Podstawy Przetwarzania Danych

Wykład 3: Redukcja wymiarowości

dr inż. Marcin Luckner
mluckner@mini.pw.edu.pl

Wydział Matematyki i Nauk Informacyjnych

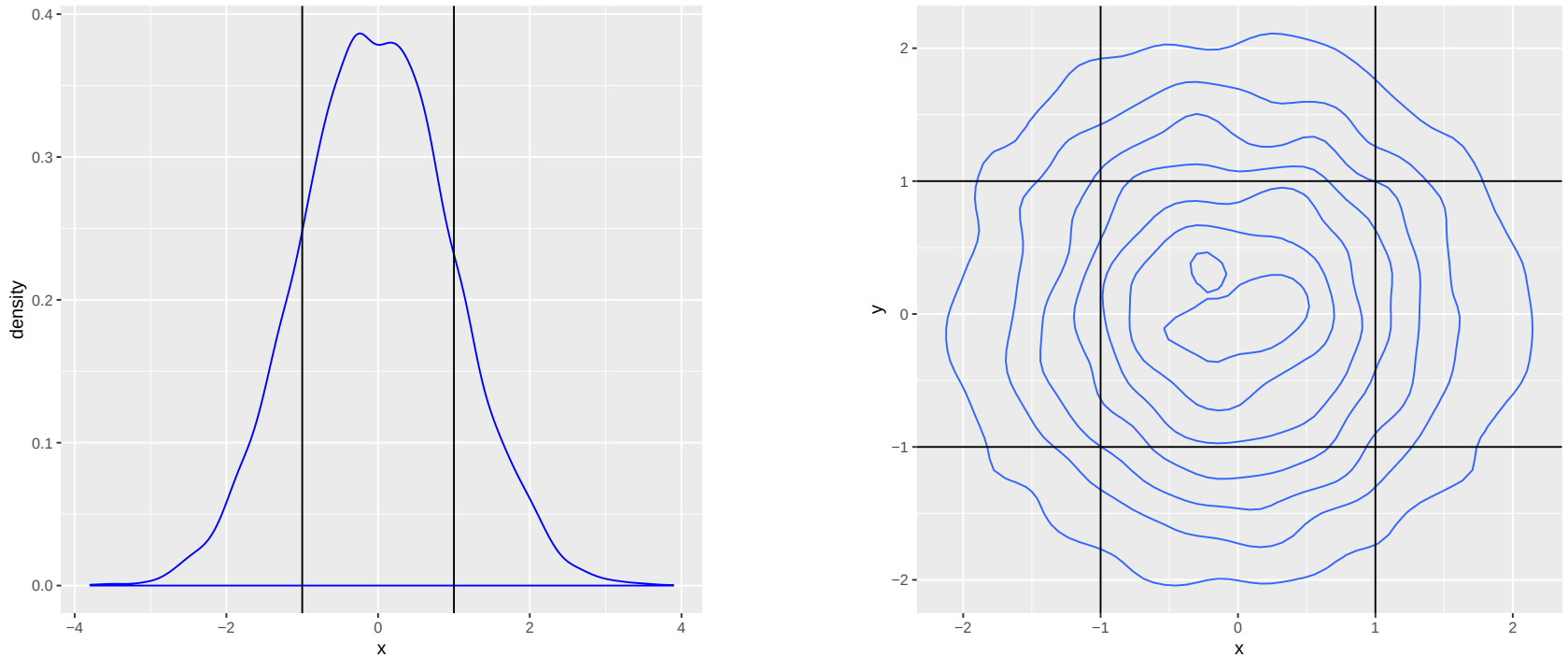
Wersja 1.2
18 października 2021

Fundusze Europejskie Wiedza Edukacja Rozwój
Rzeczpospolita Polska
Politechnika Warszawska
Unia Europejska Europejski Fundusz Społeczny

Projekt „NERW 2 PW. Nauka – Edukacja – Rozwój – Współpraca” współfinansowany jest ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

Zadanie 10 pn. „Modyfikacja programów studiów na kierunkach prowadzonych przez Wydział Matematyki i Nauk Informacyjnych”, realizowane w ramach projektu „NERW 2 PW. Nauka – Edukacja – Rozwój – Współpraca”, współfinansowanego jest ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

Zagadnienie wielowymiarowości



Rysunek 1: Rozkład normalny dla jednego i dwóch wymiarów

- Ile elementów rozkładu $X \sim \mathcal{N}(0,1)$ mieści się w hiperkuli o promieniu środku (0,0) i promieniu 1?
  - Jedna zmienna: ok. 68 procent,
  - Dwie zmienne: ok. 26 procent,
  - Dziesięć zmiennych: ok. 0.02 procenta.

Zmienne skorelowane

- Jeżeli zmienne są skorelowane to wpływają na jakość modelu decyzyjnego.
- Używając do opisu kilku cech, z których część jest w sposób niejawny powiązana z tym samym czynnikiem, zwiększamy wpływ danego czynnika na proces decyzyjny.

