Zmienne kanoniczne

Obserwujemy wektor losowy

$$(Y_1, \ldots, Y_m, X_1, \ldots X_k)'$$

Badamy istnienie zależności między

$$\boldsymbol{Y} = (Y_1, \ldots, Y_m)' \qquad \boldsymbol{X} = (X_1, \ldots X_k)'$$

Macierz kowariancji

$$\mathcal{C} = \begin{bmatrix} \mathcal{C}_{11} & \mathcal{C}_{12} \\ \mathcal{C}_{12}' & \mathcal{C}_{22} \end{bmatrix}$$

Wektory $\boldsymbol{Y}$ i $\boldsymbol{X}$ są niezależne, jeżeli

$$\mathcal{C}_{12} = \mathbf{0}$$

$$\mathcal{C}_{12} = \mathbf{0}$$

$$\Updownarrow$$

$$(\forall \boldsymbol{a} \in \mathbb{R}^m,\ \boldsymbol{b} \in \mathbb{R}^k)\ \boldsymbol{a}'\mathcal{C}_{12}\boldsymbol{b} = 0$$

$$\Updownarrow$$

$$(\forall \boldsymbol{a} \in \mathbb{R}^m, \boldsymbol{b} \in \mathbb{R}^k)$$

$$\varrho(\boldsymbol{a}, \boldsymbol{b}) = \frac{\boldsymbol{a}'\mathcal{C}_{12}\boldsymbol{b}}{\sqrt{(\boldsymbol{a}'\mathcal{C}_{11}\boldsymbol{a})(\boldsymbol{b}'\mathcal{C}_{22}\boldsymbol{b})}} = 0$$

$$\Updownarrow$$

$$\max\left\{\varrho(\boldsymbol{a}, \boldsymbol{b}),\ \boldsymbol{a} \in \mathbb{R}^m,\ \boldsymbol{b} \in \mathbb{R}^k\right\} = 0$$

$$\mathcal{C}_{11}^{-1}\mathcal{C}_{12}\mathcal{C}_{22}^{-1}\mathcal{C}_{12}'$$

$$\begin{array}{ll} u_1 = \boldsymbol{a}_1' \boldsymbol{Y} & v_1 = \boldsymbol{b}_1' \boldsymbol{X} \\ \vdots & \vdots \\ u_s = \boldsymbol{a}_s' \boldsymbol{Y} & v_s = \boldsymbol{b}_s' \boldsymbol{X} \end{array}$$

$$s = \min(m, k)$$

1. zmienne u są parami nieskorelowane

2. zmienne v są parami nieskorelowane

3. współczynniki korelacji ϱ_i między zmiennym u_i oraz v_i dla $i = 1, \ldots, s$ tworzą ciąg malejący

4. zmienne u_i oraz v_j dla $i \neq j$ są nieskorelowane

Zmienne kanoniczne: $u_1, \ldots, u_s, v_1, \ldots, v_s$

Korelacje kanoniczne: $\varrho_1, \ldots, \varrho_s$

$$\begin{bmatrix} 1 & \cdots & 0 & \varrho_1 & \cdots & 0 \\ \cdot & \cdots & \cdot & \cdot & \cdots & \cdot \\ 0 & \cdots & 1 & 0 & \cdots & \varrho_s \\ \varrho_1 & \cdots & 0 & 1 & \cdots & 0 \\ \cdot & \cdots & \cdot & \cdot & \cdots & \cdot \\ 0 & \cdots & \varrho_s & 0 & \cdots & 1 \end{bmatrix}$$

Korelacje kanoniczne

Równanie wyznacznikowe

$$|\mathcal{C}_{12}'\mathcal{C}_{11}^{-1}\mathcal{C}_{12} - \lambda\mathcal{C}_{22}| = 0$$

$\lambda_i,\ i = 1, \ldots, s$ — rozwiązania równania
Korelacje kanoniczne ϱ_i są pierwiastkami z liczb λ_i

$$\varrho_i = \sqrt{\lambda_i},\ i = 1, \ldots, s$$

Zmienne kanoniczne

Wektory $\boldsymbol{a}_i$ oraz $\boldsymbol{b}_i$ tworzące i–tą parę zmiennych kanonicznych są rozwiązaniami układu równań

$$(\mathcal{C}_{12}\mathcal{C}_{22}^{-1}\mathcal{C}_{12}' - \lambda_i\mathcal{C}_{11})\boldsymbol{a}_i = 0$$
$$(\mathcal{C}_{12}\mathcal{C}_{11}^{-1}\mathcal{C}_{12}' - \lambda_i\mathcal{C}_{22})\boldsymbol{b}_i = 0$$

Przykład. Badano zależność wyników testów inteligencji Wechslera (podtest powtarzania cyfr i słownika) dla dorosłych od wieku i ilości lat spędzonych w szkole (lata edukacji).

