Analiza składowych głównych-wersja populacyjna

Niech $\mathbf{X}$ będzie p- wymiarowym wektorem losowym o wektorze wartości oczekiwanych $\mathbf{m}$ i macierzy kowariancji $\mathbf{\Sigma}$. Z uwagi na fakt, że interesować będą nas tylko wariancje i kowariancje możemy przyjąć, że $\mathbf{m=0}$ lub przyjąć że pracujemy na zmiennych scentrowanych $\mathbf{X}-\mathbf{m}$. Przyjmujemy że $\mathbf{m=0}$. Celem analizy jest określenie kolejnych tzw. składowych głównych wektora losowego $\mathbf{X}$. Interesować nas będą kombinacje liniowe wektora losowego $\mathbf{X}$, czyli iloczyny skalarne $\mathbf{a}^T\mathbf{X}=\langle \mathbf{a},\mathbf{X}\rangle$, gdzie $\mathbf{a}$ jest dowolnym ustalonym wektorem w przestrzeni R^p. Będziemy zakładać, że wektor $\mathbf{a}$ jest jednostkowy tzn. $\|\mathbf{a}\|=\langle \mathbf{a},\mathbf{a}\rangle=1$ a kombinację liniową $\mathbf{a}^T\mathbf{X}=\langle \mathbf{a},\mathbf{X}\rangle$ będziemy nazywać standaryzowaną kombinacją liniową. Pierwsza składowa główna powstaje przez znalezienie takiego jednostkowego wektora $\mathbf{f}_1 \in R^p$, że

$$V(\langle \mathbf{f}_1,\mathbf{X}\rangle)=V(\mathbf{f}_1^T\mathbf{X})=\max_{\|\mathbf{a}\|=1} V(\mathbf{a}^T\mathbf{X}).$$

Poszukujemy zatem standaryzowanej kombinacji liniowej o największej wariancji.

Ale $V(\mathbf{a}^T\mathbf{X})=\mathbf{a}^T\mathbf{\Sigma}\mathbf{a}=\langle \mathbf{a},\mathbf{\Sigma}\mathbf{a}\rangle$. Rozwiązujemy więc zagadnienie wyznaczenia maksimum formy kwadratowej $\langle \mathbf{a},\mathbf{\Sigma}\mathbf{a}\rangle$ na sferze jednostkowej $\|\mathbf{a}\|=1$. Sfera w R^p jest zbiorem zwartym a forma kwadratowa jest funkcją ciągłą, więc zadanie ma rozwiązanie. Tworzymy więc funkcję Lagrange'a

$$L(\mathbf{a},\lambda)=\langle \mathbf{a},\mathbf{\Sigma}\mathbf{a}\rangle+\lambda(1-\langle \mathbf{a},\mathbf{a}\rangle)$$

Wyznaczając pochodne cząstkowe

$$\begin{cases} \dfrac{\partial}{\partial a_1}L(\mathbf{a},\lambda)=0 \\ \vdots \\ \dfrac{\partial}{\partial a_p}L(\mathbf{a},\lambda)=0 \\ \dfrac{\partial}{\partial \lambda}L(\mathbf{a},\lambda)=0 \end{cases} \quad \text{otrzymujemy układ} \quad \begin{cases} (\mathbf{\Sigma}-\lambda\mathbf{I})\mathbf{a}=\mathbf{0} \\ \|\mathbf{a}\|=1 \end{cases}.$$

Poszukiwany wektor $\mathbf{a}$ jest więc jednostkowym wektorem własnym macierzy kowariancji $\mathbf{\Sigma}$. $V(\mathbf{a}^T\mathbf{X})=\mathbf{a}^T\mathbf{\Sigma}\mathbf{a}=\langle \mathbf{a},\mathbf{\Sigma}\mathbf{a}\rangle=\lambda\langle \mathbf{a},\mathbf{a}\rangle=\lambda$. Maksymalną wariancję otrzymamy wybierając jednostkowy wektor własny $\mathbf{f}_1$ odpowiadający maksymalnej wartości własnej λ_1 macierzy kowariancji $\mathbf{\Sigma}$. Pierwszą składową główną jest wiec zmienna losowa $F_1=\mathbf{f}_1^T\mathbf{X}=\langle \mathbf{f}_1,\mathbf{X}\rangle$ a jej

wariancja jest równa $V(F_1) = V(\mathbf{f}_1^T\mathbf{X}) = \mathbf{f}_1^T\boldsymbol{\Sigma}\mathbf{f}_1 = \langle \mathbf{f}_1, \boldsymbol{\Sigma}\mathbf{f}_1 \rangle = \lambda \langle \mathbf{f}_1, \mathbf{f}_1 \rangle = \lambda_{\max}$ maksymalnej wartości własnej.

Druga składowa główna $F_2 = \langle \mathbf{f}_2, \mathbf{X} \rangle$ powstaje przez znalezienie takiego jednostkowego wektora $\mathbf{f}_2 \in R^p$, ortogonalnego do znalezionego $\mathbf{f}_1$ (więc cov(F_1,F_2), że

$$V(\langle \mathbf{f}_2, \mathbf{X} \rangle) = V(\mathbf{f}_2^T\mathbf{X}) = \max_{\|\mathbf{a}\|=1, \langle \mathbf{a}, \mathbf{f}_1 \rangle = 0} V(\mathbf{a}^T\mathbf{X}).$$

Tworzymy funkcję Lagrange'a

$$L(\mathbf{a}, \lambda, \mu) = \langle \mathbf{a}, \boldsymbol{\Sigma}\mathbf{a} \rangle + \lambda(1 - \langle \mathbf{a}, \mathbf{a} \rangle + \mu \langle \mathbf{f}_1, \mathbf{a} \rangle$$

i z warunku koniecznego istnienia ekstremum

$$\begin{cases} \frac{\partial}{\partial a_1} L(\mathbf{a}, \lambda, \mu) = 0 \\ \vdots \\ \frac{\partial}{\partial a_p} L(\mathbf{a}, \lambda, \mu) = 0 \\ \frac{\partial}{\partial \lambda} L(\mathbf{a}, \lambda, \mu) = 0 \\ \frac{\partial}{\partial \mu} L(\mathbf{a}, \lambda, \mu) = 0 \end{cases} \text{ otrzymujemy układ równań } \begin{cases} (\boldsymbol{\Sigma} - \lambda \mathbf{I})\mathbf{a} + \mu \mathbf{f}_1 = \mathbf{0} \\ \|\mathbf{a}\| = 1 \\ \langle \mathbf{f}_1, \mathbf{a} \rangle = 0 \end{cases}$$

Stąd mnożąc skalarnie pierwsze wektorowe równanie przez $\mathbf{a}$ otrzymujemy $\langle (\boldsymbol{\Sigma} - \lambda \mathbf{I})\mathbf{a} + \mu \mathbf{f}_1, \mathbf{a} \rangle = 0$ $\Leftrightarrow \langle (\boldsymbol{\Sigma} - \lambda \mathbf{I})\mathbf{a}, \mathbf{a} \rangle + \mu \langle \mathbf{f}_1, \mathbf{a} \rangle = 0 \Leftrightarrow \langle (\boldsymbol{\Sigma} - \lambda \mathbf{I})\mathbf{a}, \mathbf{a} \rangle = 0 \Leftrightarrow (\boldsymbol{\Sigma} - \lambda \mathbf{I})\mathbf{a} = \mathbf{0}, \|\mathbf{a}\| = 1, \mu = 0$.

