Algorytmy eksploracji danych: Wykład 1

Wojciech Kempa

Politechnika Śląska
Wydział Matematyki Stosowanej

Koncepcja składowych głównych (1)

Załóżmy, że pewne zjawisko (obiekt) opisane jest za pomocą zmiennych (cech) $X_1, ..., X_p$. Ponieważ p w praktyce może być duże, a przez to może mieć niekorzystny wpływ na statystyczną analizę zjawiska, chcielibyśmy zredukować liczbę zmiennych, tak jednak, aby zachować tak dużo informacji o badanym zjawisku jak to tylko możliwe. Jedną z technik statystycznych stosowanych do rozwiązania tego zagadnienia jest analiza składowych głównych.

Koncepcja składowych głównych (2)

Zasadnicze cele **analizy składowych głównych** (ang. *Principal Components Analysis (PCA)*) można określić w następujący sposób:

- redukcja liczby zmiennych opisujących dane zjawisko poprzez wprowadzenie pewnych zmiennych „sztucznych” (nieobserwowalnych), wzajemnie ortogonalnych, będących pewnymi kombinacjami liniowymi zmiennych wyjściowych (obserwowalnych);

Koncepcja składowych głównych (2)

Zasadnicze cele **analizy składowych głównych** (ang. *Principal Components Analysis (PCA)*) można określić w następujący sposób:

- redukcja liczby zmiennych opisujących dane zjawisko poprzez wprowadzenie pewnych zmiennych „sztucznych” (nieobserwowalnych), wzajemnie ortogonalnych, będących pewnymi kombinacjami liniowymi zmiennych wyjściowych (obserwowalnych);
- wykrycie struktury i ogólnych prawidłowości w związkach pomiędzy zmiennymi;

Koncepcja składowych głównych (2)

Zasadnicze cele **analizy składowych głównych** (ang. *Principal Components Analysis (PCA)*) można określić w następujący sposób:

- redukcja liczby zmiennych opisujących dane zjawisko poprzez wprowadzenie pewnych zmiennych „sztucznych” (nieobserwowalnych), wzajemnie ortogonalnych, będących pewnymi kombinacjami liniowymi zmiennych wyjściowych (obserwowalnych);
- wykrycie struktury i ogólnych prawidłowości w związkach pomiędzy zmiennymi;
- ocena (weryfikacja) wykrytych powiązań i prawidłowości;

Koncepcja składowych głównych (2)

Zasadnicze cele **analizy składowych głównych** (ang. *Principal Components Analysis (PCA)*) można określić w następujący sposób:

- redukcja liczby zmiennych opisujących dane zjawisko poprzez wprowadzenie pewnych zmiennych „sztucznych” (nieobserwowalnych), wzajemnie ortogonalnych, będących pewnymi kombinacjami liniowymi zmiennych wyjściowych (obserwowalnych);
- wykrycie struktury i ogólnych prawidłowości w związkach pomiędzy zmiennymi;
- ocena (weryfikacja) wykrytych powiązań i prawidłowości;
- opis zjawiska za pomocą nowych współrzędnych (składowych głównych).

Wyznaczanie składowych głównych (1)

Weźmy pod uwagę zmienne $X_1, X_2, ..., X_p$. Chcielibyśmy zredukować ich liczbę, zachowując jednocześnie tak dużo zmienności (informacji) jaką niosą ze sobą jak to tylko możliwe. Tworzymy zatem nowe nieobserwowalne zmienne, które będą kombinacjami liniowymi zmiennych oryginalnych (obserwowalnych). Będą to tzw. **składowe główne**. Niech Z_1 będzie pierwszą składową główną, tj.

$$Z_1 \stackrel{def}{=} a_{1,1}X_1 + a_{1,2}X_2 + ... + a_{1,p}X_p. \quad (1)$$

Równanie (1), mimo, że jest podobne do równania regresji wielorakiej, różni się od niego dość istotnie. Nie wprowadzamy w nim rozróżnienia zmiennych niezależnych i zmiennej zależnej, brak jest w nim wyrazu wolnego oraz składnika losowego (reszty). By zachować tak dużo zmienności wyjściowego układu zmiennych obserwowalnych $X_1, ..., X_p$ jak to tylko możliwe, wyznaczamy współczynniki $a_{1,j}$, $j = 1, ..., p$, równania (1) tak, aby wariancja $S^2(Z_1)$ zmiennej Z_1 była maksymalna.

