Testowanie hipotez związanych ze składowymi głównymi

Oznaczenia i założenia

$\mathbf{X} = \left[X_1, X_2, \ldots, X_p\right]^{\mathrm{T}}$ - wektor losowy (wektor cech)

$\mathbf{X}$ ma rozkład $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} > \mathbf{0}$

$\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$, gdzie $N > p$ – próba z rozkładu tego wektora

$\mathbf{S}$ – nieobciążony estymator macierzy kowariancji $\boldsymbol{\Sigma}$

$$\mathbf{S} = \frac{1}{N-1}\sum_{i=1}^{N}\left(\boldsymbol{X}_i - \overline{\boldsymbol{X}}\right)\left(\boldsymbol{X}_i - \overline{\boldsymbol{X}}\right)^{\mathrm{T}}$$

Oznaczenia i założenia cd.

$\lambda_1 \geq \ldots \geq \lambda_p$ wartości własnych macierzy $\mathbf{\Sigma}$

$l_1 > l_2 > .. > l_p$ – wartości własne macierzy $\mathbf{S}$

Testowanie hipotez związanych ze składowymi głównymi

Test sferyczności macierzy kowariancji $\mathbf{\Sigma}$ służy do weryfikacji hipotezy, że wszystkie wartości własne macierzy kowariancji $\mathbf{\Sigma}$ są równe

$$H_0: \quad \lambda_1 = \ldots = \lambda_p$$

Przyjęcie tej hipotezy oznacza, że wszystkie składowe główne mają tę samą wariancję i równy wkład do zmienności całkowitej. Nie można wówczas zredukować wymiaru przez przejście do składowych głównych.

Równość wszystkich *p* wartości własnych?

$$\lambda_1 = \ldots = \lambda_p \qquad (= \lambda) \quad ?$$

TAK ↓

Wszystkie składowe główne mają tę samą wariancję,
zatem równy wkład do zmienności całkowitej;
nie można zredukować wymiaru przez przejście do składowych głównych.

NIE ↓

Równość *p*-1 najmniejszych wartości własnych?
I są one znacząco mniejsze od pierwszej wartości własnej?

$$\lambda_2 = \ldots = \lambda_p \qquad (= \lambda) \quad ?$$

i jeszcze $\lambda_1 >> \lambda$?

Równość p-1 najmniejszych wartości własnych?
I są one znacząco mniejsze od pierwszej wartości własnej?

$$\lambda_2 = \ldots = \lambda_p \quad (= \lambda) \quad ?$$

i jeszcze $\lambda_1 >> \lambda$?

TAK ↓

NIE →

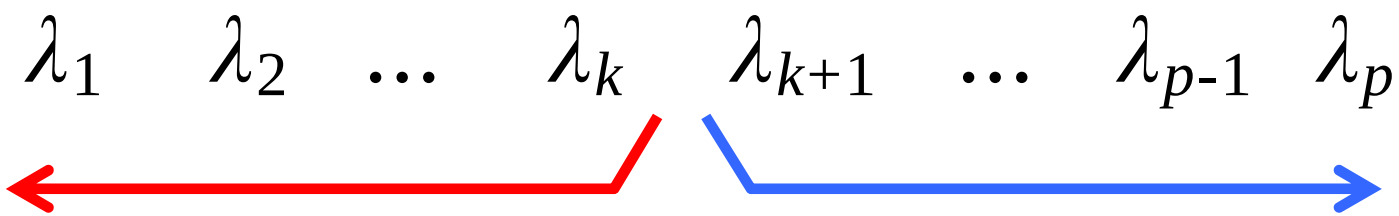
Ostatnich p-1 składowych głównych ma tę samą wariancję, zatem równy wkład do zmienności całkowitej i wkład każdej z nich jest znacząco mniejszy od wkładu pierwszej; **można zredukować wymiar, bo większość zmienności w próbie jest wyjaśniona przez pierwszą składową główną.**

Równość p-2 najmniejszych wartości własnych?

$$\lambda_3 = \ldots = \lambda_p \quad (= \lambda) \quad ?$$

itd.

