Analiza dużych zbiorów danych raport nr 5

Dominik Mika

13 października 2021

Zadanie 1

W pierwszym zadaniu wygenerujemy macierz danych $X_{500\times100}$ zgodnie z formułą:

$$\begin{gathered} F_{500\times3}: \quad \text{dla} i \in \{1,\cdots,3\} \quad F_{,i} \sim N(0, I_{500\times500}) \\ W_{100\times3}: \quad W_{,j} \sim (4-j)N(0, I_{100\times100}) \\ E_{500\times100}: \quad E_{i,j} \sim N(0,1) \\ X = FW^T + E \end{gathered}$$

Dla tak wygenerowanej macierzy wykonamy PCA stosując odpowiednio:

a) Rozkład spektralny macierzy X^TX

b) Rozkład macierzy X na wartości osobliwe

Następnie wyznaczymy współczynniki korelacji między pierwszymi składowymi głównymi a odpowiednimi kolumnami macierzy F. Dodatkowo wyznaczymy estymator wariancji błędu σ^2 zgodnie z formułą:

$$\hat{\sigma}^2 = \frac{1}{n(p-3)} \sum_{i=1}^{3} \lambda_i,$$

gdzie λ_i należy do trzech największych wartości własnych macierzy X^TX. Po tych obliczeniach wybierzemy jedną z powyższych metod wyznaczania PCA i powtórzymy ten eksperyment 100 razy w sytuacji gdy $n = 500$ oraz w sytuacji gdy $n = 2000$. Na koniec narysujemy wykresy pudełkowe współczynników korelacji między składowymi głównymi a odpowiednimi kolumnami macierzy F oraz estymatora σ^2.

Rozwiązanie

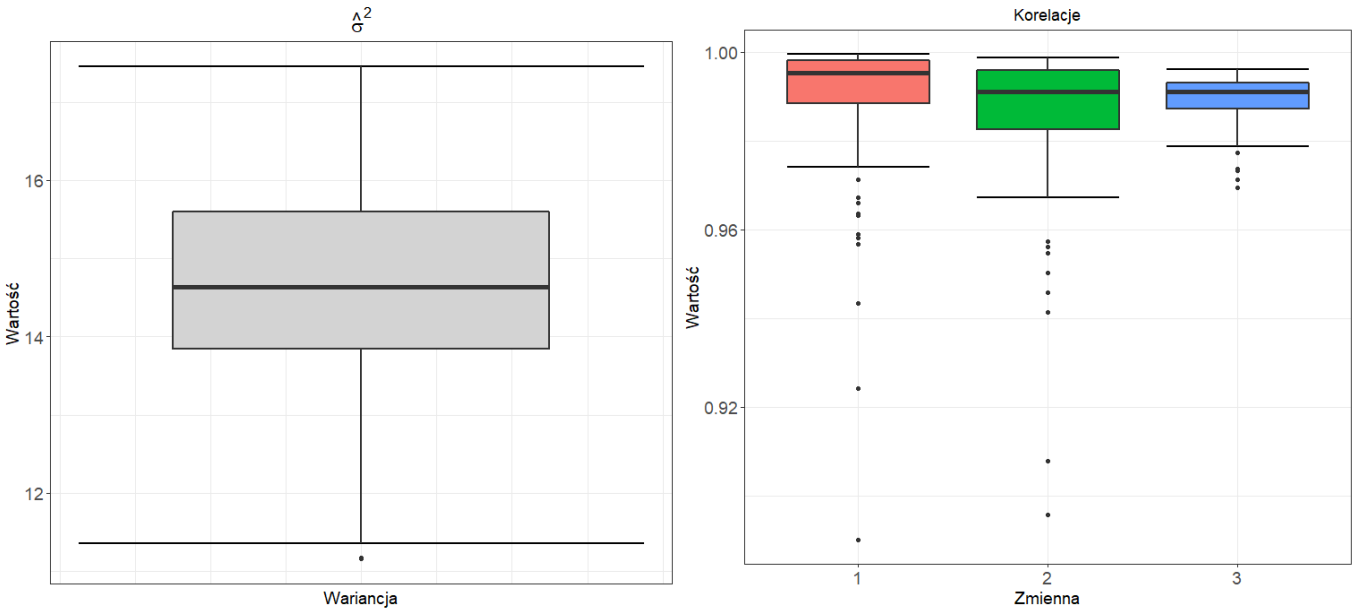
Wyestymowana wariancja z powyższego wzoru dla pojedyńczego powtórzenia wyniosła: 16.574. Co jest dość słabym wynikiem, ponieważ teoretyczna wariancja wynosi 1, więc widzimy, że wartość tego estymatora jest dużo większa od faktycznej wartości.

