Teoria analizy dużych zbiorów - Lista VI

Ryzyko estymacyjne

MM

14 czerwiec 2017

Spis treści

1 Wstęp

W nimniejszym raporcie umieszczone zostały rozwiązania szstej listy zadań z przedmiotu **Teoria analizy dużych zbiorów** prowadzonego przez Panią Profesor Małgorzatę Bogdan we współpracy z Panem Michałem Kosem. Na tejże liście poruszony został problem estymacji średniej w przypadku wielowymiarowego rozkładu normalnego. Poniżej przedstawimy cztery estymatory używane w kolejnych ćwiczeniach.

1.1 Założenia i definicje

Większośc poniższych definicji zostało wprowadzonych w raposrcie dot. listy piątej, jednakże przypomnimy je tutaj dla porządku.

Zakładamy, że dysponujemy pojedynczą obserwacją X z p-wymiarowego rozkładu normalnego $N(\mu, I)$, gdzie μ to wektor średnich, a Σ to macierz kowariancji.

Do oceny estymatorów użyjemy estymatora błędu średniokwadratowego [MSE] zdefinowanego następująco

$$\mathrm{MSE} = \frac{1}{p}\sum_{i=1}^{p}(X_i - \hat{X}_i)^2$$

gdzie X_i oraz $\hat{X}_i$ to odpowiednio i-ta współrzędna i estymator jej średniej.

1.2 Estymatory

1.2.1 Estymator największej wiarygodności [MLE]

Najprostszy estymator, to estmator największej wiarygodności, który w przypadku wielowymiarowego rozkładu normalnego jest średnią obserwacji X. Mamy więc

$$\hat{\mu}_{MLE} = X \qquad \text{[średnia]}.$$

1.2.2 Estymator Jamesa-Steina [JS]

Estymator Jamesa-Steina, to estmator który, zgodnie z teorią, powinien wykazywać mniejszy błąd średniokwadratowy niż estymator największej wiarygodności. Zadany jest on wzorem

$$\hat{\mu}_{JS} = \left(1 - \frac{p-2}{\|X\|^2}\right)X.$$

1.2.3 Estymator Jamesa-Steina z modyfikacja Mary Ellen Bock (1975) [MEB]

Jest to modyfikacja estymatora JS, która pozwala na estymację, gdy zmienne są od siebie zależne. Zadany jest on poprzez

$$\hat{\mu}_{MEB} = \left(1 - \frac{\hat{p}-2}{X^T\Sigma^{-1}X}\right)X,$$

gdzie $\hat{p} = \frac{Tr(\Sigma)}{\lambda_{max}(\Sigma)}$, a $Tr(\Sigma)$ i $\lambda_{max}(\Sigma)$ to odpowiednio ślad i największa wartość własna macierzy Σ.

1.2.4 Estymator odcięciowy [HT]

Jest to nietypowy estymator, używany w głównie w mieszaninach rzadkich. Opera się on na zasadzie odcięcia:

$$\hat{\mu}_H(x_i) = x_1 \mathbb{1}_{\{|x_i|>\lambda\}},$$

gdzie λ jest dobierana tak, aby zapewniać kontrolę odpowiednich błędów przy założeniu, że dla zdecydowanej większości zmiennych $\mu = 0$. Widać, że estymator ten nadaje się do mieszanin rzadkich, do testowania problemów w takich mieszaninach. Poziom λ dobierany jest w zależności o postawionego celu, może być oparty np. o korekte Bonferroniego czy procedurę Benaminiego-Hochberga. Poniżej przedstawimy postać tych odcieć dla testowania hipotez, p_i to p-wartość na i-tej współrzędnej, $p_{(i)}$ to oczywiście i-ta uporządkowana p-wartość:

- Dla kontrolowania FWER używa się korekty Bonferroniego, udowodnione zostało, że dla korekty postaci $\lambda = q/n$ kontrola FWER wynosi q,
- Dla kontrolowania FDR stosuje się procedurę BH(q), czyli $\lambda = p(i_0)$, gdzie $i_0 = \arg\max_i(p(i) \leq \frac{i}{n}q$.

