Teoria analizy dużych zbiorów - Lista III

Test Higher Criticism i detekcja rzadkich sygnałów

MM

11 maja 2017

Spis treści

1 Wstęp

W nimniejszym raporcie umieszczone zostały rozwiązania trzeciej listy zadań z przedmiotu **Teoria analizy dużych zbiorów** prowadzonego przez Panią Profesor Małgorzatę Bogdan we współpracy z Panem Michałem Kosem. Jest to kontynuacja zagadnień poruszanych na poprzednich listach, tym razem głównym tematem będzie statystyka **Higher Criticism** oraz tzw. rzadkie mieszaniny.

Poniżej wprowadzimy podstawowe pojęcia używane w późniejszych rozważaniach.

1.1 Testy bazujące na dystrybuancie empirycznej

Dystrybuanta empiryczna $p_1, ..., p_n$ dana jest wzorem

$$\hat{F}_n(t) = \frac{1}{n}\#\{i : p_i \leq n\}$$

1.1.1 Test Kolmogorova-Smirnova (K-S)

Statystyka testu K-S zadana jest wzorem

$$KS = \sup_t |\hat{F}_n(t) - t|$$

1.1.2 Test Andersona-Darlinga (A-D)

Statystyka testu A-D dana jest poprzez

$$A^2 = n\int_0^1 \frac{\left(\hat{F}_n(t) - t\right)^2}{t(1-t)}dt$$

Jest to specjalna wersja statystyki Cramera-von Misesa z funkcją wagi $\omega(t) = \left[t(1-t)\right]^{-1}$. Użyteczna jest następująca zależność

$$A^2 = -n - \sum_{i=1}^{n} \frac{2i-1}{n}\Big[\log(p_{(i)} + \log(1 - p_{(n+1-i)})\Big]$$

1.1.3 Test Tukey'a - Tukey's Second-Level Significance Testing (H-C)

Statystyka testowa dana jest poprzez

$$HC_n(t) = \frac{\hat{F}_n(t) - t}{\sqrt{t(1-t)/n}}$$

Obliczamy jej zgeneralizowaną wersję (wprowadzoną przez Donoho i Jin[2004]) zadaną wzorem

$$HC_n^* = \max_{0 \leq t \leq \alpha_0} \frac{\hat{F}_n(t) - t}{\sqrt{t(1-t)/n}}$$

W pożniejszych rozważaniach przyjmiemy $\alpha_0 = 0.5$.

Oczywiście wszystkie powyższe testy odrzucają hipotezy zerowe dla dużych wartości statystyk.

1.2 Rzadkie mieszaniny

Wprowadzimy teraz zadanie badane na kolejnych stronach. Rozważmy problem testowania:

$$H_{0,i} : X_i \sim \mathcal{N}(0,1)$$

$$H_{1,i} : X_i \sim \mathcal{N}(\mu_i, 1) \qquad \mu_i > 0$$

W powyższym zagadnieniu dopuszczamy pewną liczbę zagadnień, nieznaną, które pochodzą z hipotezy H_1. Zamiast tego zakładamy, że próby pochodzą z mieszaniny rozkładów $\mathcal{N}(0,1)$ oraz $\mathcal{N}(\mu,1)$. Na podstawie powyższego budujemy nowe doświadczenie:

$$H_0 : X_i \sim \mathcal{N}(0,1)$$

$$H_1 : X_i \sim (1-\epsilon)\mathcal{N}(0,1) + \epsilon\mathcal{N}(\mu,1) \qquad \mu > 0$$

W takim doświadczeniu test ilorazu największej wiarygodności jest postaci:

$$L = \prod_{i=1}^{n} \Big[(1-\epsilon) + \epsilon \exp(\mu X_i + \mu^2/2)\Big]$$

Ustalmy zależność ϵ oraz μ od n

$$\epsilon_n = n^{-\beta} \qquad \frac{1}{2} \leq \beta \leq 1$$

$$\mu_n = \sqrt{2r \log n} \qquad 0 < r < 1$$

Widać, że zagadnienia badane na poprzednich listach to sytuacje graniczne, tzn. gdy $\beta = 1/2$ lub $\beta = 1$ i $r = 1$, odpowiadające odpowiednio problemowi małych efektów i problemowi igły w stogu siana.

