Algorytmy eksploracji danych: Wykład 2

Wojciech Kempa

Politechnika Śląska
Wydział Matematyki Stosowanej

Założenia analizy składowych głównych (1)

Przeprowadzenie analizy składowych głównych powinno być poprzedzone weryfikacją poniższych warunków.

- Należy ocenić **zasadność** zastosowania tej metody.

Wszystkie współczynniki korelacji pomiędzy wyjściowymi zmiennymi $X_1, ..., X_p$ powinny być istotnie różne od zera (zaleca się, by były one większe od 0.3 co do wartości bezwzględnej). Do oceny zasadności zastosowania metody składowych głównych służy **test sferyczności Bartletta**.

W teście sferyczności weryfikacji podlega hipoteza zerowa postaci

$$H_0 : R = \mathbb{E},$$

gdzie R jest symetryczną macierzą kwadratową wymiaru p współczynników korelacji pomiędzy zmiennymi $X_1, ..., X_p$, natomiast $\mathbb{E}$ oznacza macierz jednostkową wymiaru p.

Założenia analizy składowych głównych (2)

Statystyką testową testu Bartletta jest następująca funkcja próby losowej cech $X_1, ..., X_p$:

$$\chi^2 = -\left(n - 1 - \frac{2p+5}{6}\right) \ln |R|$$
$$= -\left(n - 1 - \frac{2p+5}{6}\right) \sum_{i=1}^{p} \ln \lambda_i,$$

gdzie λ_i, $i = 1, ..., p$, są wartościami własnymi macierzy korelacji R, natomiast n oznacza liczebność próby, czyli liczbę przypadków, na podstawie których wyznaczana jest postać macierzy R. Przy założeniu prawdziwości hipotezy H_0 powyższa statystyka testowa ma rozkład chi kwadrat o $\frac{p(p-1)}{2}$ stopniach swobody. Zbiór krytyczny testu sferyczności ma postać $K = [\chi^2_{p(p-1)/2,1-\alpha}, \infty)$, gdzie $\chi^2_{p(p-1)/2,1-\alpha}$ jest kwantylem rzędu $1 - \alpha$ rozkładu chi kwadrat o $\frac{p(p-1)}{2}$ stopniach swobody, a α oznacza poziom istotności testu.

Założenia analizy składowych głównych (3)

- Zweryfikować należy **adekwatność** macierzy współczynników korelacji pomiędzy zmiennymi $X_1, ..., X_p$.

Do tego celu służy **współczynnik Kaisera-Mayera-Olkina (KMO)**, obliczany ze wzoru

$$KMO \stackrel{def}{=} \frac{\sum_{i \neq j} r_{i,j}^2}{\sum_{i \neq j} r_{i,j}^2 + \sum_{i \neq j} \widehat{r}_{i,j}^2},$$

gdzie $r_{i,j}$ jest wartością współczynnika korelacji liniowej Pearsona pomiędzy zmiennymi X_i oraz X_j obliczoną na podstawie realizacji n-elementowej próby losowej, natomiast $\widehat{r}_{i,j}$ oznacza **współczynnik korelacji cząstkowej** pomiędzy zmiennymi X_i oraz X_j.

Założenia analizy składowych głównych (4)

Współczynnik ten jest zdefiniowany w następujący sposób:

$$\widehat{r}_{i,j} \stackrel{def}{=} -\frac{R_{i,j}}{\sqrt{R_{i,i}R_{j,j}}},$$

gdzie $R_{i,j}$ oznacza dopełnienie algebraiczne elementu $r_{i,j}$ macierzy współczynników korelacji R pomiędzy zmiennymi.

Współczynnik korelacji cząstkowej $\widehat{r}_{i,j}$ opisuje siłę i kierunek zależności korelacyjnej pomiędzy zmiennymi X_i oraz X_j **przy wyeliminowaniu wpływu na tę zależność pozostałych zmiennych** wyjściowego układu.

Porównując zatem wartości $\widehat{r}_{i,j}$ oraz „zwykłego" współczynnika korelacji liniowej Pearsona $r_{i,j}$ możemy stwierdzić, jak bardzo zależność pomiędzy X_i a X_j jest wzmacniana bądź osłabiana poprzez obecność innych zmiennych.

