Teoria analizy dużych zbiorów - Lista V

Estymacja średniej rozkładu

MM

31 maja 2017

Spis treści

1 Wstęp **2**
1.1 Założenia i definicje 2
1.2 Estymatory 2

2 Zadanie 1 **3**

3 Zadanie 2 **3**

1 Wstęp

W nimniejszym raporcie umieszczone zostały rozwiązania piątej 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

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

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

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

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

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 empriyczny Bayesa

Estymator empiryczny Bayesa, to estmator który, wykorzystuje statystyke Bayesowską do estymacji, wykorzystuje pozostałe obserwacje do estymacji rozkładu a priori. Zdefiniowany jest osobno dla każdego elementu wektora μ poprzez

$$\hat{\mu}_{i_{EB}} = \bar{X} + \left(1 - \frac{p-3}{S}\right)(X_i - \bar{X}),$$

gdzie $S = \sum_{i=1}^{p}(X_i - \bar{X})^2$ a $\bar{X}$ to oczywiście średnia wszystkich obserwacji.

1.2.4 Estymator Jamesa-Steina z modyfikacja Mary Ellen Bock (1975)

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.

2 Zadanie 1

W zadaniu pierwszym porównamy pierwsze trzy estymatory w trzech róznych wypadkach:

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

Oczywiście $X = (X_1, X_2, ..., X_p) \sim N(\mu, I)$. Przyjeliśmy $p = 500$. Miarą dobroci jest uśredniony błąd średniokwadratowy dla 500 symulacji.

Wyniki są następujące:

	MLE	JS	EB
A	0.98089	0.40755	0.59173
B	1.00482	0.95984	0.97832
C	0.96304	0.96254	0.90657

Tablica 1: Estymowane błędy średniokwadratowe

Z symulacji możemy wyciągnąc następujące wnioski:

- Estymator MLE daje zdecydowanie gorsze wyniki niż pozostałe, biorąc pod uwagę MSE
- Przy założeniu rozkładu a priori parametru μ estymator JS jest lepszy lub równy estymatorowi Bajesowskiemu w mierze błędu średniokwadratowego.

Obydwa spostrzeżenia pokrywają się z teorią przedstawioną na wykładzie.

3 Zadanie 2

Zadanie drugie porusza to samo zagadnienie co zadanie pierwsze, z tym, że zakładamy tutaj, że macierz kowariancji nie jest macierzą diagonalną. W takim przypadku powinniśmy zastosować estymator MEB. Problemem jest tutaj wymaganie dot. znajomości macierzy kowariancji rozkładu.

Zakładamy, że $X = (X_1, X_2, ..., X_p) \sim N(\mu, \Sigma)$, gdzie $\Sigma_{i,i} = 1$, a $\Sigma_{i,j} = 0.7$ dla $i \neq j$. Pozostałe parametry jak w zadaniu pierwszym.

Otrzymane wyniki:

	MLE	MEB
A	0.98896	1.30034
B	1.00378	1.01831
C	0.93254	0.93441

Tablica 2: Estymowane błędy średniokwadratowe, niezerowa korelacja I

Zaobserwowana różnica pomiędzy estymatorem MLE, a MEB jest znikoma wręcz na niekorzyść MEB. Wynika to z niskiej wartości $\hat{p}$, równej 1.3158. Przytaczając teorię podaną na wykładzie, dopiero jeżeli $\hat{p} \geq 2$ to estymator MEB ma mniejszy MSE niż MLE.

Sprawdżmy co się stanie, gdy zmodyfikujemy nasz problem $X = (X_1, X_2, ..., X_p) \sim N(\mu, \Sigma)$, gdzie $\Sigma_{i,i} = 1$, a $\Sigma_{i,j} = 0.2$ dla $i \neq j$. Dla takich parametrów $\hat{p}$ jest równe 2.7778. Reszta parametrów pozostaje bez zmian.

	MLE	MEB
A	0.98319	0.70876
B	1.00452	0.98146
C	0.95432	0.95325

Tablica 3: Estymowane błędy średniokwadratowe, niezerowa korelacja II

Widoczna jest bardzo mała poprawa estymatora, co zgodne jest z teorią. Jest ona jednak na tyle mała, że warto byłoby zastanowić się na prędkościa oddalania się estymatorów od siebie. Pozostawimy to jako problem otwarty.