2 Zadanie 1

Zadanie pierwsze to zmodyfikowane zadanie drugie z listy piątej. Porównamy estymatory *MLE*, *MEB* w trzech róznych przypadkach:

- A. $\mu = 0$,
- B. μ pochodzi z rozkładu $N(0, 5I)$,
- C. $\mu_i \sim N(20, 5)$.

Zakładamy, że macierz kowariancji nie jest macierzą diagonalną. Wtedy nad estymatorer *MLE* powinien przeważać (w sensie błędu średniokwadratowego) estymator *MEB*. Głównym problemem przy jego stosowaniu jest wymagana znajomości macierzy kowariancji

Zakładamy, że $X = (X_1, X_2, ..., X_p) \sim N(\mu, \Sigma)$, gdzie $\Sigma_{i,i} = 1$, a $\Sigma_{i,j} = 0.4$ dla $i \neq j$.

Wyniki:

	MLE	MEB
A	0.99628	0.99430
B	0.98186	0.98163
C	1.01294	1.01284

Tablica 1: Estymowane błędy średniokwadratowe, niezerowa korelacja

Zaobserwowana różnica pomiędzy estymatorem MLE, a MEB jest niewielka, ale pokazuje, że estymator MEB spisuje się lepiej w powyższym problemie. tak mała różnica może wynikać z niskiej wartości $\hat{p}$, równej 2.4925. Po raz kolejny, przytaczając teorię podaną na wykładzie, dopiero jeżeli $\hat{p} \geq 2$ to estymator MEB ma mniejszy MSE niż MLE.

3 Zadanie 2

W zadaniu porównamy estymatory *MLE* i *MEB*, oraz dwie postaci reguły odcięcia w następujących przypadkach:

- A. $\mu_1 = ... = \mu_5 = 3.5, \mu_6 = ... = \mu_{500} = 0$,
- B. $\mu_1 = ... = \mu_{30} = 2.5, \mu_3 1 = ... = \mu_{500} = 0$,
- C. $\mu_1 = ... = \mu_{100} = 1.8, \mu_3 1 = ... = \mu_{500} = 0$,
- D. $\mu_1 = ... = \mu_{500} = 0.4$,
- E. $\mu_i = 3.5 * i^{-1/2}$,
- F. $\mu_i = 3.5 * i^{-1}$.

Pierwsza dwa estymatory zostały wczesniej opisane, reguła odciecia natomiast będzie użyta dwukrotnie:

- $\lambda = \lambda_{Bonf}$, przy czym próg jest ustalony tak, aby kontrolować FWER na poziomie 0.1,
- $\lambda = \lambda_{BH}$, przy czym próg jest ustalony tak, aby kontrolować FDR na poziomie 0.1

Reguła odcięcia opera się na zasadzie "keep signal, kill noise".

Otrzymane wyniki:

	MLE	JS	HTbonf	HTbh
A	0.99885	0.11231	0.07009	0.06535
B	1.00330	0.27532	0.34382	0.30843
C	1.00448	0.39758	0.66204	0.65505
D	1.00156	0.14158	0.17035	0.18031
E	1.00085	0.14691	0.16326	0.16514
F	1.00175	0.04290	0.03374	0.03415

Tablica 2: Błędy średniokwadratowe, porównanie estymatorów

Widzimy, że dla każdego z zagadnień błąd średniokwadratowy estymatora największej wiarygodność oscyluje w okolicach wariancji, co jest jasne. Estymator JS zdecydowanie zmniejsza średni bład kwadratowy, dla każdego z przypadków. Estymatory hard-threshold w problemie A zachowują się lepiej niż dwa pozostałe. Najbardziej uniwersalny wydaje sie estymator [JS], gdzy jest zdecydowanie lepszy niż [MLE], a w najgorszych wypadkach jest niewiele gorszy nic estymatory odcinające.