Tabela 1: Opis przedmiotów

Przedmiot	Wykład [h]	Laboratoria [h]	Projekt [h]	ECTS kontaktowe	ECTS praktyczne
PPD	15	30	15	2	4
PI	15	15	30	2	4
PO	15	0	30	2	3

- Punkty ECTS są wyliczane na podstawie liczby godzin zajęć.
- Mamy więc do czynienia z pozornie różnymi zmiennymi, które opisują to samo zjawisko.

Analiza składowych głównych

- Analiza składowych głównych (*Principal components analysis* (PCA)) zastępuje zbiór zmiennych mniejszym zbiorem ich kombinacji liniowych.
- Powstałe liniowe kombinacje nazywane są *składowymi głównymi*.
- Składowe główne tworzy się w oparciu o wektory własne.
- Dzięki temu osiągamy następujące efekty:
  - Składowe główne są niezależne.
  - Maksymalizację wariancji kolejnych składowych głównych.

Wartości i wektory własne

Przypomnienie

Wartości własne Niech $\mathbf{B}$ będzie macierzą wymiaru $m \times m$ i niech $\mathbf{I}$ będzie macierzą jednostkową wymiaru $m \times m$. Wartościami własnymi $\mathbf{B}$ nazywamy skalary $\lambda_1, \lambda_2, \ldots, \lambda_m$ jeżeli $\|\mathbf{B} - \lambda\mathbf{I}\| = 0$.

Wektory własne Niech $\mathbf{B}$ będzie macierzą wymiaru $m \times m$ a λ będzie wektorem własności własnych $\mathbf{B}$. Wektor $\mathbf{e}$ nazywamy wektorem własnym $\mathbf{B}$ jeżeli $\mathbf{B}e = \lambda\mathbf{e}$

Kowariancja

Przypomnienie

Kowariancja jest miarą podobieństwa dwóch zmiennych X_i i X_j

$$\sigma_{ij}^2 = \frac{\sum_{k=1}^{n}(x_{ki} - \mu_i)(x_{kj} - \mu_j)}{n},$$

gdzie μ jest średnią wartością zmiennej

- Dodatnia wartość kowariancji oznacza, że jeżeli jedna zmienna rośnie to i druga ma tendencję do wzrostu.
- Ujemna wartość kowariancji oznacza, że jeżeli jedna zmienna rośnie to druga ma tendencję do spadku.
- Jeżeli zmienne X_i i X_j są niezależne to $\sigma_{ij}^2 = 0$

Wartość kowariancji jest zależna od wielkości zmiennych!

Korelacja

Przypomnienie

Korelacja jest unormowaną kowariancją.

$$\rho_{ij} = \frac{\sigma_{ij}^2}{\sigma_{ii}\sigma_{jj}} = \frac{\sum_{k=1}^{n}(x_{ki} - \mu_i)(x_{kj} - \mu_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki} - \mu_i)^2}\sqrt{\sum_{k=1}^{n}(x_{kj} - \mu_j)^2}}$$

$$\Sigma = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \cdots & \sigma_{1n}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \cdots & \sigma_{2n}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1}^2 & \sigma_{n2}^2 & \cdots & \sigma_{nn}^2 \end{bmatrix}$$

Macierz kowariancji

$$\rho = \begin{bmatrix} \frac{\sigma_{11}^2}{\sigma_{11}\sigma_{11}} & \frac{\sigma_{12}^2}{\sigma_{11}\sigma_{22}} & \cdots & \frac{\sigma_{1n}^2}{\sigma_{11}\sigma_{nn}} \\ \frac{\sigma_{12}^2}{\sigma_{11}\sigma_{22}} & \frac{\sigma_{22}^2}{\sigma_{22}\sigma_{22}} & \cdots & \frac{\sigma_{2n}^2}{\sigma_{22}\sigma_{nn}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\sigma_{1n}^2}{\sigma_{11}\sigma_{nn}} & \frac{\sigma_{2n}^2}{\sigma_{22}\sigma_{nn}} & \cdots & \frac{\sigma_{nn}^2}{\sigma_{nn}\sigma_{nn}} \end{bmatrix}$$

Macierz korelacji

Standaryzacja

- Pierwotne zmienne $X_1, X_2, \ldots, X_m$ należy ustandaryzować tak, aby średnia każdej zmiennej wynosiła 0, a odchylenie standardowe 1.

$$Z_i = (X_i - \mu_i)/\sigma_{ii}$$

- Standaryzację możemy zapisać jako:

$$\mathbf{Z} = (\mathbf{V}^{\frac{1}{2}})^{-1}(\mathbf{X} - \mu),$$

gdzie $\mathbf{V}^{\frac{1}{2}}$ jest macierzą diagonalną macierzy kowariancji Σ.

- Standaryzacja powoduje, że macierze kowariancji i korelacji są takie same.

Wyliczenie składowych głównych

- Dla macierzy $\mathbf{Z}$ i-ta składowa główna wyliczana jest jako:

$$Y_i = \mathbf{e}_i' \mathbf{Z} = e_{i1} Z_1 + e_{i2} Z_2 + \cdots + e_{in} Z_n,$$

  gdzie $\mathbf{e}_i$ jest i-tym wektorem własnym.
- Składowa główna Y_i jest niezależna od wszystkich pozostałych składowych i maksymalizuje zmienność $Var(Y_i) = \mathbf{e}_i' \rho \mathbf{e}_i$.
- Można zauważyć, że pierwsza składowa ma większą zmienność niż wszystkie pozostałe kombinacje liniowe zmiennych.