Druga składowa główną jest wiec zmienną losową $\mathbf{f}_2^T\mathbf{X} = \langle \mathbf{f}_2, \mathbf{X} \rangle$, gdzie $\mathbf{f}_2$ jest jednostkowym wektorem własnym macierzy kowariancji $\boldsymbol{\Sigma}$ ortogonalnym do $\mathbf{f}_1$ a jej wariancja

$V(\langle \mathbf{f}_2, \mathbf{X} \rangle) = V(\mathbf{f}_2^T\mathbf{X}) = \langle \boldsymbol{\Sigma}\mathbf{f}_2, \mathbf{f}_2 \rangle = \langle \lambda \mathbf{f}_2, \mathbf{f}_2 \rangle = \lambda$ będzie maksymalna, jeśli wybierzemy wektor własny odpowiadający drugiej co do wielkości wartości własnej macierzy kowariancji.

Uwaga. Wiadomo że wektory własne odpowiadające różnym wartościom własnym są liniowo niezależne. W przypadku macierzy symetrycznej wektory własne odpowiadające różnym wartościom własnym są ortogonalne. Jeżeli wartość własna macierzy symetrycznej jest wielokrotna , to odpowiada jej podprzestrzeń własna o wymiarze równym krotności tej wartości własnej. Stosując procedurę ortogonalizacji można w tej podprzestrzeni wybrać bazę ortonormalną.

Podsumowując. Niech $\mathbf{X}$ będzie *p*- wymiarowym wektorem losowym o wektorze wartości oczekiwanych **m** i macierzy kowariancji $\mathbf{\Sigma}$ i niech wartości własne macierzy kowariancji $\mathbf{\Sigma}$ spełniają warunek $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$. Wektor $\mathbf{f}_i$ odpowiadający *i*-tej składowej głównej $F_i = \langle \mathbf{f}_i, \mathbf{X} \rangle$ jest równy *i*-temu jednostkowemu wektorowi własnemu macierzy $\mathbf{\Sigma}$ odpowiadającemu wartości własnej λ_i. Składowe główne $F_1,\ldots, F_p$ są nieskorelowane

$\operatorname{cov}(F_i, F_j) = \operatorname{cov}(\mathbf{f}_i^T \mathbf{X}\mathbf{X}^T \mathbf{f}_j) = \mathbf{f}_i^T \mathbf{\Sigma} \mathbf{f}_j = \lambda_j \mathbf{f}_i^T \mathbf{f}_j = 0$ a ich wariancje są równe

$V(F_i) = \operatorname{cov}(\mathbf{f}_i^T \mathbf{X}\mathbf{X}^T \mathbf{f}_i) = \mathbf{f}_i^T \mathbf{\Sigma} \mathbf{f}_i = \lambda_i \mathbf{f}_i^T \mathbf{f}_i = \lambda_i$, *i*=1,…,*p*. Wszystkie składowe główne tworzą nowy wektor $\mathbf{F} = \begin{bmatrix} F_1 \\ \vdots \\ F_p \end{bmatrix} = \begin{bmatrix} \mathbf{f}_1^T \\ \vdots \\ \mathbf{f}_p^T \end{bmatrix} \begin{bmatrix} \\ \mathbf{X} \\ \\ \end{bmatrix}$ o macierzy kowariancyjnej $\mathbf{\Lambda} = \operatorname{diag}\{\lambda_1,\ldots,\lambda_p\}$.

Niech $\mathbf{\Gamma} = \begin{bmatrix} \mathbf{f}_1 & \cdots & \mathbf{f}_p \end{bmatrix}$. Wektor $\mathbf{X}$ o macierzy kowariancji $\mathbf{\Sigma}$ został przekształcony przez transformację ortogonalną w nowy wektor $\mathbf{F} = \mathbf{\Gamma}^T \mathbf{X}$ ($\mathbf{F} = \mathbf{\Gamma}^T (\mathbf{X} - \mathbf{m})$ o diagonalnej macierzy kowariancji $\mathbf{\Lambda} = \operatorname{diag}\{\lambda_1,\ldots,\lambda_p\}$. Transformacja ortogonalna $\mathbf{F} = \mathbf{\Gamma}^T \mathbf{X}$ implikuje przekształcenie

$$\mathbf{\Gamma}^T \mathbf{\Sigma} \mathbf{\Gamma} = \mathbf{\Lambda}$$

macierzy kowariancyjnej z zachowaniem zarówno wyznacznika macierzy kowariancyjnej jak i jej śladu.

Stąd $$\frac{Var(F_i)}{\sum_{i=1}^{p} Var(F_i)} 100\% = \frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i} 100\%$$

informuje nas jaki procent zmienności oryginalnej zmiennej wyjaśniają poszczególne składowe główne.

Zauważmy, że $\left[\operatorname{cov}(\mathbf{X}, F_i)\right]_{(p,1)} = \left[\operatorname{cov}(\mathbf{X}, \mathbf{f}_i^T \mathbf{X})\right]_{(p,1)} = \left[\operatorname{cov}(\mathbf{X}, \mathbf{X}^T \mathbf{f}_i)\right]_{(p,1)} = \left[\mathbf{\Sigma} \mathbf{f}_i\right]_{(p,1)} = \left[\lambda_i \mathbf{f}_i\right]_{(p,1)}$

Dla standaryzowanego wektora $\mathbf{X}$

$$\left[\operatorname{cor}(\mathbf{X}, F_i)\right]_{(p,1)} = \frac{1}{\sqrt{V(F_i)}} \left[\operatorname{cov}(\mathbf{X}, \mathbf{f}_i^T \mathbf{X})\right]_{(p,1)} = \frac{1}{\sqrt{\lambda_i}} \left[\operatorname{cov}(\mathbf{X}, \mathbf{X}^T \mathbf{f}_i)\right]_{(p,1)} = \left[\sqrt{\lambda_i} \mathbf{f}_i\right]_{(p,1)}$$

Analiza składowych głównych-wersja próbkowa

Załóżmy, że dysponujemy próbą prostą z p-wymiarowego rozkładu prawdopodobieństwa. Możemy wyznaczyć wektor $\bar{\mathbf{x}}$ i próbkową macierz kowariancji $\mathbf{S}$. Jeśli oryginalne zmienne są mierzone w różnych jednostkach zwykle standaryzujemy na wstępie nasze dane i pracować będziemy z próbkową macierzą korelacyjną. Jeśli dane są jednorodne (w tych samych jednostkach) tak, że ich liniowe kombinacje mają sens, nie jest konieczna standaryzacja tylko centrowanie zmiennych. Całą analizę możemy powtórzyć wykorzystując próbkową macierz kowariancji (lub korelacji) $\mathbf{S}$ (lub $\mathbf{R}$)

Przykład

W stacji hodowli roślin IHAR w Borowie przeprowadzono doświadczenie, którego celem było zbadanie różnic pomiędzy 11 odmianami słonecznika. Był to jeden z wielu eksperymentów mających na celu wyhodowanie odmiany o dużej wartości plonu niełupek i maksymalnej zawartości tłuszczu, przy minimalnej wysokości rośliny i małej procentowej zawartości łupiny w niełupkach. Doświadczenie założono w układzie blokowym z sześcioma powtórzeniami. Zbiór danych Słonecznik zawiera średnie arytmetyczne z sześciu poletek obliczone dla 6 następujących zmiennych:

- Wysokość roślin w cm (**Wysokość**)
- Średnica koszyczka w cm (**Średnica**)
- Plon niełupek w kg z poletka o pow. 9 m^2 (**Niełupki**)
- Procentowa zawartość oleju w nasionach (**Olej**)
- Procentowa zawartość łupiny w niełupkach (**Łupiny**)
- Liczba roślin (**Liczba**)