Natomiast współczynniki korelacji między pierwszymi składowymi głównymi a odpowiednimi kolumnami macierzy F znajdują się w poniższej tabeli:

Zmienna	1	2	3
Wartość	0.989	0.979	0.996

Widzimy, że są one bliskie jedynki co oznacza, że dobrze wyznaczyliśmy składowe główne.

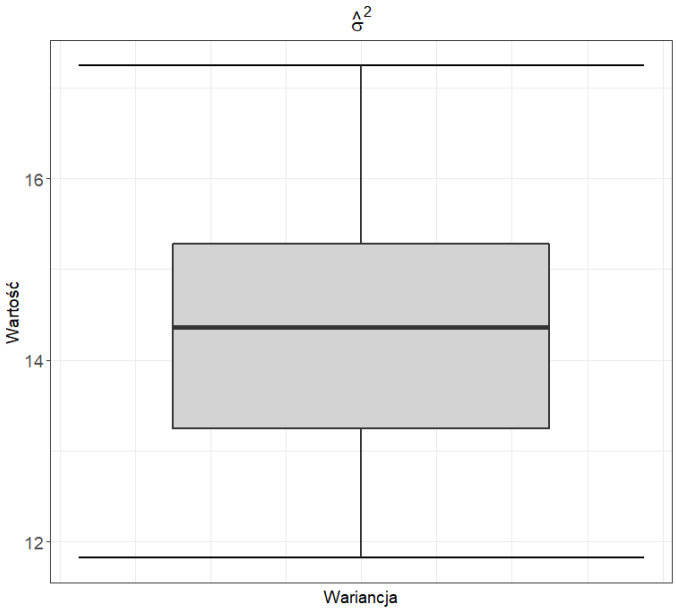
Teraz powtórzymy powyższy eksperyment 100 razy i narysujemy wykresy pudełkowe poprzednich parametrów. Wykresy znajdują się poniżej:

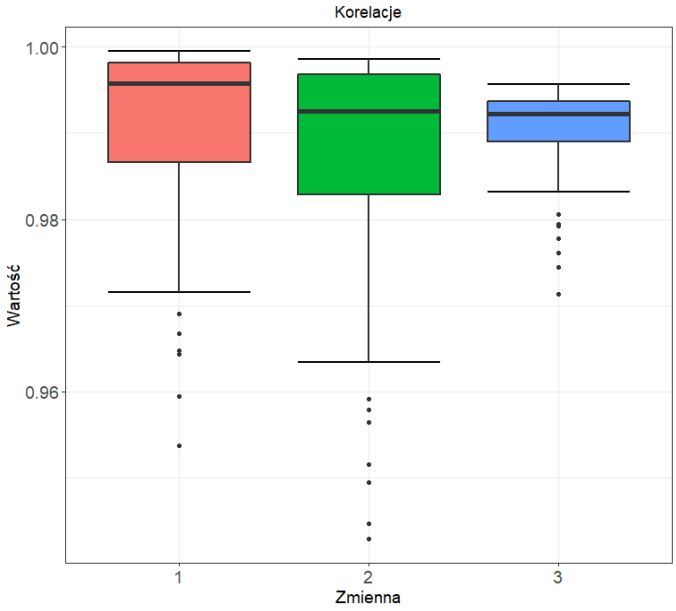


Patrząc najpierw na wykres dla wariancji widzimy, że nasz poprzedni wynik nie był przypadkiem. Wartość głównie oscyluje między 14, a 16- tyle około wynosi pierwszy i trzeci kwartyl. Średnia natomiast wynosi 14.62. Teoretyczna wartość wynosi 1, więc możemy powiedzieć, że ten estymator słabo radzi sobie z estymacją w tym przypadku.

Natomiast patrząc na wykres dla korelacji możemy zauważyć, że są one bardzo bliskie jedynki. Co jest dobrą wiadomością, ponieważ oznacza to, że dobrze wyznaczyliśmy składowe główne, które są mocno skorelowane z tymi faktycznymi.