Założenia analizy składowych głównych (5)

Współczynnik KMO przyjmuje wartości z przedziału $[0, 1]$. Im większa jego wartość, tym bardziej zasadne jest zastosowanie metody analizy składowych głównych do układu cech $X_1, ..., X_p$ (tym większy będzie potencjalny zysk z zastosowania PCA mierzony redukcją wymiaru zadania). Współczynnik ten powinien być większy od 0.5, a najlepiej, by przekraczał 0.7.

Założenia analizy składowych głównych (6)

- W przypadku niewielkiej próby istotne jest spełnienie założenia o **wielowymiarowej normalności** łącznego rozkładu prawdopodobieństwa cech $X_1, ..., X_p$. W przypadku dużej próby weryfikację tego warunku możemy pominąć.

Założenia analizy składowych głównych (6)

- W przypadku niewielkiej próby istotne jest spełnienie założenia o **wielowymiarowej normalności** łącznego rozkładu prawdopodobieństwa cech $X_1, ..., X_p$. W przypadku dużej próby weryfikację tego warunku możemy pominąć.
- Istotna jest odpowiednia **liczebność** oraz **reprezentatywność** próby, na bazie której przeprowadzana jest analiza (np. wyznaczany jest estymator macierzy kowariancji S układu cech). Absolutnym minimum jest próba licząca co najmniej 50 elementów, choć najlepiej, by jej liczebność wynosiła co najmniej 100.

Założenia analizy składowych głównych (7)

- Przed przystąpieniem do wykonania analizy PCA powinniśmy także dokonać procedury „czyszczenia" próby losowej celem wykrycia i eliminacji **wartości odstających**, których obecność może mieć istotny wpływ na wyniki końcowe (wartości odstające w wyraźny sposób wpływają na wartości wielu podstawowych miar statystycznych, takich jak np. średnia arytmetyczna).

Założenia analizy składowych głównych (7)

- Przed przystąpieniem do wykonania analizy PCA powinniśmy także dokonać procedury „czyszczenia" próby losowej celem wykrycia i eliminacji **wartości odstających**, których obecność może mieć istotny wpływ na wyniki końcowe (wartości odstające w wyraźny sposób wpływają na wartości wielu podstawowych miar statystycznych, takich jak np. średnia arytmetyczna).

- Ostatnim elementem jest **wstępna weryfikacja liczby zmiennych**. Zmienne silnie skorelowane z innymi zmiennymi powinniśmy odrzucić. Niespełnienie tego warunku może skutkować wyznaczeniem wartości własnych macierzy kowariancji cech, które będą bardzo bliskie zeru.

Górne ograniczenie dla największej wartości własnej

- Przeprowadzenie analizy składowych głównych wymaga wyznaczenia co najmniej jednej (w takim wypadku: największej) wartości własnej macierzy kowariancji S (lub ewentualnie macierzy współczynników korelacji R) wyjściowego układu zmiennych $X_1, ..., X_p$.

Górne ograniczenie dla największej wartości własnej

- Przeprowadzenie analizy składowych głównych wymaga wyznaczenia co najmniej jednej (w takim wypadku: największej) wartości własnej macierzy kowariancji S (lub ewentualnie macierzy współczynników korelacji R) wyjściowego układu zmiennych $X_1, ..., X_p$.
- Jak się okazuje, możliwe jest uzyskanie górnego oszacowania największej wartości własnej macierzy kowariancji S w zależności od elementów tej macierzy.

Górne ograniczenie dla największej wartości własnej

- Przeprowadzenie analizy składowych głównych wymaga wyznaczenia co najmniej jednej (w takim wypadku: największej) wartości własnej macierzy kowariancji S (lub ewentualnie macierzy współczynników korelacji R) wyjściowego układu zmiennych $X_1, ..., X_p$.
- Jak się okazuje, możliwe jest uzyskanie górnego oszacowania największej wartości własnej macierzy kowariancji S w zależności od elementów tej macierzy.
- Jeżeli tylko macierz kowariancji $S = (s_{i,j})$ jest nieujemnie określona, prawdziwa jest następująca nierówność:

$$\lambda_1 \leq \max_{1 \leq i \leq p} \sum_{j=1}^{p} |s_{i,j}|.$$

Idea analizy czynnikowej (1)

- Techniką redukcji wymiaru zadania podobną do analizy składowych głównych jest **analiza czynnikowa** (ang. *Factor Analysis (FA)*).

