Statystyczna Eksploracja Danych

Wykład 8 - analiza składowych głównych i skalowanie wielowymiarowe

dr inż. Julian Sienkiewicz

26 kwietnia 2019

Metody klasyfikacji bez nadzoru: charakterystyka

Metody klasyfikacji bez nadzoru: charakterystyka

- zakładamy zbiór wielowymiarowych obserwacji, leżących w przestrzeni $\mathbb{R}^p$,

Metody klasyfikacji bez nadzoru: charakterystyka

- zakładamy zbiór wielowymiarowych obserwacji, leżących w przestrzeni $\mathbb{R}^p$,
- najczęściej obserwacje nie są równomiernie rozrzucone wzdłuż wszystkich kierunków układu współrzędnych,

Metody klasyfikacji bez nadzoru: charakterystyka

- zakładamy zbiór wielowymiarowych obserwacji, leżących w przestrzeni $\mathbb{R}^p$,
- najczęściej obserwacje nie są równomiernie rozrzucone wzdłuż wszystkich kierunków układu współrzędnych,
- koncentrują się w pewnych **podprzestrzeniach** przestrzeni $\mathbb{R}^p$,

Metody klasyfikacji bez nadzoru: charakterystyka

- zakładamy zbiór wielowymiarowych obserwacji, leżących w przestrzeni $\mathbb{R}^p$,
- najczęściej obserwacje nie są równomiernie rozrzucone wzdłuż wszystkich kierunków układu współrzędnych,
- koncentrują się w pewnych **podprzestrzeniach** przestrzeni $\mathbb{R}^p$,
- kierunki, wydłuż których znajduje się większość obserwacji, nie muszą się pokrywać z osiami układu współrzędnych $\mathbb{R}^p$

Cele i ogólny opis

Cele i ogólny opis

- próba losowa pochodzi z populacji o ciągłym rozkładzie w przestrzeni $\mathbb{R}^p$ z wektorem wartości oczekiwanych $\mathbf{m}$ i macierzą kowariancji $\mathbf{S}$,

Cele i ogólny opis

- próba losowa pochodzi z populacji o ciągłym rozkładzie w przestrzeni $\mathbb{R}^p$ z wektorem wartości oczekiwanych $\mathbf{m}$ i macierzą kowariancji $\mathbf{S}$,
- $\mathbf{x}$ - dowolny wektor losowy z tej populacji ,

Cele i ogólny opis

- próba losowa pochodzi z populacji o ciągłym rozkładzie w przestrzeni $\mathbb{R}^p$ z wektorem wartości oczekiwanych $\mathbf{m}$ i macierzą kowariancji $\mathbf{S}$,
- $\mathbf{x}$ - dowolny wektor losowy z tej populacji ,
- celem analizy jest określenie **składowych głównych** wektora $\mathbf{x}$,

Cele i ogólny opis

- próba losowa pochodzi z populacji o ciągłym rozkładzie w przestrzeni $\mathbb{R}^p$ z wektorem wartości oczekiwanych $\mathbf{m}$ i macierzą kowariancji $\mathbf{S}$,
- $\mathbf{x}$ - dowolny wektor losowy z tej populacji ,
- celem analizy jest określenie **składowych głównych** wektora $\mathbf{x}$,
- interesują nas kombinacje liniowe wektora $\mathbf{x}$, czyli iloczyny skalarne $\mathbf{a}^T\mathbf{x}$, gdzie $\mathbf{a}$ jest dowolnym ustalonym wektorem z przestrzeni $\mathbb{R}^p$,

Cele i ogólny opis

- próba losowa pochodzi z populacji o ciągłym rozkładzie w przestrzeni $\mathbb{R}^p$ z wektorem wartości oczekiwanych $\mathbf{m}$ i macierzą kowariancji $\mathbf{S}$,
- $\mathbf{x}$ - dowolny wektor losowy z tej populacji ,
- celem analizy jest określenie **składowych głównych** wektora $\mathbf{x}$,
- interesują nas kombinacje liniowe wektora $\mathbf{x}$, czyli iloczyny skalarne $\mathbf{a}^T\mathbf{x}$, gdzie $\mathbf{a}$ jest dowolnym ustalonym wektorem z przestrzeni $\mathbb{R}^p$,
- wektor $\mathbf{a}$ jest jednostkowy $\|a\|^2 \equiv \mathbf{a}^T\mathbf{a} = 1$, a iloczyn $\mathbf{a}^T\mathbf{x}$ nazywamy **standaryzowaną kombinacją liniową**.