	1 WYSOKOŚĆ	2 ŚREDNICA	3 NIEŁUPKI	4 OLEJ	5 ŁUPINY	6 LICZBA
1	84,18	21,02	1,29	53,40	30,33	19,17
2	113,40	19,12	1,20	50,70	27,08	19,17
3	110,98	19,60	0,97	47,87	28,00	15,50
4	107,80	18,85	0,97	48,83	25,83	17,17
5	112,67	19,18	1,18	49,53	27,17	20,17
6	85,40	22,28	1,26	52,80	29,58	17,67
7	130,35	25,47	1,27	51,70	24,67	16,17
8	110,75	19,12	1,20	49,80	29,08	20,17
9	106,07	18,32	1,04	48,05	27,58	19,67
10	82,55	21,58	1,16	53,55	30,50	17,17
11	91,50	18,25	0,93	43,70	43,42	18,50

Metoda głównych składowych jako metoda redukcji danych

Cel ćwiczenia: Graficzne przedstawienie wyników pomiarów 6 zmiennych dla 11 rodów słonecznika przy użyciu dwóch pierwszych składowych głównych i ocena stopnia podobieństwa pomiędzy 11 odmianami słonecznika.

Model. Przyjmujemy, że zaobserwowane zmienne losowe $X_1,...,X_6$ są skorelowane, a problem polega na określeniu sześciu nowych zmiennych $F_1,...,F_6$, zwanych składowymi głównymi o następujących własnościach:

- każda składowa główna $F_1,...,F_6$ jest kombinacją liniową zmiennych $X_1,...,X_6$ tzn. $F_i = \sum_{j=1}^{6} f_{ij} X_j$, i=1,...,6 o współczynnikach f_{ij} spełniających warunek $\sum_{j=1}^{6} f_{ij}^2 = 1$,
- składowe $F_1,...,F_6$ są nieskorelowane,
- pierwsza składowa F_1 ma największą wariancję spośród wszystkich kombinacji liniowych (o unormowanych współczynnikach) zmiennych $X_1,...,X_6$, druga składowa F_2 ma największą wariancję spośród wszystkich kombinacji liniowych zmiennych $X_1,...,X_6$ nieskorelowanych z F_1, trzecia składowa F_3 ma największą wariancję spośród wszystkich kombinacji liniowych zmiennych $X_1,...,X_6$ nieskorelowanych z F_1 i F_2 itd.

Tak określone składowe posiadają tę własność, że najwięcej informacji o badanych obiektach (najwięcej zmienności) zawarte jest w F_1 a najmniej w F_6. Dla składowej F_i ten procentowy udział wyraża się wzorem

$$\frac{Var(F_i)}{\sum_{i=1}^{6} Var(F_i)} 100\%.$$

W zastosowaniach praktycznych, ograniczamy się zwykle do dwóch (ewentualnie trzech) pierwszych składowych głównych. Zwykle też w celu ułatwienia interpretacji uzyskiwanych składowych analizę prowadzi się po standaryzacji danych, co prowadzi do danych bezwymiarowych unormowanych (przeskalowanych).

Łatwo pokazać, że problem wyznaczania głównych składowych sprowadza się do problemu wyznaczania ekstremum formy kwadratowej na sferze jednostkowej, której macierzą jest macierz kowariancji (korelacji), a ten problem redukuje się do rozwiązania problemu własnego dla macierzy kowariancji (korelacji). Niech $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_6$ będzie uporządkowanym malejąco ciągiem wartości własnych. Wektor współczynników wagowych $f_i=(f_{i1},...,f_{i6})^T$ i-tej składowej głównej F_i jest unormowanym wektorem własnym odpowiadającym i-tej wartości własnej λ_i macierzy kowariancji (korelacji) zmiennych $X_1,...,X_6$. Wariancja składowej głównej F_i jest równa λ_i.

Zadanie.

1. Dla powyższego zbioru danych **wykonaj standaryzację** (Statistica "sama" standaryzuje dane") oblicz:
   - udział składowych głównych w zmienności całkowitej (wartości własne)

Nr wartości	Wartości własne macierzy korelacji i pokrewne Tylko zmienne aktywne			
	Wartość wł	% ogółu Warianc.	Skumul. Wartość wł	Skumul. %
1	2,779148	46,31914	2,779148	46,3191
2	1,473436	24,55727	4,252585	70,8764
3	1,208280	20,13799	5,460864	91,0144
4	0,483347	8,05578	5,944211	99,0702
5	0,041144	0,68573	5,985355	99,7559
6	0,014645	0,24409	6,000000	100,0000

- udział poszczególnych zmiennych w tworzeniu składowych głównych (wektory własne)

Zmienna	Wektory własne macierzy korelacji(Słonecznik) Tylko zmienne aktywne					
	Czynn. 1	Czynn. 2	Czynn. 3	Czynn. 4	Czynn. 5	Czynn. 6
WYSOKOŚĆ	-0,097747	0,684153	-0,404753	-0,409215	0,281272	0,334640
ŚREDNICA	-0,502529	0,076056	0,369965	-0,474450	-0,611154	0,078704
NIEŁUPKI	-0,507557	-0,328867	-0,216432	-0,340423	0,442006	-0,525484
OLEJ	-0,540999	-0,284655	0,020571	0,345678	0,239795	0,669980
ŁUPINY	0,409191	-0,373038	0,366237	-0,577778	0,331175	0,340253
LICZBA	0,142052	-0,444759	-0,719657	-0,195529	-0,429919	0,202593

Zmienna	Korel. czynniki-zmienne (ładunki czynn.)na podst. korelacji (Słonecznik)					
	Czynn. 1	Czynn. 2	Czynn. 3	Czynn. 4	Czynn. 5	Czynn. 6
WYSOKOŚĆ	-0,162952	0,830461	-0,444912	-0,284499	0,057053	0,040498
ŚREDNICA	-0,837755	0,092321	0,406672	-0,329852	-0,123966	0,009525
NIEŁUPKI	-0,846137	-0,399196	-0,237905	-0,236673	0,089656	-0,063593
OLEJ	-0,901888	-0,345529	0,022611	0,240326	0,048640	0,081080
ŁUPINY	0,682153	-0,452813	0,402574	-0,401689	0,067175	0,041177
LICZBA	0,236812	-0,539871	-0,791059	-0,135938	-0,087204	0,024517

Uwaga: Statgraf podaje wektory $f_i=(f_{i1},...,f_{i6})^T$ a Statistica wektory

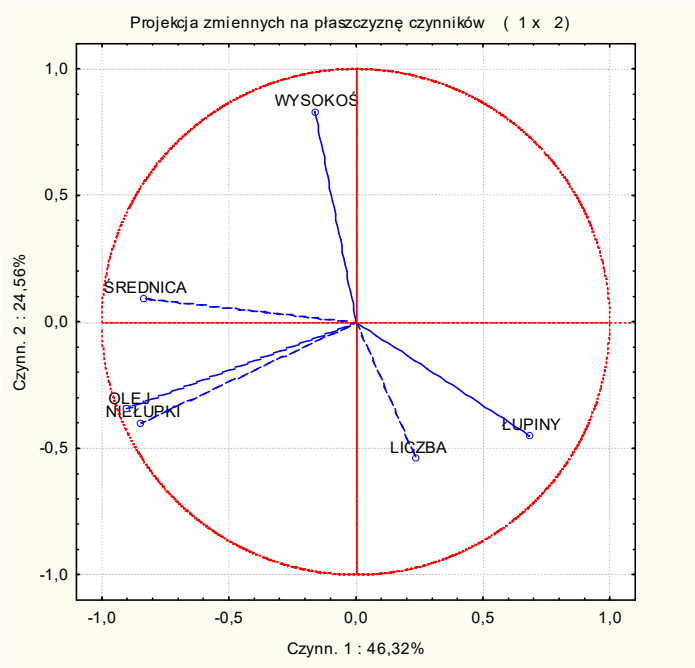
$\hat{f}_i = \frac{1}{\sqrt{\lambda_i}} f_i$ (tzw. **wartości współczynników czynnikowych**).