Teraz powtarzamy poprzedni eksperyment tylko tym razem dla $n = 2000$. Wykresy pudełkowe znajdują się poniżej:





Na wykresach dla przypadku, gdy $n = 2000$ obserwujemy podobne zjawiska co na poprzednich. Wartości wariancji są nieco mniejsze, ale nadal odstają od faktycznej wartości. Dla korelacji widzimy, że pudełka stały się szersze, ale widzimy, że skala osi y się zmieniła i wartości korelacji nawet bliżej 1 niż poprzednio.

Zadanie 2

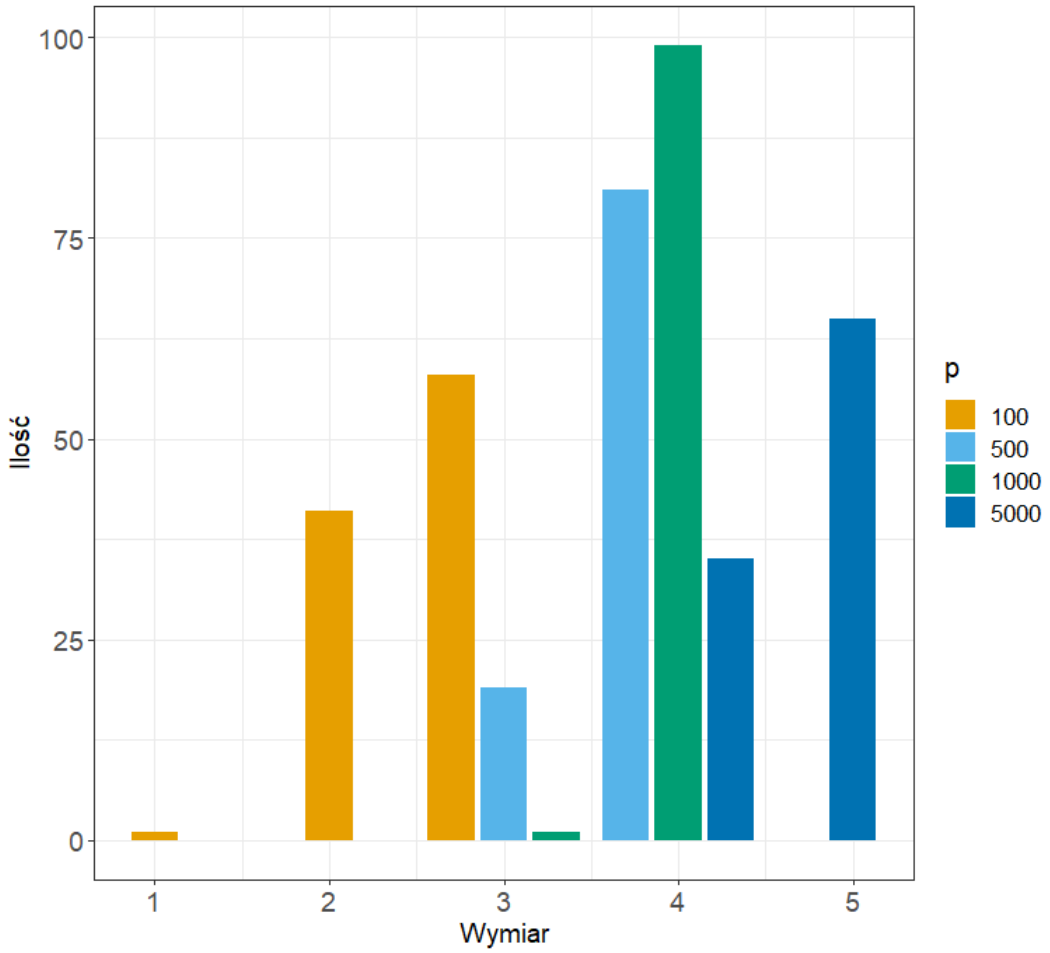
W tym zadaniu dla $n = 50$ i $p \in \{100, 500, 1000, 5000\}$ wygenerujemy 100 niezależnych replikacji macierzy $X_{n\times p}$ zgodnie ze wzorem:

$$
\begin{aligned}
F_{n\times 5}: \quad & \text{dla} \quad i \in \{1, \cdots, 5\} \quad F_{,i} \sim N(0, I_{n\times n}) \\
W_{p\times 3}: \quad & W_{,j} \sim (6-j)N(0, I_{p\times p}) \\
E_{n\times p}: \quad & E_{i,j} \sim N(0, \sigma = 10) \\
& X = FW^T + E
\end{aligned}
$$

Dla każdej kombinacji n i p narysujemy histogram wymiaru wybranego przez PESEL.