Pierwsza składowa główna

Pierwsza składowa główna powstaje powstaje poprzez znalezienie takiego jednostkowego wektora $\gamma_{(1)} \in \mathbb{R}^p$, że

$$\operatorname{var}\left(\gamma_{(1)}^T \mathbf{x}\right) = \max_{\mathbf{a} \in \mathbb{R}^p, ||\mathbf{a}||=1} \left\{ \operatorname{var}\left(\mathbf{a}^T \mathbf{x}\right) \right\}$$

Pierwsza składowa główna powstaje powstaje poprzez znalezienie takiego jednostkowego wektora $\gamma_{(1)} \in \mathbb{R}^p$, że

$$\operatorname{var}\left(\gamma_{(1)}^T \mathbf{x}\right) = \max_{\mathbf{a} \in \mathbb{R}^p, \|\mathbf{a}\|=1} \left\{ \operatorname{var}\left(\mathbf{a}^T \mathbf{x}\right) \right\}$$

Czyli szukamy standaryzowanej kombinacji liniowej o największej wariancji, inaczej — szukamy takiego kierunku $\gamma_{(1)} \in \mathbb{R}^p$, aby rzut ortogonalny wektora losowego **x** na ten kierunek dawał zmienną losową o maksymalnej wariancji.

Pierwsza składowa główna powstaje powstaje poprzez znalezienie takiego jednostkowego wektora $\gamma_{(1)} \in \mathbb{R}^p$, że

$$\operatorname{var}\left(\gamma_{(1)}^T \mathbf{x}\right) = \max_{\mathbf{a} \in \mathbb{R}^p, ||\mathbf{a}||=1} \left\{\operatorname{var}\left(\mathbf{a}^T \mathbf{x}\right)\right\}$$

Czyli szukamy standaryzowanej kombinacji liniowej o największej wariancji, inaczej — szukamy takiego kierunku $\gamma_{(1)} \in \mathbb{R}^p$, aby rzut ortogonalny wektora losowego $\mathbf{x}$ na ten kierunek dawał zmienną losową o maksymalnej wariancji.

Pierwsza składowa główna $\gamma_{(1)}$

Zmienną losową $\gamma_{(1)}^T(\mathbf{x} - \mathbf{m})$

Pierwsza składowa główna powstaje powstaje poprzez znalezienie takiego jednostkowego wektora $\gamma_{(1)} \in \mathbb{R}^p$, że

$$\operatorname{var}\left(\gamma_{(1)}^T \mathbf{x}\right) = \max_{\mathbf{a} \in \mathbb{R}^p, \|\mathbf{a}\|=1} \left\{\operatorname{var}\left(\mathbf{a}^T \mathbf{x}\right)\right\}$$

Czyli szukamy standaryzowanej kombinacji liniowej o największej wariancji, inaczej — szukamy takiego kierunku $\gamma_{(1)} \in \mathbb{R}^p$, aby rzut ortogonalny wektora losowego $\mathbf{x}$ na ten kierunek dawał zmienną losową o maksymalnej wariancji.

Pierwsza składowa główna $\gamma_{(1)}$

Zmienną losową $\gamma_{(1)}^T(\mathbf{x} - \mathbf{m})$ nazywamy **pierwszą składową główną** wektora $\mathbf{x}$. Odjęcie od $\gamma_{(1)}^T \mathbf{x}$ wartości stałej $\gamma_{(1)}^T \mathbf{m}$ nie ma, rzecz jasna, żadnego wpływu na wariancję zmiennej, natomiast później takie centrowanie okaże się przydatne.