Idea analizy czynnikowej (1)

- Techniką redukcji wymiaru zadania podobną do analizy składowych głównych jest **analiza czynnikowa** (ang. *Factor Analysis (FA)*).
- Jej zasadniczym celem jest zastąpienie oryginalnych, badanych zmiennych $X_1, ..., X_p$ mniejszą liczbą nowych zmiennych (zwanych **czynnikami**) takich, od których wyjściowe zmienne będą liniowo zależne i które, dodatkowo, najlepiej jak to tylko możliwe wyjaśniają zależności pomiędzy nimi.

Idea analizy czynnikowej (1)

- Techniką redukcji wymiaru zadania podobną do analizy składowych głównych jest **analiza czynnikowa** (ang. *Factor Analysis (FA)*).
- Jej zasadniczym celem jest zastąpienie oryginalnych, badanych zmiennych $X_1, ..., X_p$ mniejszą liczbą nowych zmiennych (zwanych **czynnikami**) takich, od których wyjściowe zmienne będą liniowo zależne i które, dodatkowo, najlepiej jak to tylko możliwe wyjaśniają zależności pomiędzy nimi.
- **Zasadność zastosowania techniki analizy czynnikowej** sprawdzamy, podobnie jak w przypadku analizy składowych głównych, wykonując test sferyczności Bartletta oraz obliczając współczynnik Kaisera-Mayera-Olkina (KMO).

Idea analizy czynnikowej (2)

Ogólny model analizy czynnikowej ma następującą postać:

$$\begin{cases} X_1 = a_{1,1}F_1 + ... + a_{1,k}F_k + \varepsilon_1, \\ X_2 = a_{2,1}F_1 + ... + a_{2,k}F_k + \varepsilon_2, \\ \dots, \\ X_p = a_{p,1}F_1 + ... + a_{p,k}F_k + \varepsilon_p, \end{cases} \tag{1}$$

gdzie $F_1, ..., F_k$ to tzw. **czynniki wspólne** (na ogół jest ich mniej niż zmiennych wyjściowych, tzn. $k < p$), współczynniki $a_{i,j}$ to tzw. **ładunki czynnikowe** (wartość $a_{i,j}$ to ładunek czynnika F_j w zmiennej X_i), natomiast ε_i są tzw. **czynnikami swoistymi (losowymi)**. Z reguły zmienne oryginalne (obserwowane) $X_1, ..., X_p$ przed rozpoczęciem analizy są standaryzowane.

Idea analizy czynnikowej (3)

- W modelu analizy czynnikowej przyjmuje się założenie o nieskorelowaniu między sobą poszczególnych czynników wspólnych F_j oraz czynników swoistych ε_i.

Idea analizy czynnikowej (3)

- W modelu analizy czynnikowej przyjmuje się założenie o nieskorelowaniu między sobą poszczególnych czynników wspólnych F_j oraz czynników swoistych ε_i.
- Dodatkowo, czynniki wspólne są nieskorelowane z czynnikami swoistymi. O czynnikach wspólnych zakłada się także, że są zmiennymi losowymi o zerowej wartości oczekiwanej (średniej) i wariancji równej 1.

Idea analizy czynnikowej (3)

- W modelu analizy czynnikowej przyjmuje się założenie o nieskorelowaniu między sobą poszczególnych czynników wspólnych F_j oraz czynników swoistych ε_i.
- Dodatkowo, czynniki wspólne są nieskorelowane z czynnikami swoistymi. O czynnikach wspólnych zakłada się także, że są zmiennymi losowymi o zerowej wartości oczekiwanej (średniej) i wariancji równej 1.

- Czynniki swoiste są zmiennymi losowymi o rozkładach normalnych z zerową wartością oczekiwaną. Wariancję i-tego czynnika swoistego nazywamy **wariancją swoistą** i oznaczamy przez Ψ_i, $i = 1, ..., p$.

Idea analizy czynnikowej (3)

- W modelu analizy czynnikowej przyjmuje się założenie o nieskorelowaniu między sobą poszczególnych czynników wspólnych F_j oraz czynników swoistych ε_i.
- Dodatkowo, czynniki wspólne są nieskorelowane z czynnikami swoistymi. O czynnikach wspólnych zakłada się także, że są zmiennymi losowymi o zerowej wartości oczekiwanej (średniej) i wariancji równej 1.