Objaśnienie oznaczeń



„odcięte” *waw*

waw porównywane w hipotezie H_k

k

W hipotezie:
H_0 porównujemy p-0 = p (wszystkie) *waw*;
H_1 porównujemy p-1 najmniejszych wartości własnych,
H_2 porównujemy p-2 najmniejszych wartości własnych, itd.
...
H_k porównujemy p-k najmniejszych wartości własnych

Sekwencyjne testowanie hipotez

$$\mathrm{H}_k: \quad \lambda_{k+1} = \ldots = \lambda_p, \quad \text{dla} \quad k = 0, 1, , \ldots, p-2,$$

gdzie $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$ są wartościami własnymi **Σ.**

Test hipotezy $\mathbf{H_0}$

$$H_0: \quad \lambda_1 = \ldots = \lambda_p$$

jest oparty na statystyce

$$V_0 = \frac{l_1 \cdot l_2 \cdot \ldots \cdot l_p}{\left(\frac{1}{p} \sum_{i=1}^{p} l_i \right)^p}$$

gdzie

$l_1 > l_2 > \ldots > l_p$ - wartości własne macierzy $\boldsymbol{S}$

Test hipotezy $\mathbf{H_0}$ **cd.**

$$H_0: \quad \lambda_1 = \ldots = \lambda_p$$

Test asymptotyczny odrzuca hipotezę H_0 na poziomie istotności α, gdy

$$-\left(N-1-\frac{2p^2+p+2}{6p}\right) ln V_0 > c(\alpha; r)$$

gdzie:

$$r = \tfrac{1}{2}(p+2)(p-1)$$

$c(\alpha; r)$ takie, że $P\{\chi_r^2 \geq c(\alpha; r)\} = \alpha$

Test hipotezy $\mathbf{H}_k$ *- wynik Bartletta (1954)*

Test asymptotyczny hipotezy

$$\mathrm{H}_k: \quad \lambda_{k+1} = \ldots = \lambda_p \qquad (= \lambda, \quad \text{nieznane})$$

jest oparty na statystyce

$$-\left(\mathrm{N} - 1 - k - \frac{2q^2 + q + 2}{6q}\right) ln\,\mathrm{V}_k$$

która ma rozkład $\chi^2_{(q+2)(q-1)/2}$, $q = p\text{-}k$, $k = 0, 1, \ldots, p\text{-}2$.

$$\mathrm{V}_k = \frac{l_{k+1} \cdot \ldots \cdot l_p}{\left(\frac{1}{p-k}\sum_{i=k+1}^{p} l_i\right)^{p-k}}$$

Test hipotezy H_k – wynik Lawleya (1956)

z dodatkową informacją o dokładności aproksymacji rozkładem chi-kwadrat pochodzącą od Jamesa (1969)

Test asymptotyczny hipotezy

$$\mathrm{H}_k : \quad \lambda_{k+1} = \ldots = \lambda_p \qquad (= \lambda, \quad \text{nieznane})$$

oparty jest na statystyce

$$P_k = -\left(\mathrm{N} - 1 - k - \frac{2q^2 + q + 2}{6q} + \sum_{i=1}^{k} \frac{\bar{l}_q}{\left(l_i - \bar{l}_q\right)^2} \right) ln \mathrm{V}_k$$

Test hipotezy H_k – wynik Lawleya (1956) cd.

Statystyka testowa

$$P_k = -\left(N - 1 - k - \frac{2q^2 + q + 2}{6q} + \sum_{i=1}^{k} \frac{\bar{l}_q}{\left(l_i - \bar{l}_q\right)^2} \right) ln \mathrm{V}_k$$

gdzie:

$$\bar{l}_q = \frac{1}{q} \sum_{i=k+1}^{p} l_i, \quad q = p - k,$$

$$\mathrm{V}_k = \frac{l_{k+1} \cdot \ldots \cdot l_p}{\left(\frac{1}{p-k} \sum_{i=k+1}^{p} l_i \right)^{p-k}}$$

Test hipotezy H_k – wynik Lawleya (1956) cd.

Statystyka testowa

$$P_k = -\left(N - 1 - k - \frac{2q^2 + q + 2}{6q} + \sum_{i=1}^{k} \frac{\bar{l}_q}{\left(l_i - \bar{l}_q \right)^2} \right) ln \mathrm{V}_k$$

odrzuca hipotezę H_k na poziomie istotności α, gdy

$$P_k > c(\alpha; r)$$

gdzie:

$$r = \tfrac{1}{2}(q+2)(q-1), \quad q = p - k$$

$c(\alpha; r)$ takie, że $P\left\{ \chi_r^2 \geq c(\alpha; r) \right\} = \alpha$

Zakończenie sekwencyjnego testowania hipotez

Przypuśćmy, że dla pewnego k hipoteza H_k została przyjęta.

Zatem możemy przyjąć, że $q = p\text{-}k$ najmniejszych wartości własnych macierzy $\mathbf{\Sigma}$ ma tę samą wartość λ. Jeżeli wartość λ jest znacząco mniejsza od pozostałych wartości własnych, to można pominąć ostatnich q składowych głównych i zostawić tylko k pierwszych, tym samym redukując wymiar przestrzeni danych.

Na jakiej podstawie zdecydować, o małej wartości λ?