Rozwiązanie

Dla lepszego zobrazowania, wyniki przedstawimy na jednym wykresie słupkowym, który znajduje się poniżej:



Pierwsze co obserwujemy na powyższym wykresie, to to, że wraz ze wzrostem p szacowany wymiar się zwiększa i zbliża się do tego faktycznego, czyli 5. Dla $p = 100$ nigdy nie wyznaczyliśmy dobrego wymiaru, raczej był wybierany wymiar 2 lub 3. Dla $p = 500$ i $p = 1000$ najczęściej wybierany wymiar to 4. Tylko w przypadku $p = 5000$ w większości wybieraliśmy dobry wymiar, ale nie zawsze.

Zadanie 3

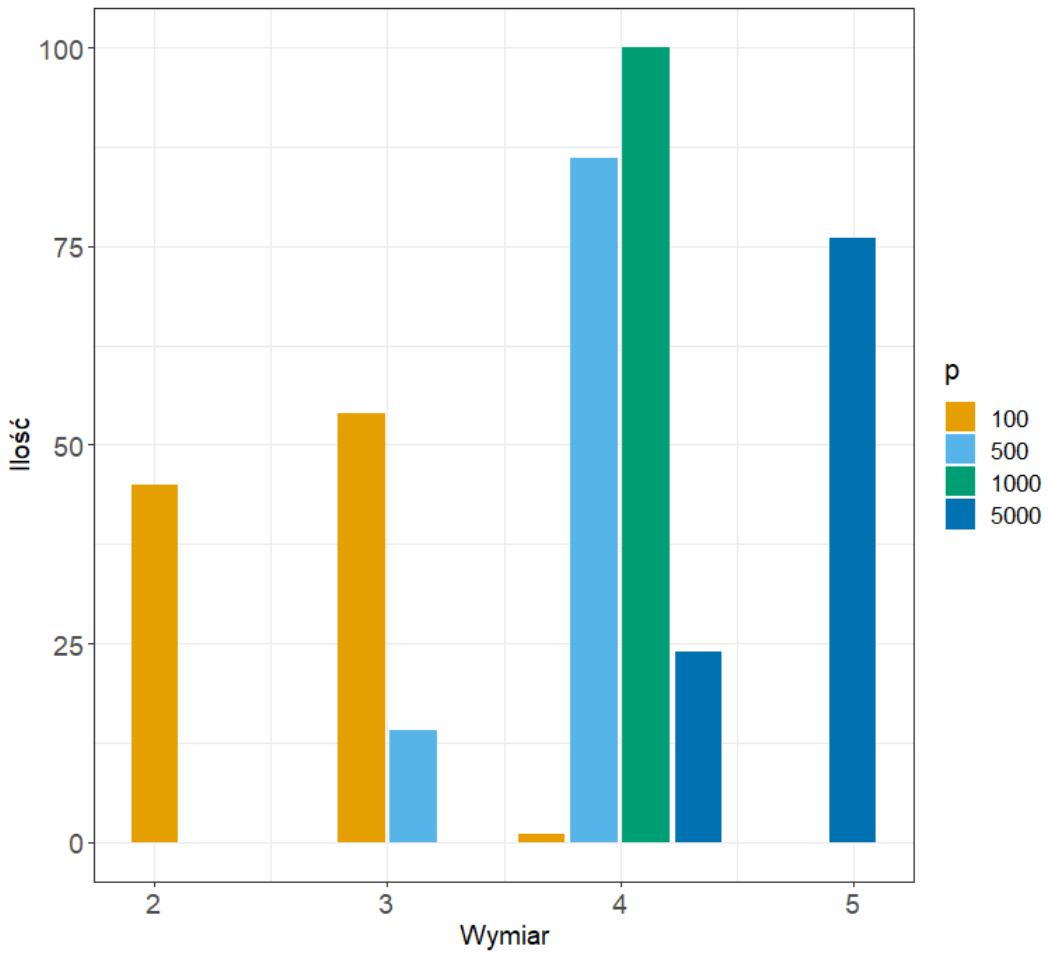
W ostatnim zadaniu powtórzymy poprzednie, gdy elementy macierzy błędów E generowane są jako:

a) Niezależne zmienne losowe z rozkładu wykładniczego z $\lambda = 0.1$,

b) Niezależne zmienne losowe z rozkładu Cauchy'ego.