Wpływ zmiennych na składowe

- Wpływ danej zmiennej na wyliczoną składową możemy przedstawić w postaci korelacji cząstkowej.
- Jest ona zależna od wektorów i wartości własnych macierzy korelacji ρ:

$$Corr(Y_i, Z_j) = e_{ij}\sqrt{\lambda_i}$$

- Ponieważ $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_m$ to pierwsze składowe mają potencjalnie najsilniejszy związek ze zmiennymi.

Zmienność

- Całkowita zmienność w standaryzowanym zbiorze danych nie zmienia się dla składowych i jest wyrażana przez sumę wartości własnych, a także liczbę zmiennych.

$$\sum_{i=1}^{m} Var(Y_i) = \sum_{i=1}^{m} Var(Z_i) = \sum_{i=1}^{m} \lambda_i = m$$

- Możemy oceniać część zmienności zmiennej $\mathbf{Z}$ wyjaśnianą przez składową Y_i poprzez iloraz λ_i/m.
- Zatem pierwsze składowe najlepiej tłumaczą zmienność zmiennej $\mathbf{Z}$.

Dane z Urzędu Miasta Stołecznego Warszawy

- Analizujemy dane z Urzędu Miasta Stołecznego Warszawy wyliczone dla 776 obszarów miasta

  powMieszkalna suma powierzchni całkowitych wszystkich budynków mieszkalnych,
  powUsługowa powierzchnia całkowita budynków usługowych,
  powHandlowa powierzchnia całkowita budynków handlowych,
  lMieszkańców całkowita liczba ludności,
  lDzieci liczba ludności w grupie wiekowej 0-14 lat,
  lMiejscPracy całkowita liczba miejsc pracy.

Składowe PCA

Tabela 2: Budowa składowych

	PC1	PC2	PC3	PC4	PC5	PC6
powMieszkalna	-0.5622	0.1521	0.0745	0.0585	-0.7950	-0.1406
powUsługowa	-0.1678	0.6388	0.1371	0.6750	0.2960	0.0418
powHandlowa	-0.2538	-0.0645	-0.9486	0.1513	0.0922	-0.0158
lMieszkańców	-0.5194	-0.2973	0.2056	-0.0428	0.4391	-0.6364
lDzieci	-0.5059	-0.3458	0.1705	0.0423	0.1782	0.7496
lMiejscPracy	-0.2562	0.5973	-0.0671	-0.7172	0.2175	0.1067

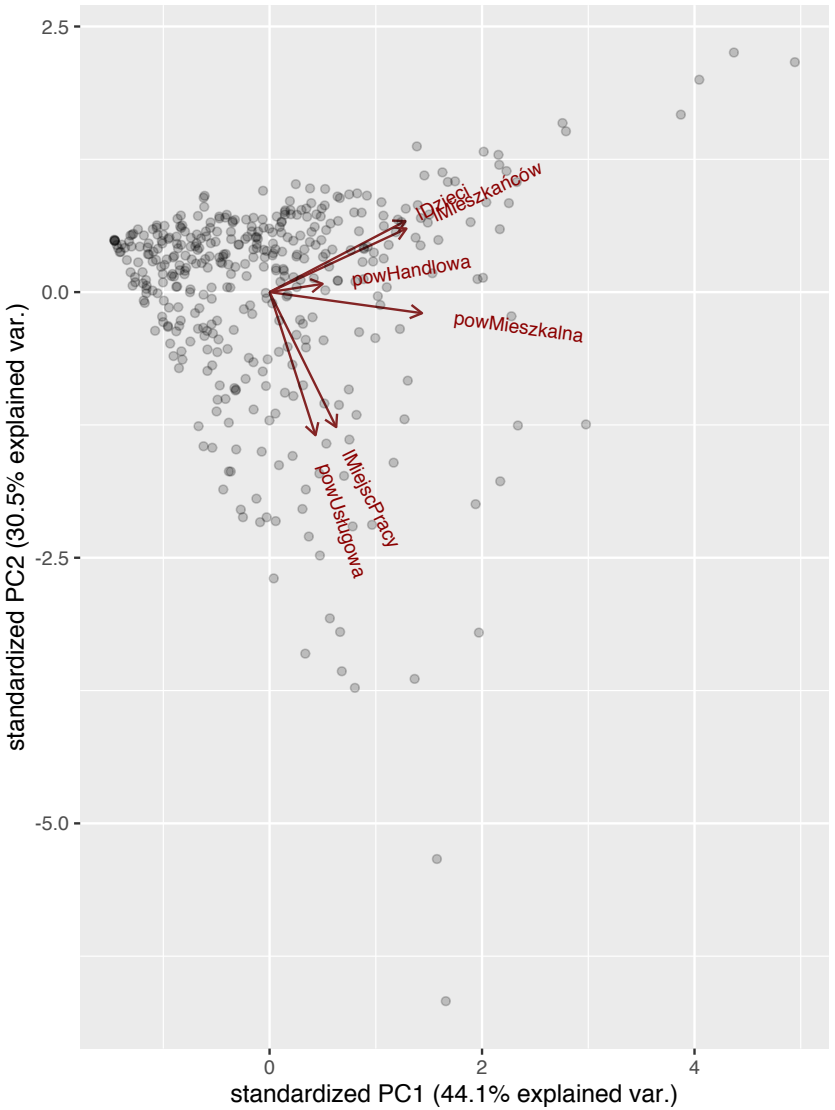
Analiza wyliczonych PCA

Tabela 3: Istotność składowych

	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	1.6259	1.3708	0.9511	0.5260	0.4220	0.3435
Proportion of Variance	0.4406	0.3132	0.1508	0.0461	0.0297	0.0197
Cumulative Proportion	0.4406	0.7538	0.9045	0.9507	0.9803	1.0000

