc. Ser. B Stat. Meth. 64 (2002), 479-498.

[8] W. B. Wu, *On false discovery control under dependence,* Ann. Stat. 36 (2009), 364-380.