$\hat{\hat{f}}_i = \sqrt{\lambda_i} f_i$ (tzw. **ładunki czynnikowe - współrzędne czynnikowe**)

Statistica podaje jeszcze tzw. **zasoby zmienności wspólnej** ($\cos^2$) $w_i = \lambda_i f_i^2$ gdzie f_i^2 jest wektorem, którego współrzędne są kwadratami współrzędnych wektora f_i.

- **składowe główne** F_i (Statistica podaje standaryzowane $\hat{F}_i = \frac{1}{\sqrt{\lambda_i}} F_i$)

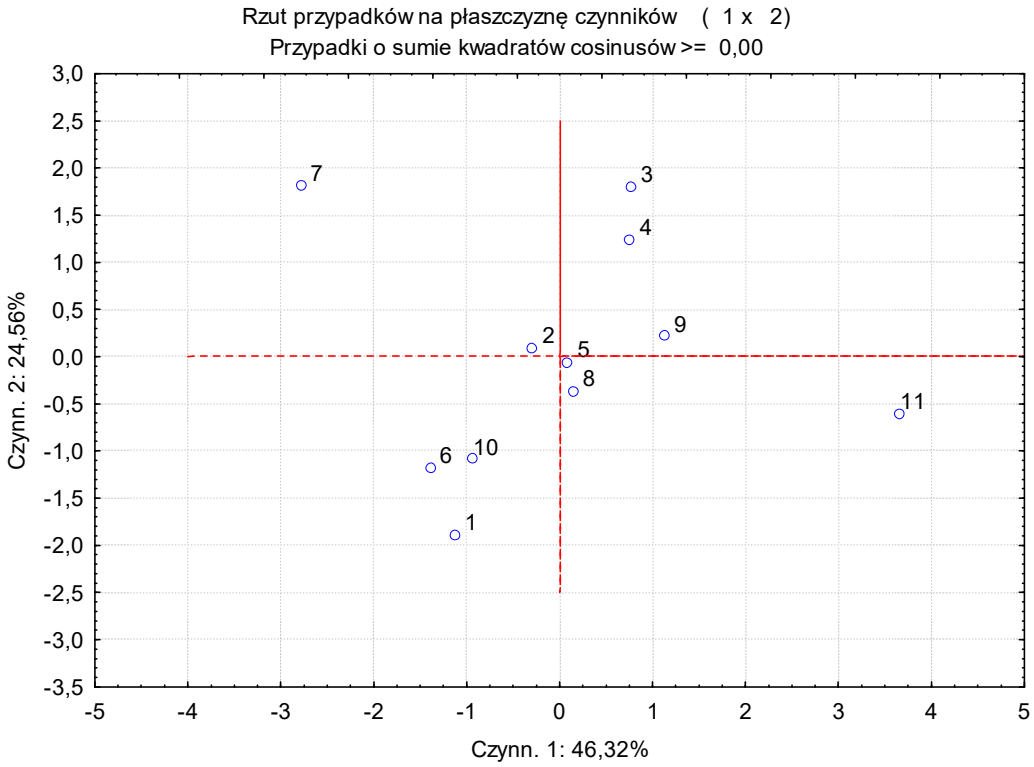
2. Przedstaw graficznie
   - udział zmiennych $X_1,...,X_6$ w tworzeniu dwóch pierwszych składowych głównych F_1 i F_2



Wnioski. Na Czynnik 1 największy wpływ (tu ujemny) mają zmienne Olej, Niełupki, Średnica a dodatni (Łupiny, Liczba).Duża ujemna wartość czynnika I świadczy o wysokiej wartości hodowlanej Cynnik II jest związany z wysokością rośliny . Duża dodatnia wartość czynnika 2 oznacza dużą wysokość rośliny.

- rozmieszczenie 11 rodów słonecznika w układzie dwóch pierwszych składowych głównych F_1 i F_2 .

	Współrzędne czynnikowe przypadków, na podst. korelacji (Słonec					
Przyp	Czynn. 1	Czynn. 2	Czynn. 3	Czynn. 4	Czynn. 5	Czynn. 6
1	-1,12644	-1,88265	0,05166	0,12223	0,047068	-0,047497
2	-0,29554	0,09399	-1,15713	0,03843	0,380604	0,042304
3	0,76836	1,80045	1,06056	0,59756	0,244204	-0,129810
4	0,75376	1,23646	0,11577	1,00731	-0,117092	0,057040
5	0,08521	-0,05687	-1,54535	-0,17453	-0,075051	-0,028293
6	-1,39245	-1,17846	0,89637	0,09118	-0,080767	-0,234356
7	-2,78568	1,81730	0,53235	-1,21600	-0,125063	0,051900
8	0,14088	-0,36400	-1,39534	-0,35036	0,122073	0,039920
9	1,12762	0,21781	-1,04240	0,38890	-0,384803	-0,022647
10	-0,93662	-1,08510	1,31439	0,62129	-0,015071	0,247392
11	3,66087	-0,59894	1,16911	-1,12601	0,003898	0,024047



Wnioski Analizując ostatni rysunek można zauważyć podział rodów na 2 grupy. W skład pierwszej grupy wchodzą rody o numerach 1, 6, 10. Jest to grupa która charakteryzuje się dużą zawartością oleju w nasionach stosunkowo dużą wartością plonu niełupek (zobacz wagi f_1 dla zmiennej F_1) i małą wysokością roślin (zobacz wagi f_2 dla zmiennej F_2). Zgodnie z przyjętym wcześniej kryterium są to rody o dużej wartości hodowlanej. Do grupy drugiej należą rody o numerach 2, 3, 4, 5, 8 i 9. Są to rody o niższej wartości hodowlanej. Rody 7 i 11 wyraźnie odbiegają od pozostałych odmian: 7-ze względu na dużą wysokość przy dużej zawartości oleju natomiast 11- ze względu na małą wysokość i małą zawartość oleju.