- Odchylenie standardowe PCA.
- Udział w tłumaczeniu zmienności $\frac{\sigma_i}{\sum_{j=1}^{n} \sigma_j}$.
- Skumulowane wytłumaczenie zmienności $\sum_{i=1}^{k} \frac{\sigma_i}{\sum_{j=1}^{n} \sigma_j}$.

Wizualizacja głównych składowych



- Istnieje zależność między powierzchnią mieszkań, liczbą mieszkańców i liczbą dzieci.
- Istnieje zależność między powierzchnią usługową, a liczbą miejsc pracy.

Rysunek 2: Wizualizacja dwóch składowych

Czy składowe są lepsze od zmiennych?

- Otrzymujemy taką samą liczbę składowych jak zmiennych.
- Jednakże są one uporządkowane względem istotności i możemy usunąć kilka ostatnich składowych.
- Składowe tworzą kombinację liniową zmiennych, są więc trudniejsze do interpretacji niż zmienne.

Kryteria wyboru składowych

- Kryterium własności własnej
- Kryterium części wyjaśnionej wariancji
- Kryterium wykresu osypiskowego
- Kryterium minimalnego zasobu zmienności wspólnej

Kryterium własności własnej

- Każda składowa powinna tłumaczyć zmienność równą przynajmniej jednej zmiennej podstawowej.
- Dlatego można wybrać składowe o wartościach większych niż 1.
- Metoda może powodować wybór zbyt małej liczby składowych dla mniej niż 20 zmiennych i zbyt dużej dla powyżej 50 zmiennych.

Tabela 4: Kryterium własności własnej składowych

	PC1	PC2	PC3	PC4	PC5	PC6
σ	**1.6259**	**1.3708**	0.9511	0.5260	0.4220	0.3435

Kryterium części wyjaśnionej wariancji

- Określamy jaka część zmienności ma zostać wyjaśniona przez składowe główne.
- Następnie dobieramy składowe, aż do momentu gdy zostanie osiągnięta oczekiwana wartość.
- Jaką wartość zmienności będziemy wyjaśniać?
  - W badaniach socjologicznych cenne są już związki o niskiej wartości 60%, w naukach przyrodniczych stosuje się wyższe wymagania 90-95%.
  - Jeżeli celem jest redukcja wymiaru danych to maksymalizujemy część wyjaśnionej wariancji przy ograniczeniach z innych kryteriów.

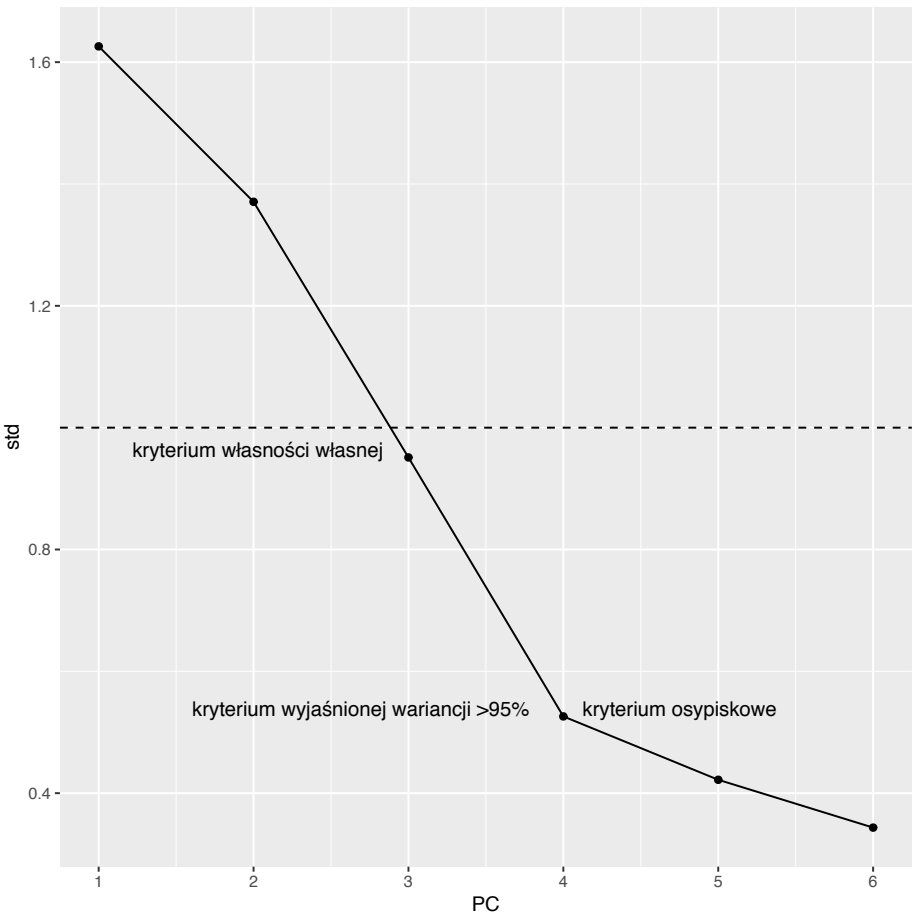
Tabela 5: Kryterium części wyjaśnionej wariancji