Wyznaczanie składowych głównych (2)

Stosujemy jednak ograniczenie $\sum_{j=1}^{p} a_{1,j}^2 = 1$, czyli normalizujemy wektor współczynników

$$\mathbf{a}_1 \stackrel{def}{=} [a_{1,1}, ..., a_{1,p}]^T,$$

przyjmując $|\mathbf{a}_1| = 1$. Rezygnacja z tego warunku spowodowałaby, że $S^2(Z_1)$ dążyłoby do nieskończoności, gdyby którekolwiek z $a_{1,j}$, $j = 1, ..., p$, dążyło do nieskończoności.

Przyjmując oznaczenie $\mathbf{X} \stackrel{def}{=} [X_1, X_2, ..., X_p]^T$, mamy następującą równość:

$$S^2(Z_1) = S^2(\mathbf{a}_1{}^T\mathbf{X}) = \mathbf{a}_1{}^T S \mathbf{a}_1 = \sum_{i=1}^{p}\sum_{j=1}^{p} a_{1,i} a_{1,j} s_{i,j},$$

gdzie $S = (s_{i,j})$ jest macierzą kowariancji układu zmiennych wyjściowych $X_1, ..., X_p$.

Wyznaczanie składowych głównych (3)

Mamy

$$s_{i,j} \stackrel{def}{=} Cov(X_i, X_j) = \frac{1}{n}\sum_{k=1}^{n}(X_{i,k} - \overline{X}_i)(X_{j,k} - \overline{X}_j),$$

gdzie $X_{i,k}$ jest wartością zmiennej X_i uzyskaną dla k-tego elementu p-wymiarowej n-elementowej próby losowej zmiennych (cech) $X_1, ..., X_p$, a $\overline{X}_i$ jest średnią wartością cechy X_i w tej próbie.

Wyznaczanie składowych głównych (4)

Poszukujemy zatem takiego wektora $\mathbf{a}_1$, dla którego $S^2(Z_1)$ będzie maksymalne, przy czym $\mathbf{a}_1{}^T\mathbf{a}_1 = 1$ (jest to warunek normalizacyjny równoważny temu, że wektor $\mathbf{a_1}$ jest wektorem jednostkowym). W rozwiązaniu powyższego zagadnienia wykorzystamy rachunek różniczkowy, uwzględniając warunek normalizacyjny w postaci mnożnika Lagrange'a l_1. Obliczając pochodną względem wektora $\mathbf{a_1}$, otrzymujemy

$$\frac{\partial}{\partial \mathbf{a}_1}\left[S^2(Z_1) + l_1(1 - \mathbf{a}_1{}^T\mathbf{a}_1)\right] = \frac{\partial}{\partial \mathbf{a}_1}\left[\mathbf{a}_1{}^T S\mathbf{a}_1 + l_1(1 - \mathbf{a}_1{}^T\mathbf{a}_1)\right]$$
$$= 2(S - l_1\mathbb{E})\mathbf{a}_1.$$

Wyznaczanie składowych głównych (5)

Poszukiwane współrzędne wektora $\mathbf{a}_1$ muszą zatem spełniać układ p równań liniowych postaci (warunek konieczny istnienia ekstremum)

$$(S - l_1\mathbb{E})\mathbf{a_1} = \mathbf{0}, \tag{2}$$

gdzie $\mathbf{0} \stackrel{def}{=} \underbrace{[0, 0, ..., 0]}_{p}^{T}$.