W programie R dostępnych jest kilka funkcji realizujących PCA

- prcomp() i princcomp() [w pakiecie *stats*]
- PCA() [pakiet FactoMineR]
- dudi.pca [pakiet ade4]
- epPCA() [pakiet ExPosition]

```
# Plik Słonecznik.R
library(openxlsx)
dane<-read.xlsx("C:/Users/User/Documents/Dydaktyka/Modele
liniowe/Ćwiczenia/Słonecznik.xlsx",colNames = TRUE)
colnames(dane)<-c("Wysokość","Średnica","Niełupki","Olej","Łupiny","Liczba")
library(FactoMineR)
library(factoextra)
```

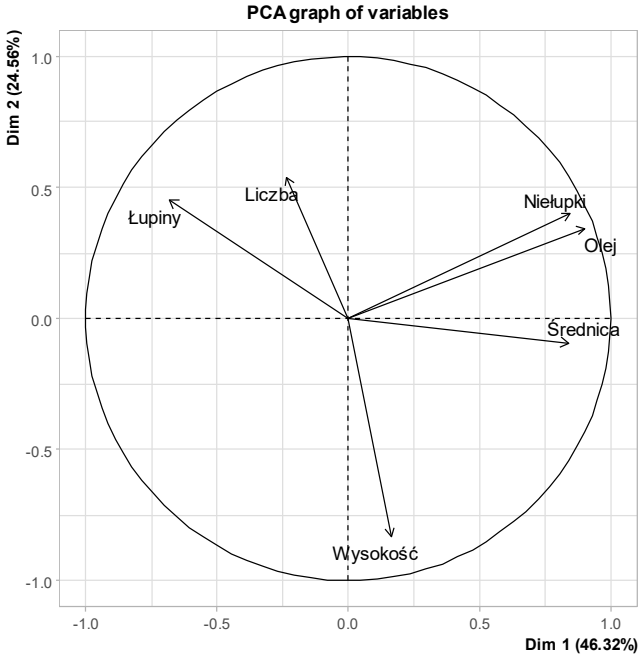
```
#sprawdzenie SVD  X=UDV'
wyn<-svd(as.matrix(dane),nu=6,nv=6)
wyn$u%*%diag(wyn$d)%*%t(wyn$v)
wyn$d
wyn$v #macierz wektorów własnych macierzy X'X

#wyznaczymy wartości własne i wektory własne macierzy X'X
sp<-eigen(t(as.matrix(dane))%*%as.matrix(dane))
#wartości osobliwe zestawione w D z (SVD) są pierwiastkami wartości własnych X'X z procedury eigen
sp
sqrt(sp$values)
sp$vectors


res.pca<-PCA(dane,scale.unit = TRUE,ncp=6,graph=TRUE)
print(res.pca)
```

Results for the Principal Component Analysis (PCA)
The analysis was performed on 11 individuals, described by 6 variables
*The results are available in the following objects:

```
   name              description
1  "$eig"            "eigenvalues"
2  "$var"            "results for the variables"
3  "$var$coord"      "coord. for the variables"
4  "$var$cor"        "correlations variables - dimensions"
5  "$var$cos2"       "cos2 for the variables"
6  "$var$contrib"    "contributions of the variables"
7  "$ind"            "results for the individuals"
8  "$ind$coord"      "coord. for the individuals"
9  "$ind$cos2"       "cos2 for the individuals"
10 "$ind$contrib"    "contributions of the individuals"
11 "$call"           "summary statistics"
12 "$call$centre"    "mean of the variables"
13 "$call$ecart.type" "standard error of the variables"
14 "$call$row.w"     "weights for the individuals"
15 "$call$col.w"     "weights for the variables"
```



$coord współrzędne czynnikowe $\sqrt{\lambda_i}\mathbf{f}_i$ (ładunki czynnikowe)

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6
Wysokość	0.1629518	-0.83046094	0.44491169	0.2844993	0.05705276	0.040497536
Średnica	0.8377552	-0.09232127	-0.40667214	0.3298524	-0.12396564	0.009524624
Niełupki	0.8461367	0.39919606	0.23790538	0.2366726	0.08965589	-0.063593137
Olej	0.9018880	0.34552892	-0.02261147	-0.2403264	0.04863971	0.081079831
Łupiny	-0.6821527	0.45281254	-0.40257432	0.4016895	0.06717510	0.041176798
Liczba	-0.2368117	0.53987114	0.79105919	0.1359377	-0.08720415	0.024517479

$cor Korelacje pomiędzy czynnikami a składowymi głównymi

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6
Wysokość	0.1629518	-0.83046094	0.44491169	0.2844993	0.05705276	0.040497536
Średnica	0.8377552	-0.09232127	-0.40667214	0.3298524	-0.12396564	0.009524624
Niełupki	0.8461367	0.39919606	0.23790538	0.2366726	0.08965589	-0.063593137
Olej	0.9018880	0.34552892	-0.02261147	-0.2403264	0.04863971	0.081079831
Łupiny	-0.6821527	0.45281254	-0.40257432	0.4016895	0.06717510	0.041176798
Liczba	-0.2368117	0.53987114	0.79105919	0.1359377	-0.08720415	0.02451747

Ponieważ pracujemy ma macierzy korelacyjnej współrzędne czynnikowe są równe współczynnikom korelacji

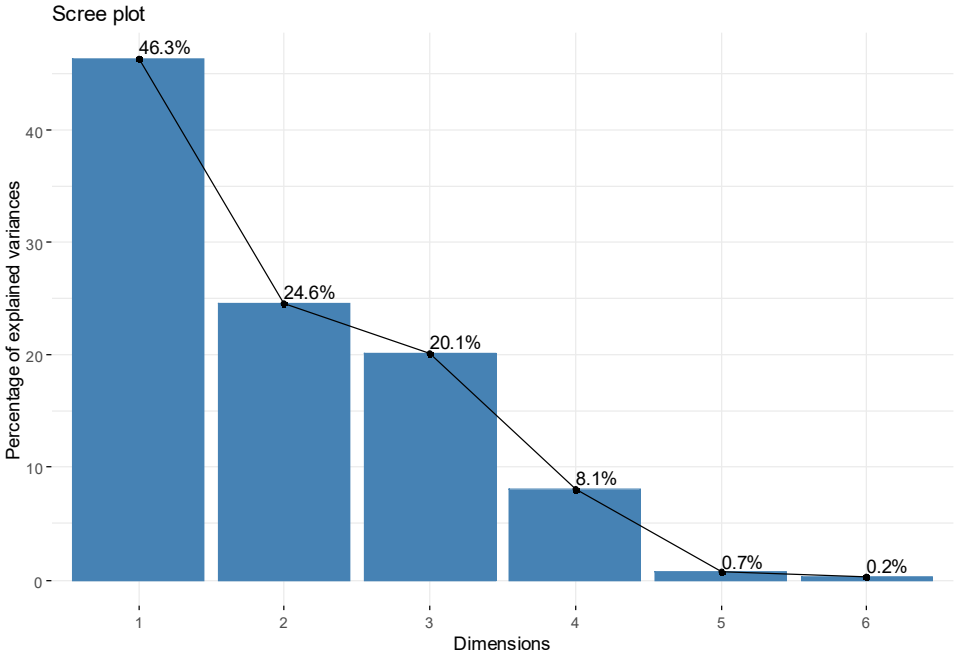
$cos2 zasoby zmienności wspólnej $(\sqrt{\lambda_i}\mathbf{f}_i)^2$ kwadraty powyższych kolumn .