- Czynniki swoiste są zmiennymi losowymi o rozkładach normalnych z zerową wartością oczekiwaną. Wariancję i-tego czynnika swoistego nazywamy **wariancją swoistą** i oznaczamy przez Ψ_i, $i = 1, ..., p$.
- Wariancja swoista Ψ_i wyraża część zmienności (wariancji) zmiennej X_i niezależną od pozostałych zmiennych oryginalnych, związaną z czynnikiem swoistym ε_i.

Idea analizy czynnikowej (4)

- W analizie czynnikowej punkty reprezentujące n-elementową próbę losową **możemy utożsamiać z wektorami o wspólnym początku** w początku p-wymiarowego układu współrzędnych i końcach w punktach $(X_{1,i}, ..., X_{p,i})$, $i = 1, ..., n$.

Idea analizy czynnikowej (4)

- W analizie czynnikowej punkty reprezentujące n-elementową próbę losową **możemy utożsamiać z wektorami o wspólnym początku** w początku p-wymiarowego układu współrzędnych i końcach w punktach $(X_{1,i}, ..., X_{p,i})$, $i = 1, ..., n$.
- Wektory te, geometrycznie rzecz ujmując, mogą grupować się w pewne „wiązki". Celem analizy czynnikowej jest zatem wykrycie i opisanie za pomocą jak najmniejszej liczby współrzędnych tej właśnie „wewnętrznej" struktury wektorów stanowiących próbę.

Idea analizy czynnikowej (4)

- W analizie czynnikowej punkty reprezentujące n-elementową próbę losową **możemy utożsamiać z wektorami o wspólnym początku** w początku p-wymiarowego układu współrzędnych i końcach w punktach $(X_{1,i}, ..., X_{p,i})$, $i = 1, ..., n$.
- Wektory te, geometrycznie rzecz ujmując, mogą grupować się w pewne „wiązki". Celem analizy czynnikowej jest zatem wykrycie i opisanie za pomocą jak najmniejszej liczby współrzędnych tej właśnie „wewnętrznej" struktury wektorów stanowiących próbę.
- Jak tego dokonać? Najlepiej tak, by nowe osie (współrzędne) przebiegały wzdłuż najbardziej „wyraźnych wiązek" wektorów opisujących oryginalną próbę. Zadanie to ma jednak nieskończenie wiele rozwiązań, ponieważ nie można w sposób jednoznaczny wyznaczyć początku takiego nowego układu współrzędnych (tymi nowymi współrzędnymi będą właśnie czynniki wspólne F_i, $i = 1, ..., k$).

Idea analizy czynnikowej (5)

- Drugi etap analizy czynnikowej polega zatem na optymalnym wyborze takiego układu (a co za tym idzie, takich czynników wspólnych), by w jak najprostszy sposób dokonać ich interpretacji (z reguły bowiem jest to bardzo trudne).

Idea analizy czynnikowej (5)

- Drugi etap analizy czynnikowej polega zatem na optymalnym wyborze takiego układu (a co za tym idzie, takich czynników wspólnych), by w jak najprostszy sposób dokonać ich interpretacji (z reguły bowiem jest to bardzo trudne).

- Dokonujemy tego, wykorzystując tzw. **rotacje** nowych osi współrzędnych. Sposobów wykonania takiej optymalnej rotacji jest kilka i zależą one od przyjętego kryterium odniesienia.

Matematyczny opis techniki FA (1)

Przejdźmy teraz do precyzyjnego matematycznego opisu techniki FA. Przyjmując następujące oznaczenia:

$$\mathbf{X} \stackrel{def}{=} [X_1, X_2, ..., X_p]^T, \qquad \mathbf{F} \stackrel{def}{=} [F_1, F_2, ..., F_k]^T,$$

$$A \stackrel{def}{=} \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,k} \\ a_{2,1} & a_{2,2} & \dots & a_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p,1} & a_{p,2} & \dots & a_{p,k} \end{bmatrix}, \quad \varepsilon \stackrel{def}{=} [\varepsilon_1, \varepsilon_2, ..., \varepsilon_p]^T,$$

możemy zapisać model (1) w następującej postaci macierzowej:

$$\mathbf{X} = A \cdot \mathbf{F} + \varepsilon.$$

Matematyczny opis techniki FA (2)