Wyznaczanie składowych głównych (6)

Ponieważ rozwiązanie tego układu musi być niezerowe, zatem l_1 musi być liczbą taką, by $\det(S - l_1\mathbb{E}) = 0$.
Wnioskujemy stąd, że l_1 musi być wartością własną macierzy kowariancji S, natomiast $\mathbf{a}_1$ - odpowiadającym tej wartości wektorem własnym. Wykorzystując fakt, że ${\mathbf{a}_1}^T\mathbf{a}_1 = 1$, otrzymujemy teraz, mnożąc równanie (2) lewostronnie przez ${\mathbf{a}_1}^T$,

$$
\begin{aligned}
{\mathbf{a}_1}^T(S - l_1\mathbb{E})\mathbf{a}_1 = 0,\\
{\mathbf{a}_1}^T S\mathbf{a}_1 - l_1{\mathbf{a}_1}^T\mathbf{a}_1 = 0,
\end{aligned}
$$

a stąd

$$
l_1 = {\mathbf{a}_1}^T S\mathbf{a}_1 = S^2(Z_1).
$$

Ponieważ wariancja $S^2(Z_1)$ zmiennej Z_1 (pierwszej składowej głównej) miała być maksymalna, zatem l_1 musi być **największą wartością własną macierzy kowariancji** S, zaś $\mathbf{a}_1$ - **wektorem własnym odpowiadającym największej wartości własnej**.

Wyznaczanie składowych głównych (12)

W analogiczny sposób definiujemy kolejne składowe główne: współczynniki kolejnych składowych będą współrzędnymi wektorów własnych odpowiadających kolejnym największym wartościom własnym macierzy kowariancji S wyjściowego układu zmiennych $X_1, ..., X_p$. Ogólnie zatem, ponieważ macierz S jest macierzą stopnia p, zdefiniować możemy p składowych głównych $Z_1, ..., Z_p$ dla układu cech $X_1, ..., X_p$ w następujący sposób:

$$
\begin{aligned}
Z_1 &= a_{1,1}X_1 + a_{1,2}X_2 + ... + a_{1,p}X_p, \\
Z_2 &= a_{2,1}X_1 + a_{2,2}X_2 + ... + a_{2,p}X_p, \\
&\dots \\
Z_{p-1} &= a_{p-1,1}X_1 + a_{p-1,2}X_2 + ... + a_{p-1,p}X_p, \\
Z_p &= a_{p,1}X_1 + a_{p,2}X_2 + ... + a_{p,p}X_p,
\end{aligned}
$$

gdzie $\mathbf{a}_i \stackrel{def}{=} [a_{i,1}, ..., a_{i,p}]^T$, $i \in \{1, ..., p\}$, jest **wektorem własnym macierzy S odpowiadającym i-tej największej wartości własnej λ_i tej macierzy**.

Cechy składowych głównych

Podsumowując powyższe rozważania, możemy sformułować następujące uwagi i wnioski:

- składowe główne $Z_1, ..., Z_p$ wyjściowego układu cech $X_1, ..., X_p$ są nieobserwowalnymi („sztucznymi") zmiennymi będącymi liniowymi kombinacjami zmiennych oryginalnych;

Cechy składowych głównych

Podsumowując powyższe rozważania, możemy sformułować następujące uwagi i wnioski:

- składowe główne $Z_1, ..., Z_p$ wyjściowego układu cech $X_1, ..., X_p$ są nieobserwowalnymi („sztucznymi") zmiennymi będącymi liniowymi kombinacjami zmiennych oryginalnych;
- współczynniki każdej kolejnej składowej są wyznaczane tak, by zmaksymalizować całkowitą zmienność (wariancję) wyjściowego układu cech, która nie została wyjaśniona przez poprzednią składową (poprzednie składowe);

Cechy składowych głównych

Podsumowując powyższe rozważania, możemy sformułować następujące uwagi i wnioski:

- składowe główne $Z_1, ..., Z_p$ wyjściowego układu cech $X_1, ..., X_p$ są nieobserwowalnymi („sztucznymi") zmiennymi będącymi liniowymi kombinacjami zmiennych oryginalnych;
- współczynniki każdej kolejnej składowej są wyznaczane tak, by zmaksymalizować całkowitą zmienność (wariancję) wyjściowego układu cech, która nie została wyjaśniona przez poprzednią składową (poprzednie składowe);
- współczynniki składowych głównych tworzą wektory wzajemnie ortogonalne, co sprawia, że składowe są ze sobą nieskorelowane;

Cechy składowych głównych

Podsumowując powyższe rozważania, możemy sformułować następujące uwagi i wnioski:

- składowe główne $Z_1, ..., Z_p$ wyjściowego układu cech $X_1, ..., X_p$ są nieobserwowalnymi („sztucznymi") zmiennymi będącymi liniowymi kombinacjami zmiennych oryginalnych;
- współczynniki każdej kolejnej składowej są wyznaczane tak, by zmaksymalizować całkowitą zmienność (wariancję) wyjściowego układu cech, która nie została wyjaśniona przez poprzednią składową (poprzednie składowe);
- współczynniki składowych głównych tworzą wektory wzajemnie ortogonalne, co sprawia, że składowe są ze sobą nieskorelowane;
- wariancja i-tej składowej Z_i jest równa i-tej co do wielkości wartości własnej λ_i macierzy kowariancji S układu cech $X_1, ..., X_p$.

Ładunki czynnikowe (1)

Ponieważ wszystkie składowe główne $Z_1, ..., Z_p$ wyjaśniają całkowitą zmienność układu $X_1, ..., X_p$, zatem zmienność ta jest równa

$$\sum_{i=1}^{p} S^2(X_i) = \sum_{i=1}^{p} S^2(Z_i) = \sum_{i=1}^{p} \lambda_i.$$

Część wariancji (w ujęciu procentowym) wyodrębniona przez i-tą składową główną wynosi zatem

$$\frac{\lambda_i}{\lambda_1 + ... + \lambda_p} \cdot 100\%.$$

Definiuje się także tzw. **ładunki czynnikowe**, czyli współczynniki korelacji pomiędzy daną zmienną obserwowalną X_k a kolejnymi składowymi głównymi. W praktyce zatem zmienne obserwowalne o najwyższych wartościach ładunków czynnikowych dla danej składowej głównej wnoszą największy „wkład" w konstrukcję tej składowej.

Ładunki czynnikowe (2)

Jak się okazuje, wartości ładunków czynnikowych można obliczyć z następującego wzoru:

$$R(X_i, Z_j) = R(Z_j, X_i) = \frac{\sqrt{\lambda_j} a_{j,i}}{\sqrt{s_{i,i}}},$$

gdzie $R(X_i, Z_j)$ oznacza współczynnik korelacji pomiędzy i-tą zmienną a j-tą składową, natomiast $s_{i,i} = Cov(X_i, X_i) = S^2(X_i)$.

Ładunki czynnikowe (3)

Czasem analizę składowych głównych prowadzi się w oparciu o macierz współczynników korelacji R układu cech $X_1, ..., X_p$ (zamiast macierzy kowariancji S). Oczywiście, wartości własne macierzy R będą inne niż wartości własne macierzy S. Nie zmieni się natomiast procentowy udział konkretnej składowej w wyjaśnieniu całkowitej wariancji układu. Oparcie analizy na macierzy korelacji jest rekomendowane w przypadku dużych różnic pomiędzy wartościami poszczególnych zmiennych w próbie losowej. Z drugiej strony w takiej sytuacji pomóc może również standaryzacja wartości zmiennych (odjęcie średniej arytmetycznej i podzielenie przez odchylenie standardowe).

Ładunki czynnikowe (4)

Wartości ładunków czynnikowych w przypadku wykorzystania macierzy R obliczamy ze wzoru

$$R(X_i, Z_j) = R(Z_j, X_i) = \sqrt{\lambda_j} a_{j,i},$$

gdzie $i, j \in \{1, ..., p\}$.