Jeśli nie jest napisane inaczej wszystkie symulacje zostały przeprowadzone dla 500 replikacji

2 Zadanie 1

Zadaniem jest estymacja wartości krytycznej testu H-C dla $p \in 5000, 50000, 500000$ na poziomie istotności $\alpha = 0.05$. Estymacja przy użyciu zadanej wyżej postaci statystyki HC_n^* jest skomplikowana numerycznie. Zauważmy, że możemy zamiast szukać po zbiorze wszystkich α możemy przebiegać zbiór p-wartości. Wtedy statystyka będzie miała postać

$$HC_n^* = \max_{p_i \leq \alpha_0} \frac{\hat{F}_n(p_i) - p_i}{\sqrt{p_i(1-p_i)/n}}$$

P-wartości do symulacji wyliczamy z wzoru

$$p_i = 2\Phi(-|X_i|)$$

Gdzie X_i pochodzą z odpowiedniego rozkładu normalnego. Zakładamy, że przy H_0 p-wartości powinny miec rozkład jednostajny, dlatego porównamy wyniki naszych symulacji z symulacjami, gdzie p-warotści będą właśnie z takiego rozkładu pochodziły.

	$2\Phi(-\|X_i\|)$	$\mathcal{U}[0,1]$
5000	4.87	4.96
50000	4.74	4.81
500000	4.61	4.65

Tablica 1: Moce testów dla każdego z przypadków

Okazało się, że dopiero dla 10 000 symulacji wyniki zaczęły się do siebie zbliżać, tutaj przedstawionmo wynika dla 100 000 symulacji. Dla 1000 symulacji różnice pomiędzy wartościami krytycznymi osiągały nawet 0.5. Prawdziwa wartość krytyczna jest trudna do estymowania, ale możemy ustalić, ze z pewnym prawdopodobieństwem leży w przedziale $(4.7 - \theta, 4.7 + \theta)$.

3 Zadanie 2

W kolejnym zadanie należy porównać moce testów H-C, Bonferonniego, χ^2, K-S oraz A-D przy testowaniu problemów z zadania czwartego z listy pierwszej, ponadto nalezy dołożyć jedno zagadnienie testowania. Nie bedziemy przytaczać statystyk testowych, zostało to zrobione we wstępie. Próbkę symulujemy z rozkładu normalnego o wariancji równej jeden, a $p = 5000$. Liczba sumulacji to także 5000.

Dla przypomniania, będziemy rozpatrywać następujące problemy testowania (hipoteza zerowa jest wspólna dla każdego z nich):

$$H_0 : \mu_1 = ... = \mu_{5000} = 0$$

- A. $H_1 : \mu_1 = 1.2\sqrt{2 \log p}, \mu_2 = ... = \mu_{5000} = 0$
- B. $H_1 : \mu_1 = ... = \mu_{1000} = 0.15\sqrt{2 \log p}, \mu_{1001} = ... = \mu_{5000} = 0$
- C. $H_1 : \mu_1 = ... = \mu_{100} = 2, \mu_{101} = ... = \mu_{5000} = 0$

Jak to w takich przypadkach, zasymulowaliśmy próbki z rozkładu przy hipetezie alternatywnej i sprawdziliśmy jak często została odrzucona hipoteza zerowa, oto wyniki:

	A	B	C
HC	0.71	0.42	0.93
Bonf	0.71	0.12	0.55
KS	0.05	1.00	0.49
AD	0.05	1.00	0.86
χ^2	0.08	0.98	0.98

Tablica 2: Estymowane moce testów

Widzimy, że każdy z testów nadaje się do innego typu testowania. W przypadku igły w stogu siana najlepiej najmocniejszym testem jest korekta Bonferroniego. Nieźle radzi sobie także H-C. Pozostałe testy mają bardzo małą moc i nie nadają się do tego typu zagadnień, bardzo często popełnialibyśmy błąd drugiego rodzaju. Do testowanie zagadnień wielu małych efektów nadaja się wszystkie trzy wymienone wyżej testy, bez Korekty Bonferonniego oraz H-C. Ponadto, dla zagadnienia trzeciego, dobrze spisują się testy H-C, χ^2, A-D, przy czym ten ostatni delikatnie odstaje od pozostałych. H-C wydaje się być najbardziej "uniwersalny" ze wszystkich pięciu.

4 Zadanie 3

W pierwszej częsci tego zadania należy wysymulować wartości krytyczne dla testu N-P w testowaniu w mieszaninach rzadkich z różnymi parametrami, w drugiej należy porównać jego moc do mocy testów z poprzedniego zadania. Będziemy badać moce testów dla wszystkich możliwych kombinacji następujących zbiorów: $p \in \{5000, 5000, 50000\}$, $\beta \in \{0.6, 0.8\}$, $r \in \{0.1, 0.2, 0.3, 0.4\}$. Sposób parametryzacji wprowadziliśmy we wstępie.