Wyznaczmy teraz wartość wariancji zmiennej losowej X_i, $i = 1, ..., p$. Mamy

$$\begin{aligned}\mathbf{Var}(X_i) &= \mathbf{Var}(a_{i,1}F_1 + ... + a_{i,k}F_k + \varepsilon_i) \\ &= a_{i,1}^2\mathbf{Var}(F_1) + ... + a_{i,k}^2\mathbf{Var}(F_k) + \Psi_i \\ &= \sum_{j=1}^{p} a_{i,j}^2 + \Psi_i.\end{aligned}$$

Z kolei kowariancja zmiennych X_i oraz X_j (dla $i \neq j$) będzie równa

$$\begin{aligned}Cov(X_i, X_j) &= \mathbf{E}[(X_i - \mathbf{E}(X_i))(X_j - \mathbf{E}(X_j))] \\ &= \mathbf{E}(X_iX_j) - \mathbf{E}(X_i)\mathbf{E}(X_j) = \sum_{r=1}^{k} a_{i,r}a_{j,r}.\end{aligned}$$

Matematyczny opis techniki FA (3)

W konsekwencji więc macierz kowariancji S układu zmiennych $X_1, ..., X_p$ można przedstawić w następujący sposób:

$$S = A \cdot A^T + \mathbf{\Psi},$$

gdzie

$$\mathbf{\Psi} \stackrel{def}{=} \begin{bmatrix} \Psi_1 & \dots & \dots & 0 \\ 0 & \Psi_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \Psi_p \end{bmatrix}$$

jest macierzą kowariancji czynników swoistych.

Matematyczny opis techniki FA (4)

Elementy diagonalne macierzy $A \cdot A^T$

$$h_i^2 \stackrel{def}{=} \mathbf{Var}(X_i) - \Psi_i = \sum_{j=1}^{k} a_{i,j}^2$$

nazywamy **zasobami zmienności wspólnej**. W szczególności, wartość h_i^2 jest zasobem zmienności wspólnej zmiennej X_i, czyli częścią wariancji tej zmiennej zależną od pozostałych zmiennych oryginalnych.

Całkowita zmienność wyjściowego układu cech $X_1, ..., X_p$ jest zatem równa **sumie zasobów wspólnych poszczególnych zmiennych i ich wariancji swoistych**, czyli

$$\sum_{i=1}^{p} \mathbf{Var}(X_i) = \sum_{i=1}^{p} \left(\sum_{j=1}^{k} a_{i,j}^2 + \Psi_i \right) = \sum_{i=1}^{p} \left(h_i^2 + \psi_i \right).$$

Matematyczny opis techniki FA (5)

Z drugiej jednak strony, zmieniając kolejność sumowania w sumie

$$\sum_{i=1}^{p}\sum_{j=1}^{k} a_{i,j}^2,$$

możemy zapisać

$$\sum_{j=1}^{k}\sum_{i=1}^{p} a_{i,j}^2 = \sum_{j=1}^{k} f_j^2,$$

gdzie

$$f_j^2 \stackrel{def}{=} \sum_{i=1}^{p} a_{i,j}^2$$

możemy określić jako **wariancję czynnika** F_j, czyli część zmienności wyjściowego układu cech wyjaśnianą przez czynnik (nową zmienną) F_j.

Matematyczny opis techniki FA (6)

- Podobnie jak to było w przypadku analizy składowych głównych, wartość $a_{i,j}$ jest równa współczynnikowi korelacji pomiędzy zmienną X_i a czynnikiem wspólnym F_j (stąd nazywamy ją **ładunkiem czynnikowym**).

Matematyczny opis techniki FA (6)

- Podobnie jak to było w przypadku analizy składowych głównych, wartość $a_{i,j}$ jest równa współczynnikowi korelacji pomiędzy zmienną X_i a czynnikiem wspólnym F_j (stąd nazywamy ją **ładunkiem czynnikowym**).
- Kwadrat ładunku czynnikowego (wyrażony w procentach) interpretować można jako część wariancji oryginalnej zmiennej X_i wyjaśnioną za pomocą czynnika F_j.