Wartości ładunków czynnikowych (obliczanych na podstawie macierzy korelacji) podniesione do kwadratu wyrażają procentowy udział danej składowej w wyjaśnieniu zmienności (wariancji) danej zmiennej wyjściowej. Na przykład, jeśli $R(X_1, Z_1) = 0.9$, to pierwsza składowa Z_1 wyjaśnia 81% wariancji cechy X_1.

Redukcja wymiaru za pomocą składowych głównych

Kolejne składowe główne wyjaśniają coraz mniejszą część całkowitej wariancji wyjściowego układu. Z reguły już pierwsze dwie składowe wyjaśniają większość całkowitej zmienności cech $X_1, ..., X_p$. Sens analizy składowych głównych wiąże się zatem z możliwością **istotnej redukcji liczby zmiennych** służących do opisu badanego zagadnienia: zamiast p zmiennych możemy zastosować np. 2-3 składowe główne, które „wyjaśnią" 80% zmienności wyjściowego układu (20-procentowa strata informacji). Wskazana jest także właściwa interpretacja składowych jako zmiennych nieobserwowalnych, co jest niezwykle ważne w całościowej analizie badanego zjawiska. Dokonujemy tego w oparciu o wartości ładunków czynnikowych związanych ze zmiennymi i składowymi.

Kryteria wyboru ilości składowych (1)

Analiza składowych głównych pozwala na zastąpienie układu obserwowalnych wyjściowych zmiennych $X_1, ..., X_p$ układem zmiennych „sztucznych" (nieobserwowalnych) $Z_1, ..., Z_p$, zwanych składowymi głównymi, które kolejno wyjaśniają coraz mniejszą część całkowitej zmienności (wariancji) układu wyjściowego. Aby jednak dokonać redukcji wymiaru zadania, należy zrezygnować z niektórych składowych, pozostawiając tylko te, które pozwalają na relatywnie niewielką stratę wyjściowej informacji. W praktycznym użyciu są trzy podstawowe kryteria wyboru ilości składowych w analizie: kryterium procentowe, kryterium Kaisera oraz kryterium Cattella.

Kryteria wyboru ilości składowych (2)

- **Kryterium procentowe** zakłada pozostawienie tylko takiej ilości początkowych składowych głównych wyznaczonych dla układu cech, by łączny udział procentowy wyjaśnianej przez nie wariancji przekroczył pewien ustalony próg, np. 75% lub 80%.

Kryteria wyboru ilości składowych (2)

- **Kryterium procentowe** zakłada pozostawienie tylko takiej ilości początkowych składowych głównych wyznaczonych dla układu cech, by łączny udział procentowy wyjaśnianej przez nie wariancji przekroczył pewien ustalony próg, np. 75% lub 80%.
- **Kryterium Kaisera** rekomenduje pozostawienie tylko tych składowych, którym odpowiadają wartości własne macierzy korelacji większe od 1.

Kryteria wyboru ilości składowych (2)

- **Kryterium procentowe** zakłada pozostawienie tylko takiej ilości początkowych składowych głównych wyznaczonych dla układu cech, by łączny udział procentowy wyjaśnianej przez nie wariancji przekroczył pewien ustalony próg, np. 75% lub 80%.
- **Kryterium Kaisera** rekomenduje pozostawienie tylko tych składowych, którym odpowiadają wartości własne macierzy korelacji większe od 1.
- **Kryterium Cattella** opiera się na tzw. **wykresie osypiska**, na którym kolejne wartości własne macierzy kowariancji bądź korelacji przedstawia się za pomocą łamanej. W analizie uwzględniamy tylko te składowe, które odpowiadają wartościom własnym położonym na lewo od punktu, w którym rozpoczyna się łagodny spadek wartości własnych (osypisko). Sam punkt rozpoczęcia łagodnego spadku (dokładniej: składową odpowiadającą tej wartości własnej) możemy uwzględniać bądź nie.