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6
Wysokość	0.02655329	0.689665377	0.1979464124	0.08093985	0.003255018	1.640050e-03
Średnica	0.70183376	0.008523216	0.1653822259	0.10880260	0.015367479	9.071846e-05
Niełupki	0.71594733	0.159357495	0.0565989716	0.05601394	0.008038179	4.044087e-03
Olej	0.81340197	0.119390237	0.0005112785	0.05775676	0.002365821	6.573939e-03
Łupiny	0.46533226	0.205039199	0.1620660857	0.16135443	0.004512494	1.695529e-03
Liczba	0.05607979	0.291460844	0.6257746395	0.01847905	0.007604564	6.011068e-04

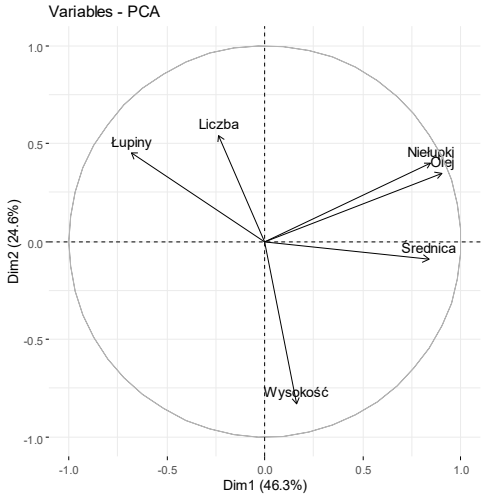
res.pcavarcontrib

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6
Wysokość	0.9554471	46.8065939	16.38250039	16.745715	7.911368	11.1983762
Średnica	25.2535546	0.5784584	13.68741341	22.510264	37.350878	0.6194318
Niełupki	25.7613926	10.8153632	4.68426107	11.588772	19.536908	27.6133030
Olej	29.2680293	8.1028431	0.04231459	11.949345	5.750163	44.8873049
Łupiny	16.7436996	13.9157145	13.41296202	33.382757	10.967681	11.5771858
Liczba	2.0178768	19.7810269	51.79054852	3.823147	18.483002	4.1043983

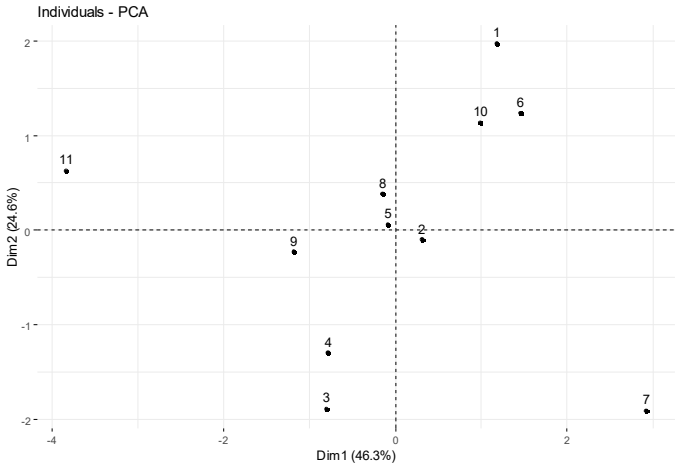
Wektor Dim.1 to wektor którego współrzędne są kwadratami współrzędnych pierwszego wektora własnego pomnożone prze 100%. Mówi nam o tym jaki procentowy udział w tworzeniu pierwszego wektora własnego mają poszczególne zmienne ($cos2/lambda1)*100% itd.



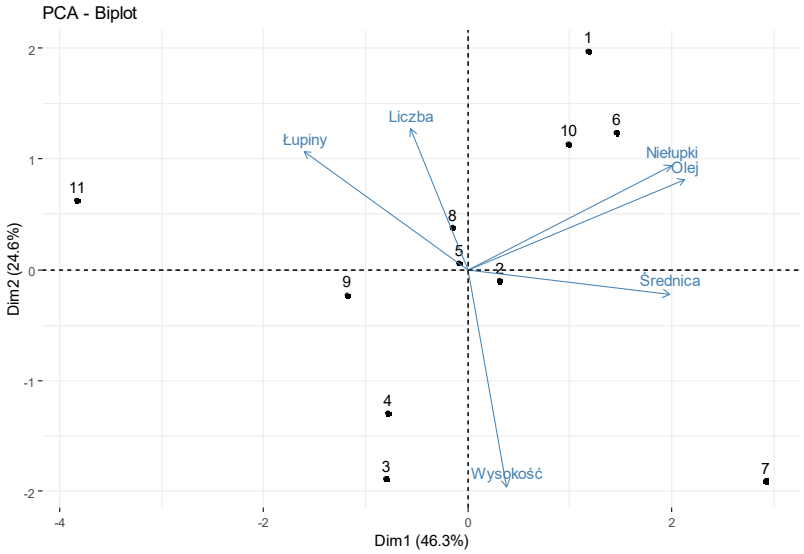
fviz_pca_var(res.pca)



fviz_pca_ind(res.pca)



```
fviz_pca_biplot(res.pca)
```



Regresja z wykorzystaniem PCA

Plik PCA i PLS Ćwicz10(faraway9)

```
library(faraway)
data (meatspec)#Spektrometria mięsa do oznaczania zawartości tłuszczu
#zmienne V1-V100 - absorpcja w zakresie 100 długości fali
# 101-sza zmienna fat zawiera dane dotyczące tłuszczu
# pierwszych 172 obserwacji potraktujemy jako zbiór uczący a resztę 173 d0 215 tako zbiór testowy
help("meatspec")
model1 <- lm (fat~.,meatspec[1:172,]) # dopasowujemy model pełnu na zbiorze uczącym
summary(model1)$r.squared
rmse <- function(x,y) sqrt(mean((x-y)^2))# funkcja wyznaczająca RMSE
rmse (model1$fit, meatspec$fat [1:172]) # rms dla zbioru uczącego
        0.6903167
rmse (predict(model1,meatspec[173:215,]),meatspec$fat[173:215])# rms dla zbioru testowego
        3.814
```

Widać że wyestymowany model ma słabe własności predykcyjne

Wykonamy procedurę PCA na zbiorze uczącym

```
library(MVA)
meatpca <- prcomp(meatspec[1:172,-101])
round (meatpca$sdev,3) #pierwiastki z wartości własnych
  [1] 5.055 0.511 0.282 0.168 0.038 0.025 0.014 0.011 0.005 0.003 0.002 0.002 0.001 0.001 0.001 0.000 0.000
 [18] 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
```

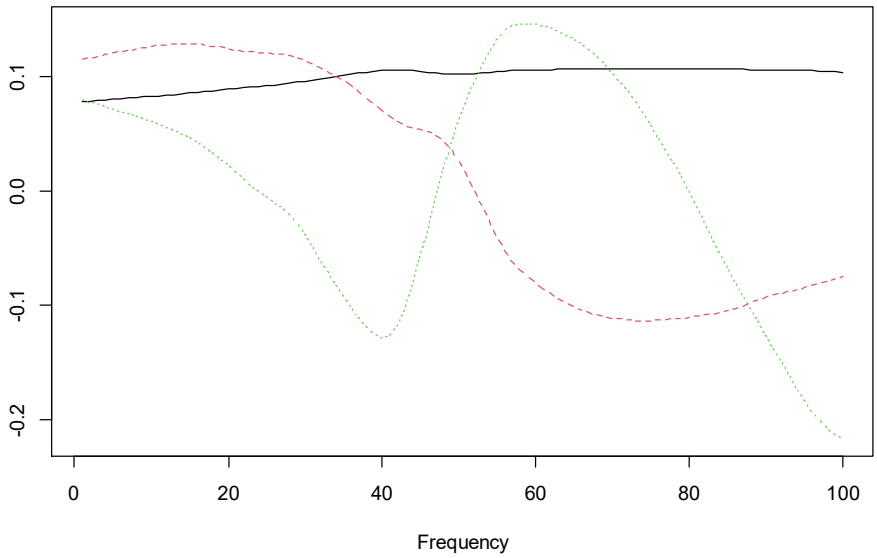
```
 [35] 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
 [52] 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
 [69] 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
 [86] 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
```

```
round (meatpca$rotation[,1:3],3)# trzy pierwsze składowe główne
```