	PC1	PC2	PC3	PC4	PC5	PC6
$\sum_{i=1}^{k} \frac{\sigma_i}{\sum_{j=1}^{n} \sigma_j}$	**0.441**	**0.754**	**0.905**	**0.951**	0.980	1.000

Kryterium wykresu osypiskowego

- Wykres osypiskowy stanowi graficzną prezentację wartości własnych względem numeru składowej.
- Ponieważ pierwsze składowe wyjaśniają większość zmienności, to wykres opada szybko dla pierwszych składowych, a następnie staje się płaski.
- Możemy odciąć składowe zawarte w płaskiej części wykresu.

Wybór składowych - przykład



Rysunek 3: Kryteria selekcji

- Kryterium własności własnej nakazuje wybrać dwie składowe, ale ze względu na małą liczbę zmiennych i wartość std trzeciej składowej bliskiej 1 wybieramy 3.
- Kryterium wyjaśnionej wariancji wielkości 95% nakazuje wybrać 4 składowe.
- Kryterium osypiskowe wybiera 4 składowe, bo potem spadek staje się łagodny.

Kryterium minimalnego zasobu zmienności wspólnej

- Zasób zmienności wspólnej jest częścią wariancji danej zmiennej, która jest uwzględniona w wybranych składowych.
- Wyliczamy go sumując kwadraty wag danej zmiennej dla wybranych składowych.
- Wartość zasobu zmienności wspólnej mniejsza niż 0.5 może oznaczać zbyt małą reprezentację zmiennej w składowych.

Wybór liczby składowych

Tabela 6: Kryterium minimalnego zasobu zmienności wspólnej

	PC1	PC2	PC3	PC4	PC5	PC6
powMieszkalna	0.32	0.34	0.34	0.35	0.98	1.00
powUsługowa	0.03	0.44	0.45	0.91	1.00	1.00
powHandlowa	0.06	0.07	0.97	0.99	1.00	1.00
lMieszkańców	0.27	0.36	0.40	0.40	0.60	1.00
lDzieci	0.26	0.38	0.40	0.41	0.44	1.00
lMiejscPracy	0.07	0.42	0.43	0.94	0.99	1.00

- Wybierając trzy składowe uwzględniamy w znaczącym stopniu tylko jedną zmienną.
- Cztery składowe uwzględniają trzy zmienne.
- Aby uwzględnić wszystkie zmienne musimy wybrać wszystkie składowe.

Weryfikacja składowych

- Ostateczną liczbę składowych możemy ustalić korzystając z mechanizmu weryfikacji.
- Wyliczamy składowe dla części danych, zbioru uczącego.
- Następnie ponawiamy tę czynność na odrębnych danych, o tym samym charakterze, stanowiących zbiór weryfikacyjny.
- Porównujemy oba wyniki, aby sprawdzić czy składowe są reprezentacyjne dla całego zbioru danych.

Weryfikacja PCA

	PC1	PC2	PC3	PC4	PC5	PC6
powMieszkalna	**-0.5622**	0.1521	0.0745	0.0585	**-0.7950**	-0.1406
powUsługowa	-0.1678	**0.6388**	0.1371	**0.6750**	0.2960	0.0418
powHandlowa	-0.2538	-0.0645	**-0.9486**	0.1513	0.0922	-0.0158
lMieszkańców	**-0.5194**	-0.2973	0.2056	-0.0428	0.4391	**-0.6364**
lDzieci	**-0.5059**	-0.3458	0.1705	0.0423	0.1782	**0.7496**
lMiejscPracy	-0.2562	**0.5973**	-0.0671	**-0.7172**	0.2175	0.1067

	PC1	PC2	PC3	PC4	PC5	PC6
powMieszkalna	**-0.5762**	0.0589	-0.0590	0.0351	0.4626	**0.6677**
powUsługowa	-0.1827	**0.6586**	-0.0852	**0.6920**	-0.1500	-0.1557
powHandlowa	-0.1857	-0.0361	**0.9761**	0.0868	-0.0296	-0.0548
lMieszkańców	**-0.5160**	-0.2798	-0.1232	-0.0300	**-0.7908**	0.1180
lDzieci	**-0.5115**	-0.3282	-0.1443	0.0852	0.3671	**-0.6840**
lMiejscPracy	-0.2688	**0.6128**	0.0212	**-0.7101**	-0.0495	-0.2125

- Tylko cztery składowe mają powtarzalny charakter.