Druga składowa główna powstaje powstaje poprzez znalezienie takiego jednostkowego wektora $\gamma_{(2)} \in \mathbb{R}^p$, że

$$\mathrm{var}\left(\gamma_{(2)}^T \mathbf{x}\right) = \max_{\mathbf{a} \in \mathbb{R}^p, ||\mathbf{a}||=1} \left\{ \mathrm{var}\left(\mathbf{a}^T \mathbf{x}\right) \right\}$$

Druga składowa główna powstaje powstaje poprzez znalezienie takiego jednostkowego wektora $\gamma_{(2)} \in \mathbb{R}^p$, że

$$\mathrm{var}\left(\gamma_{(2)}^T \mathbf{x}\right) = \max_{\mathbf{a} \in \mathbb{R}^p, ||\mathbf{a}||=1} \left\{ \mathrm{var}\left(\mathbf{a}^T \mathbf{x}\right) \right\}$$

pod warunkiem, że

$$\mathrm{E}\left[\gamma_{(1)}^T(\mathbf{x}-\mathbf{m})\gamma_{(2)}^T(\mathbf{x}-\mathbf{m})\right] = 0.$$

Druga składowa główna powstaje powstaje poprzez znalezienie takiego jednostkowego wektora $\gamma_{(2)} \in \mathbb{R}^p$, że

$$\operatorname{var}\left(\gamma_{(2)}^T \mathbf{x}\right) = \max_{\mathbf{a} \in \mathbb{R}^p, ||\mathbf{a}||=1} \left\{ \operatorname{var}\left(\mathbf{a}^T \mathbf{x}\right) \right\}$$

pod warunkiem, że

$$\mathrm{E}\left[\gamma_{(1)}^T (\mathbf{x} - \mathbf{m}) \gamma_{(2)}^T (\mathbf{x} - \mathbf{m})\right] = 0.$$

Druga składowa główna $\gamma_{(2)}$

Druga składowa główna, równa $\gamma_{(2)}^T(\mathbf{x}-\mathbf{m})$, ma największą możliwą wariancję, pod warunkiem, że jest nieskorelowana z pierwszą.

$k + 1$-sza składowa główna $\gamma_{(2)}$

$k+1$-sza składowa główna $\gamma_{(2)}$

Ogólnie, $k+1$-sza składowa główna wektora $\mathbf{x}$, równa $\gamma_{(k+1)}^T(\mathbf{x}-\mathbf{m})$, odpowiadająca standaryzowanej kombinacji liniowej $\gamma_{(k+1)}^T\mathbf{x}$, ma największą możliwą wariancję, pod warunkiem, że jest składowa ta jest nieskorelowana z wcześniejszymi składowymi głównymi: pierwszą, drugą,... i k-tą.

$k + 1$-sza składowa główna $\gamma_{(2)}$

Ogólnie, $k + 1$-sza składowa główna wektora $\mathbf{x}$, równa $\gamma_{(k+1)}^T(\mathbf{x} - \mathbf{m})$, odpowiadająca standaryzowanej kombinacji liniowej $\gamma_{(k+1)}^T\mathbf{x}$, ma największą możliwą wariancję, pod warunkiem, że jest składowa ta jest nieskorelowana z wcześniejszymi składowymi głównymi: pierwszą, drugą,... i k-tą.

Wektor $\gamma_{(i)}$, $i = 1, 2, \ldots, p$ nazywamy i-tym wektorem **ładunków** lub współczynników i-tej składowej głównej.

$k+1$-sza składowa główna $\gamma_{(2)}$

Ogólnie, $k+1$-sza składowa główna wektora $\mathbf{x}$, równa $\gamma_{(k+1)}^T(\mathbf{x}-\mathbf{m})$, odpowiadająca standaryzowanej kombinacji liniowej $\gamma_{(k+1)}^T\mathbf{x}$, ma największą możliwą wariancję, pod warunkiem, że jest składowa ta jest nieskorelowana z wcześniejszymi składowymi głównymi: pierwszą, drugą,... i k-tą.

Wektor $\gamma_{(i)}$, $i=1,2,\ldots,p$ nazywamy i-tym wektorem **ładunków** lub współczynników i-tej składowej głównej.