4.1 Część A

Wyniki symulacji:

	0.1	0.2	0.3	0.4
5000	2.82	3.72	2.39	0.14
50000	2.76	3.54	0.19	0.00
5e+05	3.15	1.57	0.00	0.00

(a) $\beta = 0.6$

	0.1	0.2	0.3	0.4
5000	1.27	1.77	2.65	2.42
50000	1.23	1.59	2.34	2.28
5e+05	1.13	1.48	2.51	3.61

(b) $\beta = 0.8$

Tablica 3: Wartości krytyczne testu N-P

Wraz ze wzrostem parametrów r oraz β rosnie wartość krytyczna testu, co zdaje się byc intuicyjne.

4.2 Część B

Przy użyciu powyższych wartości porównamy moc testu N-P oraz testów z poprzedniego zadania

	β	p	r	NP	HC	Bonf	KS	AD	χ^2
1	0.60	5000.00	0.10	0.24	0.08	0.08	0.06	0.07	0.11
2	0.60	5000.00	0.20	0.60	0.20	0.19	0.07	0.09	0.28
3	0.60	5000.00	0.30	0.88	0.57	0.46	0.10	0.13	0.43
4	0.60	5000.00	0.40	0.98	0.84	0.69	0.11	0.17	0.63
5	0.60	50000.00	0.10	0.25	0.07	0.06	0.05	0.06	0.11
6	0.60	50000.00	0.20	0.67	0.25	0.19	0.05	0.06	0.25
7	0.60	50000.00	0.30	1.00	0.74	0.54	0.08	0.11	0.48
8	0.60	50000.00	0.40	1.00	0.99	0.86	0.08	0.12	0.66
9	0.60	500000.00	0.10	0.26	0.07	0.07	0.04	0.07	0.15
10	0.60	500000.00	0.20	0.92	0.38	0.27	0.07	0.09	0.25
11	0.60	500000.00	0.30	1.00	0.96	0.70	0.07	0.10	0.41
12	0.60	500000.00	0.40	1.00	1.00	0.95	0.05	0.07	0.64
13	0.80	5000.00	0.10	0.08	0.04	0.05	0.05	0.05	0.07
14	0.80	5000.00	0.20	0.10	0.07	0.07	0.05	0.05	0.08
15	0.80	5000.00	0.30	0.17	0.14	0.12	0.06	0.06	0.09
16	0.80	5000.00	0.40	0.40	0.26	0.25	0.05	0.05	0.10
17	0.80	50000.00	0.10	0.03	0.06	0.06	0.06	0.05	0.06
18	0.80	50000.00	0.20	0.13	0.08	0.08	0.05	0.05	0.07
19	0.80	50000.00	0.30	0.25	0.18	0.15	0.04	0.05	0.09
20	0.80	50000.00	0.40	0.43	0.28	0.27	0.05	0.04	0.07
21	0.80	500000.00	0.10	0.03	0.06	0.06	0.05	0.05	0.07
22	0.80	500000.00	0.20	0.10	0.10	0.08	0.07	0.06	0.06
23	0.80	500000.00	0.30	0.21	0.16	0.13	0.06	0.06	0.06
24	0.80	500000.00	0.40	0.43	0.32	0.27	0.06	0.04	0.05