Podsumowanie

- Korzystając z kryteriów własności własnej, wyjaśnionej wariancji i osypiskowego uznaliśmy, że należy uwzględnić trzy lub cztery składowe w zredukowanym zbiorze danych.
- Kryterium minimalnego zasobu zmienności wspólnej pokazało, że uwzględniając cztery składowe nie będziemy dobrze reprezentować części zmiennych.
- Jednakże weryfikacja składowych na zbiorze walidacyjnym pokazała, że składowa 5 i 6 nie może być generalizowana na cały zbiór danych.
- Finalnie, redukujemy zbiór danych do czterech składowych.

Analiza czynnikowa

- Analiza czynnikowa (*factor analysis*) stanowi model danych oparty na liniowej kombinacji zmiennych.
- Na podstawie modelu analizy czynnikowej stawiamy hipotezę, że wektor zmiennych $X_1, X_2, \ldots, X_m$ można modelować jako liniowe kombinacje mniejszego zbioru k ukrytych zmiennych losowych $F_1, F_2, \ldots, F_k$ nazywanych czynnikami wspólnymi razem z składnikiem błędu $\epsilon = [\epsilon_1, \epsilon_2, \ldots, \epsilon_k]'$

$$\mathbf{X} - \mu = \mathbf{LF} + \epsilon,$$

gdzie $\mathbf{X} - \mu$ jest wektorem zmiennych przesuniętym o wektor średni tak, że jego średnią jest wektor zerowy.

Założenia i ograniczenia

- W celu ograniczenia przestrzeni rozwiązań przyjmuje się następujące założenia
  - $E(\mathbf{F}) = 1$,
  - $E(\epsilon) = 0$,
  - $Cov(\epsilon)$ jest macierzą diagonalną,
- Metoda nie daje jednoznacznych wyników. Dla ortogonalnej macierzy $\mathbf{T}$ dwa modele dadzą takie same wyniki

$$\mathbf{X} - \mu = \mathbf{LF} + \epsilon,$$

$$\mathbf{X} - \mu = (\mathbf{LT})(\mathbf{TF}) + \epsilon.$$

- Ze względu na niejednoznaczność wyników stosujemy rotację czynników.

Rotacja czynników

- Rotacja czynników (*factor rotation*) przekształca osie układu współrzędnych tak, aby zwiększać duże ładunki czynnikowe zmiennych i zmniejszać małe.
- Pomaga to w interpretacji wyodrębnionych czynników.
- Rotacja varimax maksymalizuje zmienność ładunków dla czynników w celu uzyskania kolumny złożonej z zer i jedynki dla każdej zmiennej.
- To podejście silniej wiąże czynniki z wybranymi zmiennymi i zapewnia większą niezmienniczość niż analogiczne działanie na wierszach macierzy (rotacja quartimax).

Zasadność stosowania modelu

- W celu weryfikacji modelu stosuje się test Bartletta (ang. *Bartlett's test of sphericity*) testujący zerową hipotezę, że macierz korelacji jest macierzą jednostkową (zmienne nie są skorelowane).
- Hipoteza zerowa $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$
- Statystyka testu:

$$\chi^2 = \frac{(N-k)\ln\left(S_p^2\right) - \sum_{i=1}^{k}(n_i-1)\ln\left(S_i^2\right)}{1+\frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\left(\frac{1}{n_i-1}\right)-\frac{1}{N-k}\right)},$$

  gdzie S_i^2 jest wariancją i-tej grupy, a S_p^2 jest sumaryczną wariancją $S_p^2 = \sum_{i=1}^{k}\frac{(N_i-1)s_i^2}{(N-k)}$.
- Dla p-wartości dużo większej niż 0.10 nie ma dostatecznego dowodu, że zmienne są skolerowane, więc analiza czynnikowa może być nieodpowiednia.

Czynniki FA

Tabela 7: Model złożony z trzech czynników

	Factor 1	Factor 2	Factor 3
powMieszkalna	0.70	0.49	0.31
lMieszkańców	0.89		
lDzieci	0.90		
powUsługowa		0.85	
lMiejscPracy		0.83	
powHandlowa			1.00

- Model uzyskał wartość $p = 0.0023$.
- Trzy czynniki wyjaśniają wszystkie zmienne.