Kolejne wektory ładunków $\gamma_{(i)}$ wyznaczają kolejne kierunki największej zmienności (w sensie wariancji) wektora losowego $\mathbf{x}$

Twierdzenie

Twierdzenie

Niech $\mathbf{x}$ będzie wektorem losowym o wektorze wartości oczekiwanych $\mathbf{m}$ i macierzy kowariancji $\mathbf{S}$ i nich wartości własne tej macierzy, λ_i, $i = 1, \ldots, p$, spełniają warunek:

Twierdzenie

Niech $\mathbf{x}$ będzie wektorem losowym o wektorze wartości oczekiwanych $\mathbf{m}$ i macierzy kowariancji $\mathbf{S}$ i nich wartości własne tej macierzy, λ_i, $i = 1, \ldots, p$, spełniają warunek:

$$\lambda_1 \geq \lambda_2 \geq \ldots \gamma_p > 0.$$

Twierdzenie

Niech **x** będzie wektorem losowym o wektorze wartości oczekiwanych **m** i macierzy kowariancji **S** i nich wartości własne tej macierzy, λ_i, $i = 1, \ldots, p$, spełniają warunek:

$$\lambda_1 \geq \lambda_2 \geq \ldots \gamma_p > 0.$$

Wektor $\gamma_{(i)}$ i-tej składowej głównej wektora **x**

Twierdzenie

Niech $\mathbf{x}$ będzie wektorem losowym o wektorze wartości oczekiwanych $\mathbf{m}$ i macierzy kowariancji $\mathbf{S}$ i nich wartości własne tej macierzy, λ_i, $i = 1, \ldots, p$, spełniają warunek:

$$\lambda_1 \geq \lambda_2 \geq \ldots \gamma_p > 0.$$

Wektor $\gamma_{(i)}$ i-tej składowej głównej wektora $\mathbf{x}$

$$y_i \equiv \gamma_{(i)}^T(\mathbf{x} - \mathbf{m})$$

Twierdzenie

Niech **x** będzie wektorem losowym o wektorze wartości oczekiwanych **m** i macierzy kowariancji **S** i nich wartości własne tej macierzy, λ_i, $i = 1, \ldots, p$, spełniają warunek:

$$\lambda_1 \geq \lambda_2 \geq \ldots \gamma_p > 0.$$

Wektor $\gamma_{(i)}$ i-tej składowej głównej wektora **x**

$$y_i \equiv \gamma_{(i)}^T(\mathbf{x} - \mathbf{m})$$

$i = 1, \ldots, p$, jest równy i-temu wektorowi własnemu macierzy **S**, odpowiadającemu wartości własnej λ_i.

Twierdzenie

Niech $\mathbf{x}$ będzie wektorem losowym o wektorze wartości oczekiwanych $\mathbf{m}$ i macierzy kowariancji $\mathbf{S}$ i nich wartości własne tej macierzy, λ_i, $i = 1, \ldots, p$, spełniają warunek:

$$\lambda_1 \geq \lambda_2 \geq \ldots \gamma_p > 0.$$

Wektor $\gamma_{(i)}$ i-tej składowej głównej wektora $\mathbf{x}$

$$y_i \equiv \gamma_{(i)}^T(\mathbf{x} - \mathbf{m})$$

$i = 1, \ldots, p$, jest równy i-temu wektorowi własnemu macierzy $\mathbf{S}$, odpowiadającemu wartości własnej λ_i.

Macierz $\mathbf{S}$ jest symetryczna i nieujemnie określona, więc jej wartości własne są rzeczywiste i także nieujemne. Przyjmujemy dodatnią określoność tej macierzy — jeśli tak nie jest, to rozkład prawdopodobieństwa wektorów $\mathbf{x}$ jest skupiony w podprzestrzeni $\mathbb{R}^p$ i do niej można się ograniczyć.

Dowód

Załózmy, że wektory $\gamma_{(i)}$ są kolejnymi wektorami własnymi macierzy kowariancji $\mathbf{S}$. Należy wykazać, że tak określone zmienne losowe y_i są składowymi głównymi wektora $\mathbf{x}$.