Y_1 — wynik testu powtarzania cyfr
Y_2 — wynik testu słownika
X_1 — wiek
X_2 — lata edukacji

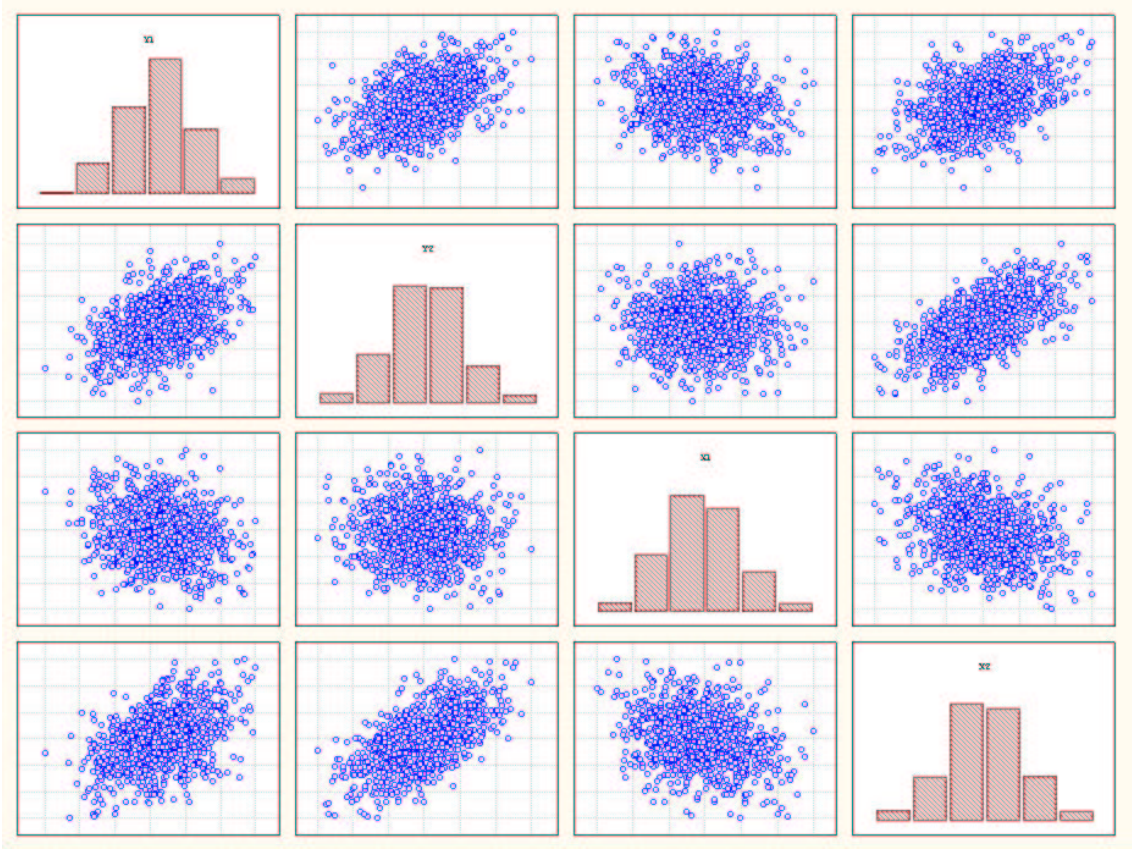
$m = k = 2$

$N = 933$

$$\mathcal{R} = \begin{bmatrix} 1 & 0.45 & -0.19 & 0.43 \\ 0.45 & 1 & -0.02 & 0.62 \\ -0.19 & -0.02 & 1 & -0.29 \\ 0.43 & 0.62 & -0.29 & 1 \end{bmatrix}$$

$$\mathcal{R}_{11} = \begin{bmatrix} 1 & 0.45 \\ 0.45 & 1 \end{bmatrix} \quad \mathcal{R}_{22} = \begin{bmatrix} 1 & -0.29 \\ -0.29 & 1 \end{bmatrix}$$