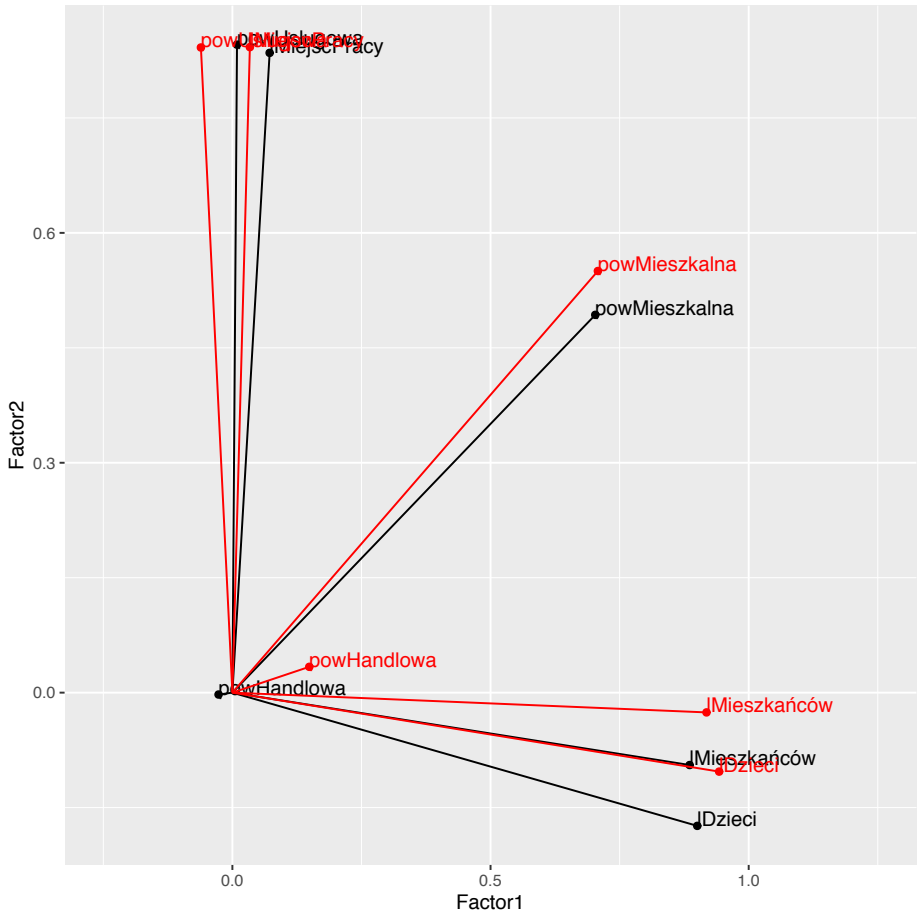
Wybór czynników

Tabela 8: Ocena czynników

	Factor 1	Factor 2	Factor 3
SS loadings	2.10	1.69	1.23
Proportion Var	0.35	0.28	0.21
Cumulative Var	0.35	0.63	0.84

- Metody wyboru czynników są podobne jak dla wyboru składowych.
- Możemy obserwować powiązanie ze zmiennymi i kumulację wariancji.

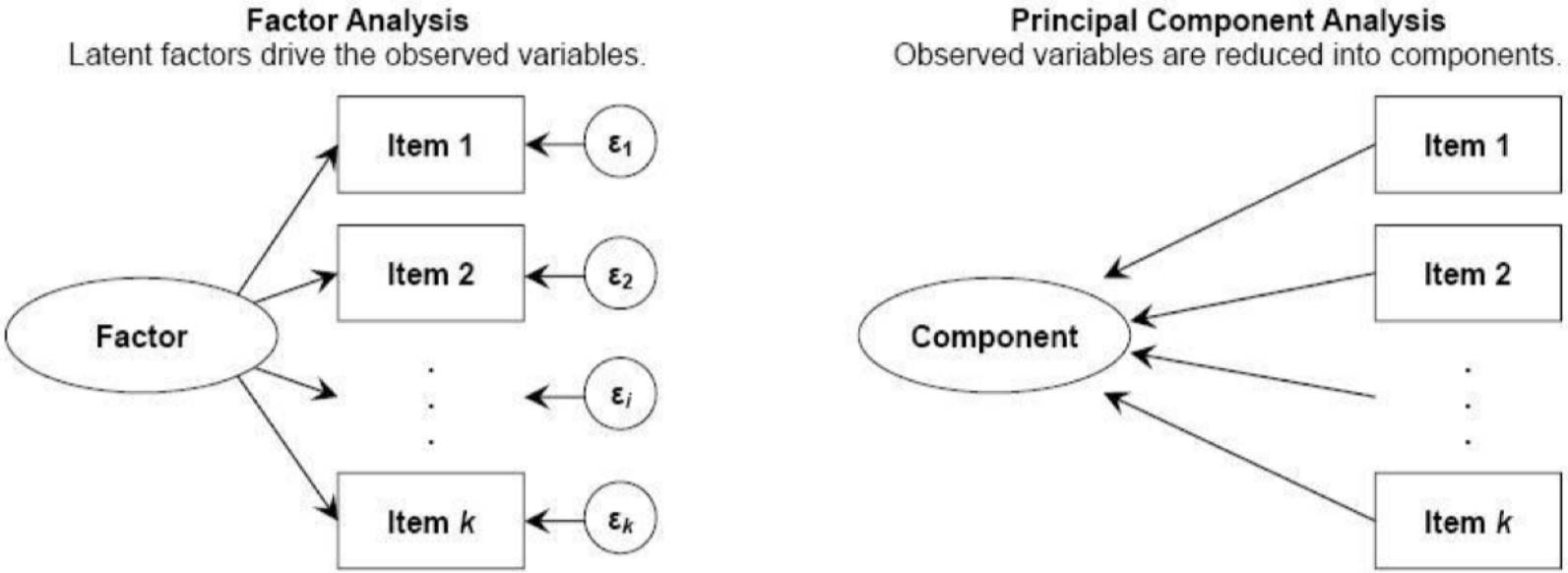
Rotacja czynników



- Rotacja spowodowała przesunięcie wektorów ładunków czynników bliżej osi współrzędnych
  - Zmiana jest szczególnie widoczna dla zmiennej lMieszkańców.