Tablica 4: Moce testów dla podanych parametrów

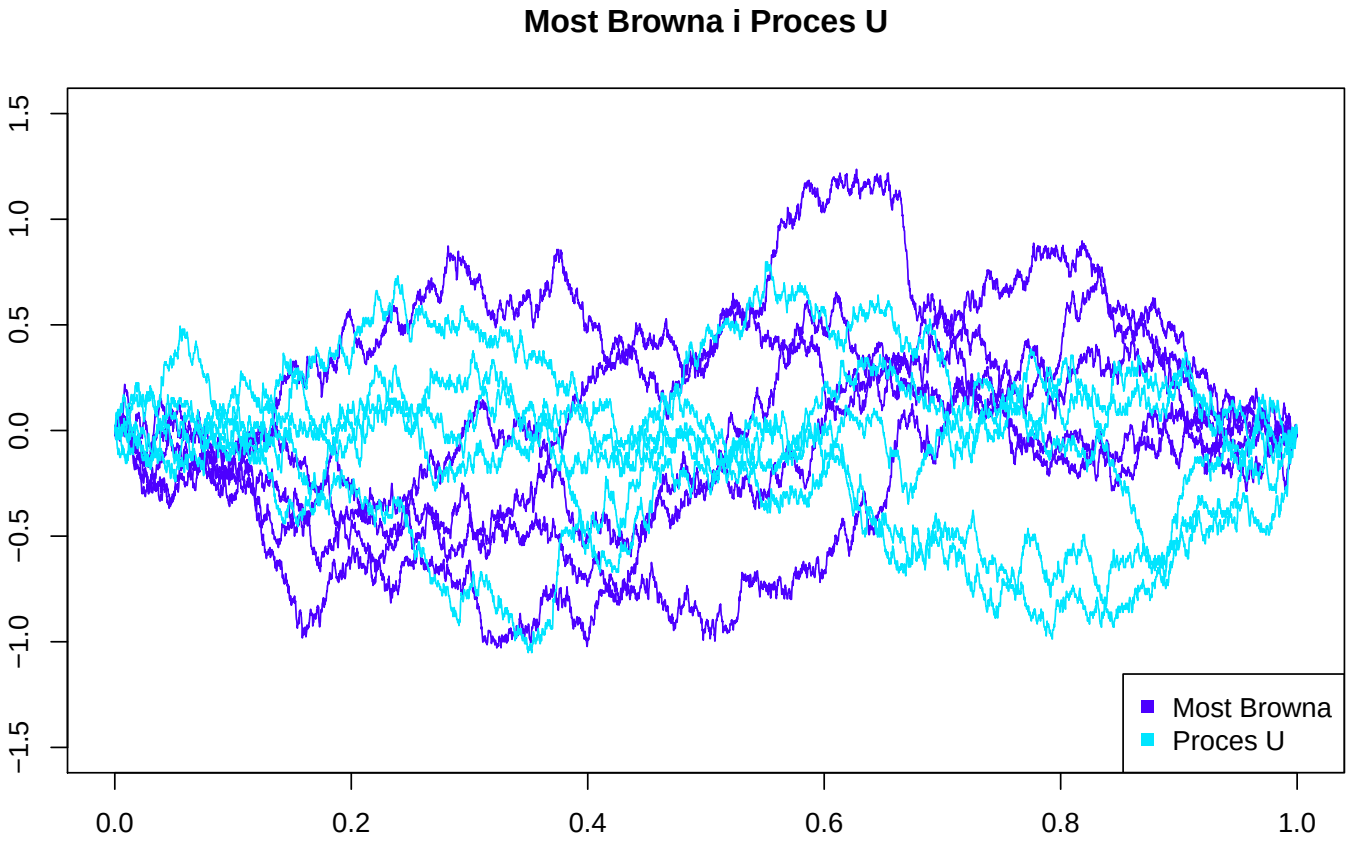
Widzimy, że wraz ze wzrostem parametrów p oraz r moce testu N-P rośnie, z kolei wzrost β powoduje spadek mocy. Jest to ooczywiste - wraz ze wzrostem β oraz p działają w przeciwny sposób na parametr ϵ naszego modelu.

5 Zadanie 4

W poleceniu ostatnim naszym zadaniem jest zasymolować dwa rodzaje trajektorii - Most Browna $B(t)$ oraz proces $U_p(t)$ zdefiniowany następująco

$$U_p(t) = \sqrt{p}(F_p(t) - (t))F_p(t) = \frac{|i : p_i \leq t|}{t}$$

gdzie p_i to p-wartości wygenerowane ze standardowego rozkładu normalnego. Na poniższym wykresie prezentujemy po 5 trajektorii z każdego z procesów.



Z wykresu, trudno jest jednoznacznie coś wywnioskować, jednakże gdyby nie legenda procesy byłoby trudno rozróżnić. W celu analizy podobieństwa porównamy kwantyle próbkowe rzędu 80% dla statystyk $T_1 = \sup_{t\in(0,1)} |B(t)|$ oraz $T_2 = sup_{t\in(0,1)}|U_p(t)|$. Wynoszą one odpowiednio 1.0744 oraz 1.0549, co potwierdza teorię przedstawioną na wykładzie mówiacą, że $T1 \xrightarrow{p\to\infty} T2$. Kwantyle zostały wyliczone dla 1000 replikacji.