$$\mathcal{R}_{12} = \begin{bmatrix} -0.19 & 0.43 \\ -0.02 & 0.62 \end{bmatrix}$$



$$|\mathcal{R}'_{12}\mathcal{R}_{11}^{-1}\mathcal{R}_{12} - \lambda\mathcal{R}_{22}| = 0$$

$$\lambda^2 - 0.4667\lambda + 0.0164 = 0$$

$$\lambda_1 = 0.4285 \quad \lambda_2 = 0.0381$$

$$\varrho_1 = \sqrt{\lambda_1} = 0.654 \quad \varrho_2 = \sqrt{\lambda_2} = 0.196$$

Zmienne kanoniczne

$$\begin{bmatrix} 0.1896 & 0.2552 \\ 0.2552 & 0.4123 \end{bmatrix} \begin{bmatrix} a_{11} \\ a_{12} \end{bmatrix} = 0.4285 \begin{bmatrix} 1 & 0.45 \\ 0.45 & 1 \end{bmatrix} \begin{bmatrix} a_{11} \\ a_{12} \end{bmatrix}$$

$$a_{11} = 0.26 \text{ oraz } a_{12} = 1 \text{ tzn. } \boldsymbol{a}_1 = (0.26, 1)'$$

$$\begin{bmatrix} 0.0415 & -0.0467 \\ -0.0467 & 0.4130 \end{bmatrix} \begin{bmatrix} b_{11} \\ b_{12} \end{bmatrix} = 0.4285 \begin{bmatrix} 1 & -0.29 \\ -0.29 & 1 \end{bmatrix} \begin{bmatrix} b_{11} \\ b_{12} \end{bmatrix}$$

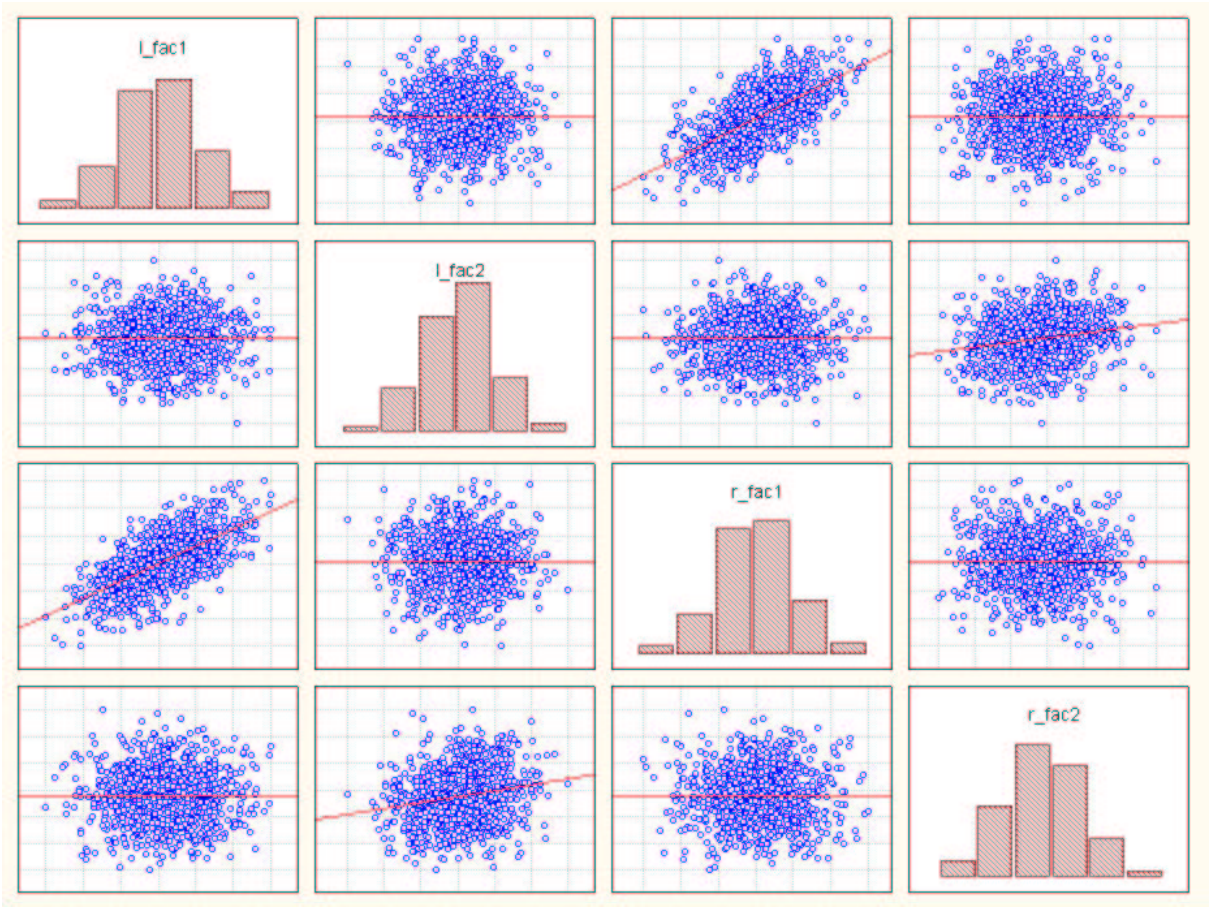
$$b_{11} = 0.20 \text{ oraz } b_{12} = 1 \text{ tzn. } \boldsymbol{b}_1 = (0.20, 1)'$$