Podstawa decyzji o małej wartości λ - przedział ufności

Jednostronny przedział ufności dla λ (Anderson, 1963)
Asymptotycznie na poziomie ufności 1-α

$$\lambda \leq \frac{\bar{l}_q}{1 - z_\alpha \sqrt{\frac{2}{(N-1)\, q}}}$$

gdzie:

$$\bar{l}_q = \frac{1}{q} \sum_{i=k+1}^{p} l_i, \quad q = p - k$$

z_α takie, że $F(z_\alpha) = 1 - \alpha$, gdzie F jest dystrybuantą rozkładu $N(0,1)$.

Inne kryterium redukcji liczby składowych

Jeżeli nie można podjąć decyzji, że pewna liczba wartości własnych ma tę samą wartość, można sprawdzić, czy zmienność wyjaśniana przez ostatnich $q = p\text{-}k$ składowych głównych

$$\sum_{i=k+1}^{p} \lambda_i$$

jest mała w porównaniu ze zmiennością całkowitą

$$\sum_{i=1}^{p} \lambda_i$$

co pozwoliłoby pominąć ostatnich $q = p\text{-}k$ składowych głównych.

Hipoteza o udziale wyjaśnionej zmienności

$$H_k^*: \quad \frac{\sum_{i=k+1}^{p} \lambda_i}{\sum_{i=1}^{p} \lambda_i} = h$$

gdzie
h $(0 < h < 1)$ – wartość ustalona przez eksperymentatora

Do weryfikacji tej hipotezy służy statystyka

$$M_k = \sum_{i=k+1}^{p} l_i - h\sum_{i=1}^{p} l_i = -h\sum_{i=1}^{k} l_i + (1-h)\sum_{i=k+1}^{p} l_i$$

Wskazówka

Przy założeniu, że $\lambda_1, \lambda_2, \ldots, \lambda_p$ są różne i $N \to \infty$,statystyka

$$\sqrt{N-1}\left(M_k + h\sum_{i=1}^{k}\lambda_i - (1-h)\sum_{i=k+1}^{p}\lambda_i \right)$$

ma asymptotycznie rozkład $N(0, \tau^2)$, gdzie

$$\tau^2 = 2h^2\sum_{i=1}^{k}\lambda_i^2 + 2(1-h)^2\sum_{i=k+1}^{p}\lambda_i^2 .$$

Po zastąpieniu λ_i przez l_i ($i = 1, 2, \ldots, p$) w wyrażeniu τ^2, wynik można wykorzystać do konstrukcji przybliżonego testu weryfikującego hipotezę H_k^* oraz do konstrukcji przedziałów ufności dla $\sum_{i=k+1}^{p}\lambda_i - h\sum_{i=1}^{p}\lambda_i$.

Przykład

Dane pochodzą z doświadczeń przeprowadzonych w 7 miejscowościach w jednym roku dla 62 odmian pszenicy ozimej. Wartości liczbowe przedstwiają średnie plony z powtórzeń.