	PC1	PC2	PC3
V1	0.079	0.115	0.080
V2	0.079	0.117	0.078
V3	0.079	0.118	0.076
V4	0.080	0.120	0.074
V5	0.080	0.121	0.072
V6	0.081	0.122	0.070
V7	0.081	0.124	0.068
V8	0.082	0.125	0.066
V9	0.082	0.126	0.064
V10	0.083	0.127	0.061
V11	0.083	0.127	0.059
V12	0.084	0.128	0.056
V13	0.084	0.129	0.053
V14	0.085	0.129	0.050
V15	0.086	0.129	0.046
V16	0.086	0.128	0.042
V17	0.087	0.128	0.038
V18	0.088	0.127	0.033
V19	0.088	0.126	0.028
V20	0.089	0.125	0.022
V21	0.090	0.124	0.016
V22	0.091	0.123	0.010
V23	0.091	0.122	0.005
V24	0.092	0.121	0.000
V25	0.093	0.121	-0.005
V26	0.093	0.121	-0.009
V27	0.094	0.120	-0.014
V28	0.095	0.119	-0.020
V29	0.096	0.117	-0.028
V30	0.096	0.114	-0.038
V31	0.097	0.111	-0.049
V32	0.098	0.108	-0.060
V33	0.100	0.104	-0.072
V34	0.100	0.101	-0.083
V35	0.101	0.097	-0.092
V36	0.102	0.093	-0.102
V37	0.103	0.088	-0.110
V38	0.104	0.082	-0.118
V39	0.105	0.076	-0.125
V40	0.106	0.071	-0.128
V41	0.106	0.065	-0.126
V42	0.106	0.061	-0.118
V43	0.106	0.058	-0.103
V44	0.106	0.056	-0.083
V45	0.105	0.054	-0.058
V46	0.104	0.052	-0.032
V47	0.103	0.049	-0.005
V48	0.103	0.044	0.019
V49	0.103	0.036	0.041
V50	0.102	0.027	0.061
V51	0.103	0.016	0.078
V52	0.103	0.003	0.095

```
V53  0.103 -0.012  0.110
V54  0.104 -0.027  0.123
V55  0.104 -0.040  0.133
V56  0.105 -0.052  0.140
V57  0.105 -0.061  0.145
V58  0.106 -0.069  0.147
V59  0.106 -0.075  0.147
V60  0.106 -0.080  0.146
V61  0.106 -0.085  0.145
V62  0.106 -0.090  0.143
V63  0.107 -0.094  0.140
V64  0.107 -0.098  0.137
V65  0.107 -0.101  0.133
V66  0.107 -0.104  0.128
V67  0.107 -0.106  0.123
V68  0.107 -0.108  0.118
V69  0.107 -0.110  0.111
V70  0.107 -0.111  0.104
V71  0.107 -0.112  0.096
V72  0.107 -0.113  0.087
V73  0.107 -0.113  0.078
V74  0.107 -0.114  0.068
V75  0.107 -0.114  0.058
V76  0.107 -0.113  0.047
V77  0.107 -0.113  0.036
V78  0.107 -0.112  0.025
V79  0.107 -0.111  0.013
V80  0.107 -0.111  0.000
V81  0.107 -0.110 -0.013
V82  0.107 -0.109 -0.027
V83  0.107 -0.107 -0.041
V84  0.107 -0.106 -0.054
V85  0.107 -0.105 -0.067
V86  0.107 -0.103 -0.080
V87  0.107 -0.100 -0.092
V88  0.106 -0.098 -0.103
V89  0.106 -0.096 -0.114
V90  0.106 -0.093 -0.126
V91  0.106 -0.091 -0.137
V92  0.106 -0.090 -0.149
V93  0.106 -0.088 -0.161
V94  0.106 -0.086 -0.173
V95  0.106 -0.084 -0.183
V96  0.105 -0.083 -0.193
V97  0.105 -0.081 -0.201
V98  0.105 -0.079 -0.208
V99  0.105 -0.077 -0.213
V100 0.104 -0.075 -0.217
```

```
matplot(1:100, meatpca$rot [,1:3],type="l",xlab="Frequency", ylab="") # wykres udziału poszczególnych
                                        #zmiennych w tworzeniu trzech pierwszych składowych głównych
```

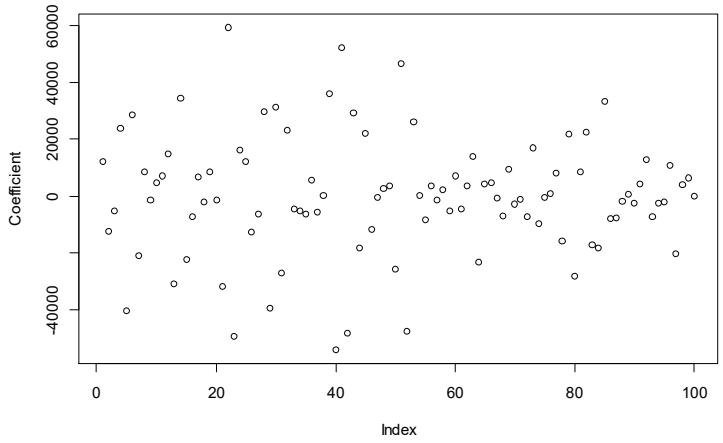


Czarna linia pokazuje udział zmiennych V1-V100 w tworzeniu pierwszej składowej głównej
Czerwona linia pokazuje udział zmiennych V1-V100 w tworzeniu drugiej składowej głównej
Zielona linia pokazuje udział zmiennych V1-V100 w tworzeniu trzeciej składowej głównej

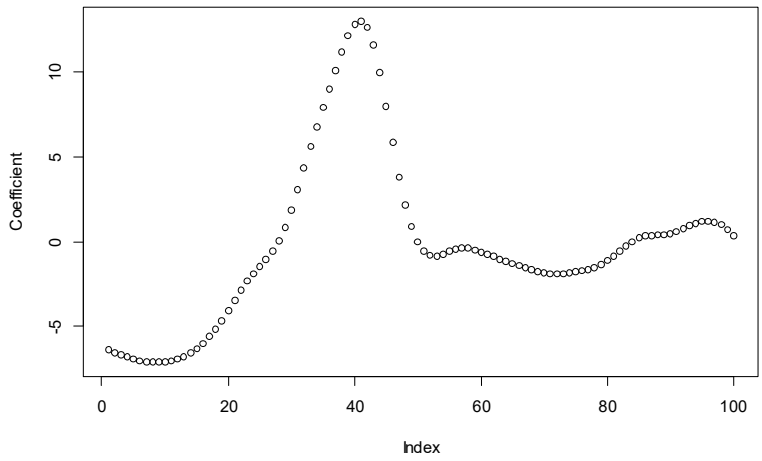
Widzimy, że pierwsza składowa główna pochodzi z prawie stałej kombinacji częstotliwości. Mierzy, czy predyktory są na ogół duże, czy małe. Druga składowa główna reprezentuje kontrast pomiędzy wyższymi i niższymi częstotliwościami. Trzecia jest trudniejsza w interpretacji.

Czasami możliwe jest, jak w tym przykładzie, nadanie znaczenia składowym głównym. Jest to zazwyczaj kwestia intuicyjnej interpretacji. W niektórych innych przypadkach nie można znaleźć żadnej interpretacji – prawie zawsze ma to miejsce, gdy predyktory mierzą zmienne w różnych skalach (takich jak wzrost i wiek danej osoby).