Pierwsza para zmiennych kanonicznych:

$$u_1 = 0.26\,(\text{powtarzanie cyfr}) + (\text{słownik})$$
$$v_1 = 0.20\,(\text{wiek}) + (\text{wykształcenie})$$

Druga para zmiennych kanonicznych:

$$u_2 = (\text{powtarzanie cyfr}) - 0.64\,(\text{słownik})$$
$$v_2 = (\text{wiek}) + 0.10\,(\text{wykształcenie})$$


l_fac1
l_fac2
r_fac1
r_fac2

Analiza składowych głównych

Obserwujemy wektor losowy $(X_1, \ldots, X_k)$

Składowe główne:
zmienne losowe $Z_1, \ldots, Z_k$ takie, że

1. są liniowymi funkcjami zmiennych $X_1, \ldots, X_k$

$$Z_i = a_{i1}X_1 + \cdots + a_{ik}X_k$$

gdzie $a_{ij}, i, j = 1, \ldots, k$ są takimi liczbami rzeczywistymi, że

$$\sum_{j=1}^{k} a_{ij}^2 = 1 \quad i = 1, \ldots, k$$

2. są wzajemnie nieskorelowane

$$\varrho_{Z_i Z_j} = 0 \quad i \neq j$$

3. wariancje składowych głównych są malejące

$$D^2 Z_1 \geq D^2 Z_2 \geq \cdots \geq D^2 Z_k$$

Konstrukcja składowych głównych

Niech $\mathcal{C}$ oznacza macierz kowariancji

Niech $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$ będą rozwiązaniami równania

$$|\mathcal{C} - \lambda \boldsymbol{I}| = O$$

gdzie $\boldsymbol{I}$ jest macierzą jednostkową
Liczby λ_i są wartościami własnymi macierzy $\mathcal{C}$

Znajdujemy k wektorów $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k$ takich, że

$$\mathcal{C}\boldsymbol{a}_i = \lambda_i \boldsymbol{a}_i \quad i = 1, \ldots, k$$