Rysunek 4: Wpływ rotacji na ładunki czynników

Porównanie metod



Rysunek 5: Porównanie modelu Analizy czynnikowej i składowych głównych [Matsunaga, 2010]

- Model analizy czynnikowej może być stosowany po potwierdzeniu jego zasadności.
- Analiza składowych głównych może być używana zawsze.

Rozkład macierzy nieujemnej

- Rozkład macierzy nieujemnej (*Nonnegative Matrix Factorisation* (NMF)) polega na rozkładzie nieujemnej macierzy $\mathbf{A} \in \Re^{m \times n}$ na dwie nieujemne macierze $\mathbf{W} \in \Re^{m \times k}$ i $\mathbf{H} \in \Re^{k \times n}$ dla $k < \min\{m, n\}$.
- W tym celu minimalizujemy funkcję

$$f(\mathbf{W}, \mathbf{H}) = \frac{1}{2} \|\mathbf{A} - \mathbf{WH}\|_F^2$$

Norma Frobeniusa

- Stosowana przy faktoryzacji norma Frobeniusa jest bezpośrednim uogólnieniem normy euklidesowej.

$$\|\mathbf{A}\|_F^2 = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}\|a_{ij}\|^2}$$

Operatory pomocnicze

- Algorytm faktoryzacji wykorzystuje notację MATLABa.
- $\mathbf{A}.*\mathbf{B}$ jest iloczynem komórek macierzy:

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 2 & 5 \\ 1 & 3 \end{pmatrix} = \begin{pmatrix} 1*2 & 2*5 \\ 3*1 & 4*3 \end{pmatrix}$$

- $\mathbf{A}./\mathbf{B}$ jest ilorazem komórek macierzy:

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 2 & 5 \\ 1 & 3 \end{pmatrix} = \begin{pmatrix} 1/2 & 2/5 \\ 3/1 & 4/3 \end{pmatrix}$$

Iteracyjny algorytm rozkładu

- Algorytm rozkładu ma charakter iteracyjny.
  1. Inicjuj $\mathbf{W}$ i $\mathbf{H}$ wartościami losowymi
  2. Przez zadaną liczbę iteracji wyliczaj

$$\mathbf{H} = \mathbf{H}. * (\mathbf{W}^T\mathbf{A}./(\mathbf{W}^T\mathbf{W}\mathbf{H} + \epsilon)$$

$$\mathbf{W} = \mathbf{W}. * (\mathbf{A}\mathbf{H}^T./(\mathbf{W}\mathbf{H}\mathbf{H}^T + \epsilon)$$

- Czynnik ϵ ma zapewnić uniknięcie dzielenia przez zero.

Właściwości algorytmu

- Algorytm jest zbieżny.
- Nie ma gwarancji zbieżności do minimum globalnego.
- Jeżeli minimum globalne leży na krawędzi analizowanej przestrzeni nie ma możliwości wykrycia tego minimum.
- Uzyskana macierz $\mathbf{H} \in \Re^{k \times n}$ redukuje wymiar cech.

Wynik działania algorytmu

- Uzyskana macierz $\mathbf{H} \in \Re^{k \times n}$ redukuje wymiar cech do k.
- Każda komórka $h_{i,j}$ określa nieujemną siłę powiązania i-tego przypadku z j-tą z k nowo powstałych cech.
- Sposób doboru parametru k nie jest oczywisty, ale powinien on być znacznie mniejszy od m.
- Zastosowania w klasyfikacji i klasteryzacji [Zdunek, 2014].

Zastosowanie NMF w klasyfikacji

- Wykrywanie Hepatocellular carcinoma (HCC) - złośliwy guz - na podstawie 57 symptomów [Ji et al., 2015].
- Zastosowano NMF by utworzyć grupy istotnych syndromów.

Feature subsets	Dimension	Classification accuracy in LSSVM (%)
FS_1	39	80.002 ± 9.95
FS_2	33	77.50 ± 12.36
OFS	49	72.50 ± 11.64

Rysunek 6: Wyniki klasyfikacji z użyciem oryginalnych syndromów (*OFS*) i danych przetworzonych z użyciem NMF (FS_1,FS_2)

Bibliografia I

[Ji et al., 2015] Ji, Z., Meng, G., Huang, D., Yue, X., and Wang, B. (2015).

NMFBFS: A nmf-based feature selection method in identifying pivotal clinical symptoms of hepatocellular carcinoma.

Comp. Math. Methods in Medicine, 2015:846942:1–846942:12.

[Matsunaga, 2010] Matsunaga, M. (2010).

How to factor-analyze your data right: Do's, don'ts, and how-to's.

International Journal of Psychological Research, 3:97–110.

[Zdunek, 2014] Zdunek, R. (2014).

Nieujemna faktoryzacja macierzy i tensorów : zastosowanie do klasyfikacji i przetwarzania sygnałów.

Projekt „NERW 2 PW. Nauka – Edukacja – Rozwój – Współpraca” współfinansowany jest ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

Zadanie 10 pn. „Modyfikacja programów studiów na kierunkach prowadzonych przez Wydział Matematyki i Nauk Informacyjnych”, realizowane w ramach projektu „NERW 2 PW. Nauka – Edukacja – Rozwój – Współpraca”, współfinansowanego jest ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.