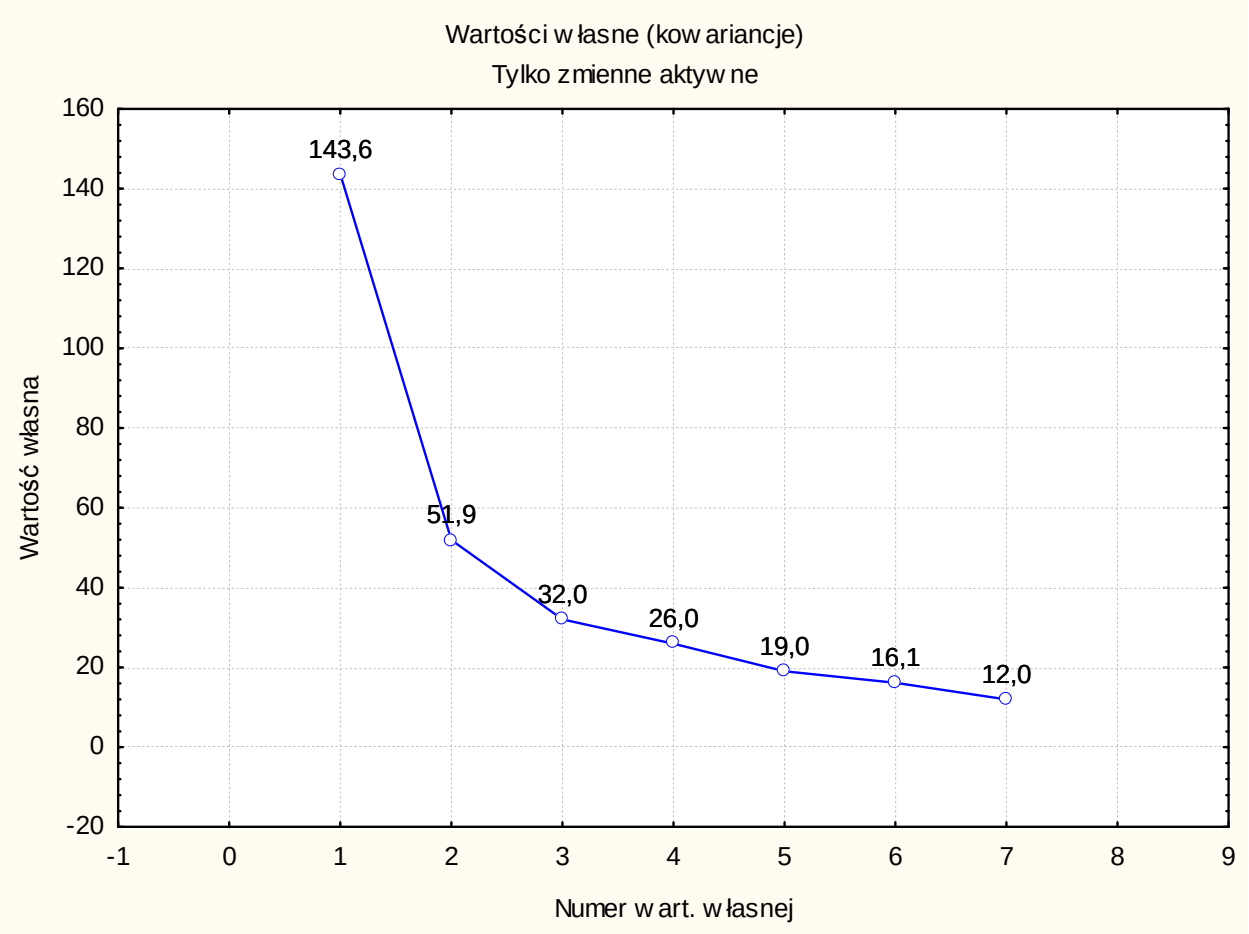
N=62, p=7

Macierz korelacji							
	M1	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**
M1	**1,00**	**0,32**	**0,42**	**0,31**	**0,41**	**0,28**	**0,34**
M2	**0,32**	**1,00**	**0,22**	**0,60**	**0,42**	**0,61**	**0,29**
M3	**0,42**	**0,22**	**1,00**	**0,17**	**0,40**	**0,13**	**0,31**
M4	**0,31**	**0,60**	**0,17**	**1,00**	**0,41**	**0,56**	**0,44**
M5	**0,41**	**0,42**	**0,40**	**0,41**	**1,00**	**0,42**	**0,49**
M6	**0,28**	**0,61**	**0,13**	**0,56**	**0,42**	**1,00**	**0,48**
M7	**0,34**	**0,29**	**0,31**	**0,44**	**0,49**	**0,48**	**1,00**

Przykład

Wartości własne macierzy kowariancji			
	lambda	% ogółu wariancji	skumulowany % ogółu wariancji
lambda 1	143,61	47,77	47,77
lambda 2	51,92	17,27	65,04
lambda 3	32,00	10,64	75,68
lambda 4	25,98	8,64	84,32
lambda 5	19,05	6,34	90,66
lambda 6	16,13	5,36	96,02
lambda 7	11,97	3,98	100,00

Przykład


Wartości w łasne (kow ariancje)
Tylko zmienne aktyw ne
Wartość własna
Numer w art. w łasnej
160
140
120
100
80
60
40
20
0
-20
-1
0
1
2
3
4
5
6
7
8
9
143,6
51,9
32,0
26,0
19,0
16,1
12,0

Przykład

Hipoteza H0 ***odrzucamy***

k=	0
q=	7
Vk=	0,085
Pk=	144,430
st sw=	27
wart kryt	40,11

Hipoteza H1 ***odrzucamy***

k=	1
q=	6
Vk=	0,494
Pk=	40,796
st sw=	20
wart kryt	31,41

Przykład

Hipoteza H2 ***przyjmujemy***

k=	2
q=	5
Vk=	0,744
Pk=	16,885
st sw=	14
wart kryt	23,68

przedział ufności
dla lambda

lambda średnia q=	21,02
z 0,05=	1,64
lambda<=	24,243

Literatura

* Krzyśko Mirosław, *Wielowymiarowa analiza statystyczna*, Wyd. UAM, Poznań 2000

- Morrison Donald Franklin, *Wielowymiarowa analiza statystyczna*, tłum. Wojciech Zieliński, PWN,Warszawa 1990

- Stanisz Andrzej, *Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny. Tom 3. Analizy wielowymiarowe*, Wyd. Statsoft, Kraków 2007