Wektory $\boldsymbol{a}_i$ są wektorami własnymi macierzy $\mathcal{C}$

Niech

$$\boldsymbol{a}_i = (a_{i1}, \ldots, a_{ik})' \quad i = 1, \ldots, k$$

Składowe główne Z_i

$$Z_i = a_{i1}X_1 + \cdots + a_{ik}X_k \quad i = 1, \ldots, k$$

Ilość składowych głównych

Kryterium wyjaśnianej zmienności

Całkowite zróżnicowanie obiektów

$$\sum_{i=1}^{k} D^2 Z_i$$

Składowa Z_i wyjaśnia

$$\frac{D^2 Z_i}{\sum_{i=1}^{k} D^2 Z_i} \cdot 100\%$$

całkowitej zmienności obiektów

Wybór ilości l składowych głównych:
procent zmienności wyjaśnianej

$$\frac{\sum_{i=1}^{l} D^2 Z_i}{\sum_{i=1}^{k} D^2 Z_i} \cdot 100\%$$

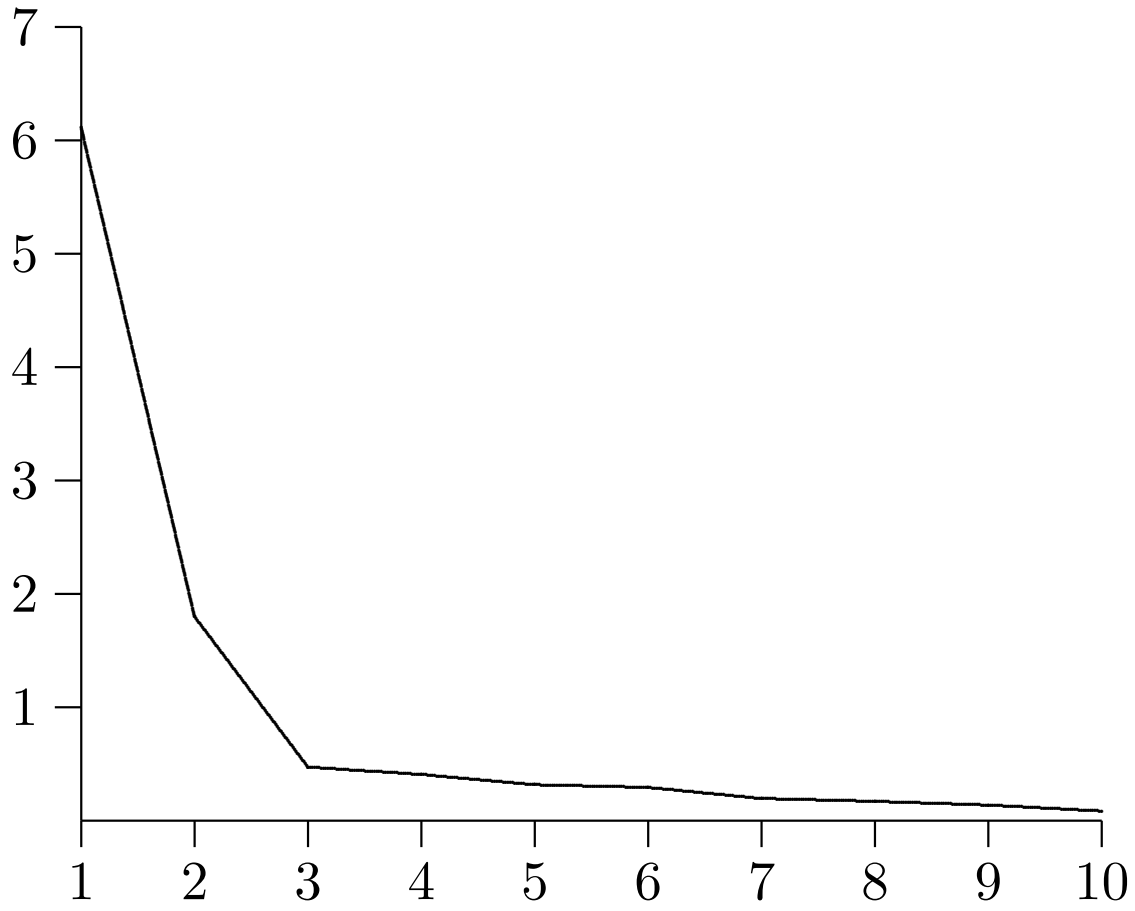
jest „dostatecznie” duży ($\geq 75\%$).

Ilość składowych głównych

Kryterium osypiska

Składowej Z_i odpowiada wartość własna λ_i macierzy kowariancji $\mathcal{R}$, przy czym

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$$



Przykład
Obserwowano trzy wymiary skorupy żółwi:
X_1 — długość
X_2 — szerokość
X_3 — wysokość