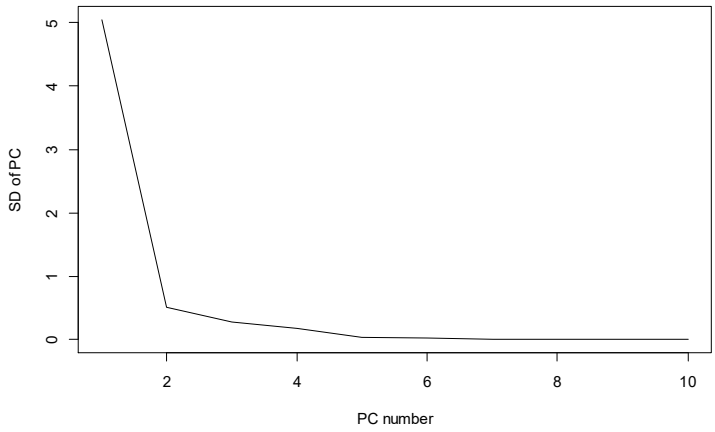
```
mode13 <- lm(fat~meatpca$x [,1:4], meatspec [1:172,])#model w oparciu o 4 składowe główne
rmse(mode13$fit,meatspec$fat [1:172])
        4.06474
plot(model1$coef[-1],ylab="Coefficient") #wykres współczynników  dla pełnego modelu
```




```
svb <- meatpca$rot[,1:4] %*% mode13$coef[-1]# wykres współczynników dla modelu opartego na PCA
plot(svb,ylab="Coefficient")
```




```
require(graphics)
plot(meatpca$sdev[1:10],type="l",ylab="SD of PC",xlab="PC number")
```

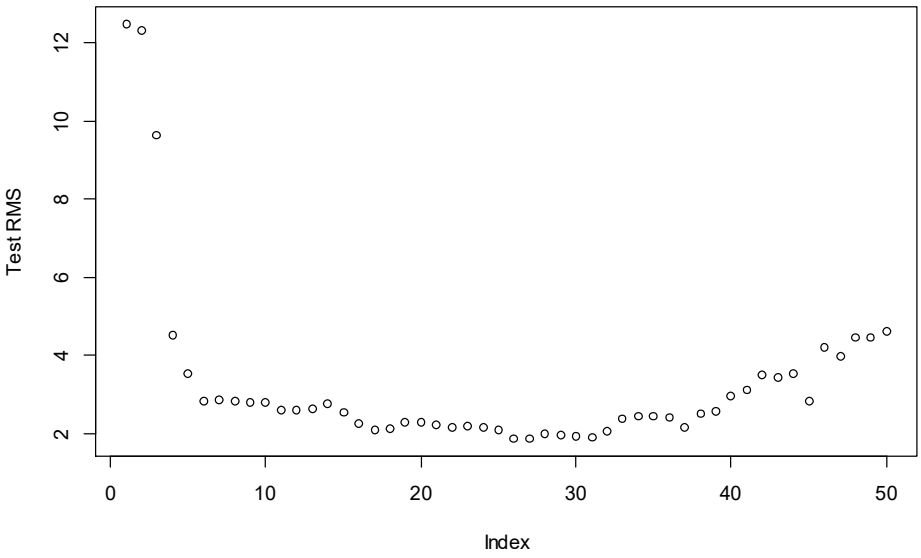



```
mm <- apply(meatspec[1:172,-101],2,mean)# wyznaczamy średnie dla 100 kolumn zbioru uczącego
tx <- as.matrix(sweep(meatspec [173:215,-101],2,mm))# tworzymy macierz zmiennych objaśniających
        # ze zbioru testowego

nx <- tx %*% meatpca$rot[,1:4] #tworzymy 4 liniowe kombinacje z 4 wektorów własnych z PCA
pv <- cbind(1,nx) %*% mode13$coef # wyznaczamy wartości przewidywane z modelu 3 dla zbioru testowego
rmse(pv,meatspec$fat[173:215])
            4.533982
```

Powtórzymy obliczenia biorąc więcej składowych głównych

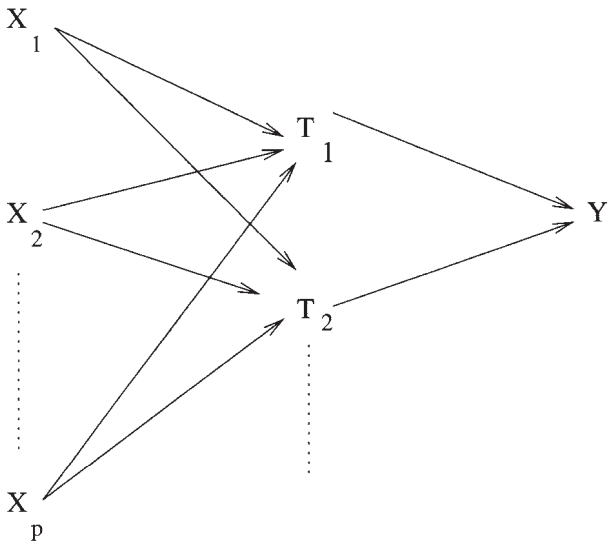
```
rmsmeat <- numeric(50)
for (i in 1:50) {
  nx <- tx %*% meatpca$rot[,1:i]
  mode13 <- lm (fat~meatpca$x[,1:i],meatspec[1:172,])
  pv <- cbind (1, nx) %*% mode13$coef
  rmsmeat[i] <- rmse(pv, meatspec$fat[173:215] )
  }
plot(rmsmeat,ylab="Test RMS")
```




```
which.min(rmsmeat)
        27
min(rmsmeat)
        1.854858
```

Najlepszy wynik predykcji na zbiorze testowym oparty jest na 27 składowych

Regresja PLS




```
library(pls)
trainx <- as.matrix(sweep(meatspec[1:172,-101],2,mm))
plsg <- plsr(meatspec$fat[1:172]~trainx,10, validation="CV")
```

```
pcrg <- pcr(meatspec$fat[1:172]~trainx,10, validation="CV")
plot (plsg$coefficients[ , , 4], ylab="Coefficient")
rmse(plsg$validation$pred,meatspec$fat[1:172])
# funkcje do wizualizacji
scoreplot(pcrg,ncomp=2,labels='names')
scoreplot(plsg,ncomp=2,labels='names')
biplot(plsg)
loadingplot(plsg)
corrplot(plsg,labels='names')



#Faraway PRA 117
g <- lm(Employed ~ ., longley)
summary(g)
round(cor(longley[,-7]),3) #macierz korelacyjna ujawnia  silne korelacje pomiędzy predyktorami
#  przeprowadzimy dekompozycję spektralną
x <- as.matrix(longley[,-7]) #macierz predyktorów
e <- eigen(t(x) %*% x)
sqrt(e$val[1]/e$val) #wartości większe od 30 sugerują współliniowość

summary(lm(x[,1] ~ x[,-1]))$r.squared # R2 dla pierwszego predyktora
1/(1-summary(lm(x[,1] ~ x[,-1]))$r.squared)# VIF dla pierwszego predyktora
vif(x) #VIF dla wszystkich predyktorów - trzeba usunąć część zmiennych z modelu
#xmienne 3 i 4  nie są silnie skorelowane z pozostałymi –niech zostaną
cor(x[,-c(3,4)])
# te zmiennensą silnie skorelowane-  zostawimy jedną np Year
summary(lm(Employed ~ Armed.Forces + Unemployed + Year,longley))

y<-longley$Employed
pcr_long <- pcr(y ~ x,4, validation="CV")
scoreplot(pcr_long,ncomp=2,labels='names')
corrplot(pcr_long,labels='names')
```