Dowód

Załózmy, że wektory $\gamma_{(i)}$ są kolejnymi wektorami własnymi macierzy kowariancji $\mathbf{S}$. Należy wykazać, że tak określone zmienne losowe y_i są składowymi głównymi wektora $\mathbf{x}$. Rozkład spektralny macierzy kowariancji ma postać $\mathbf{S} = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}^T$, gdzie $\mathbf{\Gamma}$ jest macierzą ortogonalną, ktorej kolumny są kolejnymi wektorami wlasnymi macierzy $\mathbf{S}$

Dowód

Załózmy, że wektory $\gamma_{(i)}$ są kolejnymi wektorami własnymi macierzy kowariancji **S**. Należy wykazać, że tak określone zmienne losowe y_i są składowymi głównymi wektora **x**. Rozkład spektralny macierzy kowariancji ma postać $\mathbf{S} = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}^T$, gdzie $\mathbf{\Gamma}$ jest macierzą ortogonalną, ktorej kolumny są kolejnymi wektorami wlasnymi macierzy **S**

$$\mathbf{\Gamma} = [\gamma_{(1)}, \gamma_{(1)}, \ldots, \gamma_{(p)}]$$

Dowód

Załóżmy, że wektory $\gamma_{(i)}$ są kolejnymi wektorami własnymi macierzy kowariancji $\mathbf{S}$. Należy wykazać, że tak określone zmienne losowe y_i są składowymi głównymi wektora $\mathbf{x}$. Rozkład spektralny macierzy kowariancji ma postać $\mathbf{S} = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}^T$, gdzie $\mathbf{\Gamma}$ jest macierzą ortogonalną, ktorej kolumny są kolejnymi wektorami wlasnymi macierzy $\mathbf{S}$

$$\mathbf{\Gamma} = [\gamma_{(1)}, \gamma_{(1)}, \ldots, \gamma_{(p)}]$$

oraz $\mathbf{\Lambda}$ jest diagonalna macierzą dodatnich wartości własnych λ_i.

Dowód

Załózmy, że wektory $\gamma_{(i)}$ są kolejnymi wektorami własnymi macierzy kowariancji $\mathbf{S}$. Należy wykazać, że tak określone zmienne losowe y_i są składowymi głównymi wektora $\mathbf{x}$. Rozkład spektralny macierzy kowariancji ma postać $\mathbf{S} = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}^T$, gdzie $\mathbf{\Gamma}$ jest macierzą ortogonalną, ktorej kolumny są kolejnymi wektorami wlasnymi macierzy $\mathbf{S}$

$$\mathbf{\Gamma} = [\gamma_{(1)}, \gamma_{(1)}, \ldots, \gamma_{(p)}]$$

oraz $\mathbf{\Lambda}$ jest diagonalna macierzą dodatnich wartości własnych λ_i. Zauważmy, że wariancja zmiennej losowej $y_i = \gamma_{(i)}^T(\mathbf{x} - \mathbf{m})$ to

Dowód

Załózmy, że wektory $\gamma_{(i)}$ są kolejnymi wektorami własnymi macierzy kowariancji $\mathbf{S}$. Należy wykazać, że tak określone zmienne losowe y_i są składowymi głównymi wektora $\mathbf{x}$. Rozkład spektralny macierzy kowariancji ma postać $\mathbf{S} = \mathbf{\Gamma\Lambda\Gamma}^T$, gdzie $\mathbf{\Gamma}$ jest macierzą ortogonalną, ktorej kolumny są kolejnymi wektorami wlasnymi macierzy $\mathbf{S}$

$$\mathbf{\Gamma} = [\gamma_{(1)}, \gamma_{(1)}, \ldots, \gamma_{(p)}]$$

oraz $\mathbf{\Lambda}$ jest diagonalna macierzą dodatnich wartości własnych λ_i. Zauważmy, że wariancja zmiennej losowej $y_i = \gamma_{(i)}^T(\mathbf{x} - \mathbf{m})$ to

$$\mathrm{var}(y_i) = \gamma_{(i)}^T \mathbf{S} \gamma_{(i)}$$

Dowód

Załózmy, że wektory $\gamma_{(i)}$ są kolejnymi wektorami własnymi macierzy kowariancji **S**. Należy wykazać, że tak określone zmienne losowe y_i są składowymi głównymi wektora **x**. Rozkład spektralny macierzy kowariancji ma postać $\mathbf{S} = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}^T$, gdzie $\mathbf{\Gamma}$ jest macierzą ortogonalną, ktorej kolumny są kolejnymi wektorami wlasnymi macierzy **S**