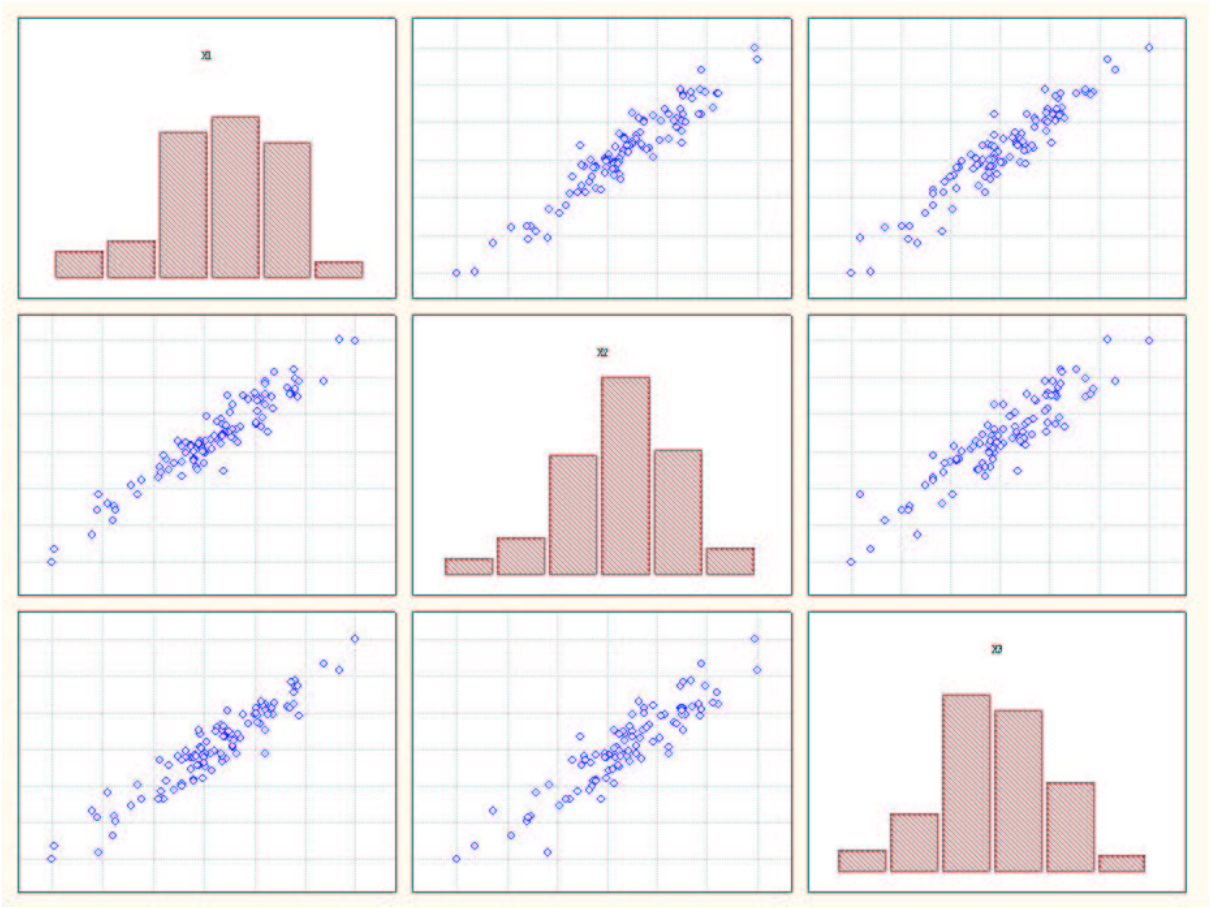
Macierz kowariancji

$$\mathcal{C} = \begin{bmatrix} 138.7663 & 79.1467 & 37.3750 \\ 79.1467 & 50.0417 & 21.6540 \\ 37.3750 & 21.6540 & 11.2591 \end{bmatrix}$$

$$\lambda_1 = 195.275 \quad \lambda_2 = 3.689 \quad \lambda_3 = 1.104$$

Wektory własne

$$\begin{aligned} \boldsymbol{a}_1 &= (0.8401, \quad 0.4919, \quad 0.2285)' \\ \boldsymbol{a}_2 &= (0.4881, -0.8694, \quad 0.0770)' \\ \boldsymbol{a}_3 &= (0.2365, \quad 0.0469, -0.9705)' \end{aligned}$$


X1
X2
X3

Składowe główne

$$Z_1 = +0.8401 \cdot X_1 + 0.4919 \cdot X_2 + 0.2285 \cdot X_3$$
$$Z_2 = +0.4881 \cdot X_1 - 0.8694 \cdot X_2 + 0.0770 \cdot X_3$$
$$Z_3 = +0.2350 \cdot X_1 + 0.0469 \cdot X_2 - 0.9705 \cdot X_3$$

Wariancje

Składowa	Wariancja	Procent zmienności
Z_1	195.275	97.60
Z_2	3.689	1.84
Z_3	1.104	0.56

Współczynniki korelacji składowych głównych z oryginalnymi zmiennymi

Zmienna	Z_1	Z_2	Z_3
Długość	0.9966	0.0796	0.0211
Szerokość	0.9717	−0.2360	0.0070
Wysokość	0.9517	0.0441	−0.3039