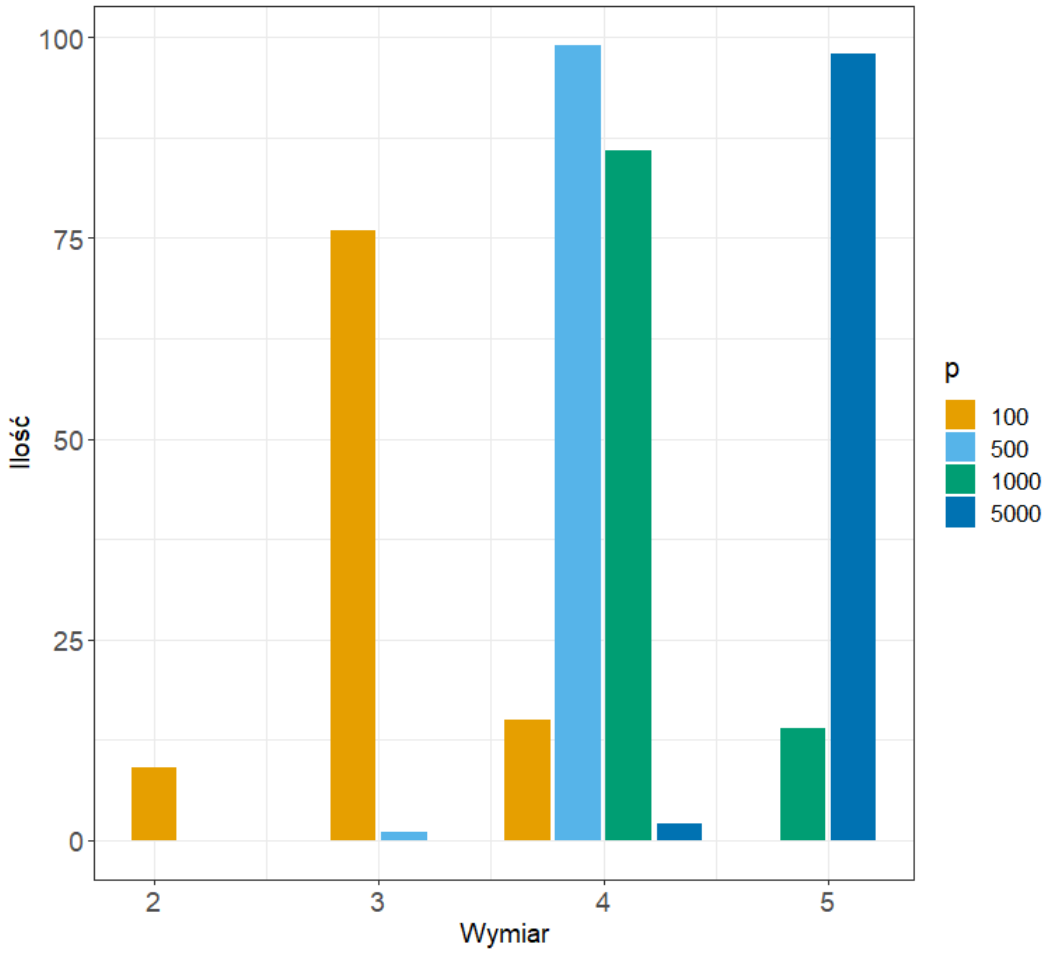
Rozwiązanie

W tym zadaniu wyniki przedstawimy analogicznie do poprzedniego. Poniższy wykres odnosi się do eksperymentu z błędem wylosowanym z rozkładu wykładniczego:



Porównując powyższy wykres z tym z poprzedniego zadania nie widzimy znaczących różnic. Możemy powiedzieć nawet, że nasze wyniki się poprawiły od tych poprzednich. Ani razu nie wybraliśmy wymiaru 1, a wyniki 4 i 5 występowały częściej.

Natomiast poniżej znajduje się wykres słupkowy, gdy błąd losowaliśmy z rozkładu Cauchy'ego:



Na powyższym wykresie widzimy, że wyniki się poprawiły. Zdecydowanie częściej wybieramy 4 i 5 wymiarów. Dla $p = 5000$ prawie zawsze dobrze wyznaczamy wymiar.

Podsumowując, możemy powiedzieć, że wraz ze wzrostem p metoda PESEL działała dużo lepiej, czyli tak jakbyśmy się spodziewali. Natomiast co ciekawe zmiana rozkładu macierzy błędu nawet poprawiła nasze wyniki. Nie musimy się więc martwić, gdy błąd nie pochodzi z rozkładu normalnego.