$$\mathbf{\Gamma} = [\gamma_{(1)}, \gamma_{(1)}, \ldots, \gamma_{(p)}]$$

oraz $\mathbf{\Lambda}$ jest diagonalna macierzą dodatnich wartości własnych λ_i. Zauważmy, że wariancja zmiennej losowej $y_i = \gamma_{(i)}^T(\mathbf{x} - \mathbf{m})$ to

$$\mathrm{var}(y_i) = \gamma_{(i)}^T \mathbf{S} \gamma_{(i)} = \gamma_{(i)}^T \left(\sum_{j=1}^{p} \lambda_j \gamma_{(j)} \gamma_{(j)}^T \right) \gamma_{(i)}$$

Dowód

Załózmy, że wektory $\gamma_{(i)}$ są kolejnymi wektorami własnymi macierzy kowariancji $\mathbf{S}$. Należy wykazać, że tak określone zmienne losowe y_i są składowymi głównymi wektora $\mathbf{x}$. Rozkład spektralny macierzy kowariancji ma postać $\mathbf{S} = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}^T$, gdzie $\mathbf{\Gamma}$ jest macierzą ortogonalną, ktorej kolumny są kolejnymi wektorami wlasnymi macierzy $\mathbf{S}$

$$\mathbf{\Gamma} = [\gamma_{(1)}, \gamma_{(1)}, \ldots, \gamma_{(p)}]$$

oraz $\mathbf{\Lambda}$ jest diagonalna macierzą dodatnich wartości własnych λ_i. Zauważmy, że wariancja zmiennej losowej $y_i = \gamma_{(i)}^T(\mathbf{x} - \mathbf{m})$ to

$$\mathrm{var}(y_i) = \gamma_{(i)}^T \mathbf{S} \gamma_{(i)} = \gamma_{(i)}^T \left(\sum_{j=1}^{p} \lambda_j \gamma_{(j)} \gamma_{(j)}^T \right) \gamma_{(i)} = \lambda_i$$

Dowód

Załózmy, że wektory $\gamma_{(i)}$ są kolejnymi wektorami własnymi macierzy kowariancji $\mathbf{S}$. Należy wykazać, że tak określone zmienne losowe y_i są składowymi głównymi wektora $\mathbf{x}$. Rozkład spektralny macierzy kowariancji ma postać $\mathbf{S} = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}^T$, gdzie $\mathbf{\Gamma}$ jest macierzą ortogonalną, ktorej kolumny są kolejnymi wektorami wlasnymi macierzy $\mathbf{S}$

$$\mathbf{\Gamma} = [\gamma_{(1)}, \gamma_{(1)}, \ldots, \gamma_{(p)}]$$

oraz $\mathbf{\Lambda}$ jest diagonalna macierzą dodatnich wartości własnych λ_i. Zauważmy, że wariancja zmiennej losowej $y_i = \gamma_{(i)}^T(\mathbf{x} - \mathbf{m})$ to

$$\mathrm{var}(y_i) = \gamma_{(i)}^T \mathbf{S} \gamma_{(i)} = \gamma_{(i)}^T \left(\sum_{j=1}^{p} \lambda_j \gamma_{(j)} \gamma_{(j)}^T \right) \gamma_{(i)} = \lambda_i$$

oraz, że

$$\mathrm{cov}(y_i, y_j) = \mathrm{E}\left[\gamma_{(i)}^T (\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^T \gamma_{(j)} \right] = \gamma_{(i)}^T \boldsymbol{S} \gamma(j) = 0$$

Dowód

Założmy, że wektory $\gamma_{(i)}$ są kolejnymi wektorami własnymi macierzy kowariancji **S**. Należy wykazać, że tak określone zmienne losowe y_i są składowymi głównymi wektora **x**. Rozkład spektralny macierzy kowariancji ma postać $\mathbf{S} = \mathbf{\Gamma \Lambda \Gamma}^T$, gdzie $\mathbf{\Gamma}$ jest macierzą ortogonalną, ktorej kolumny są kolejnymi wektorami wlasnymi macierzy **S**

$$\mathbf{\Gamma} = [\gamma_{(1)}, \gamma_{(1)}, \ldots, \gamma_{(p)}]$$

oraz $\mathbf{\Lambda}$ jest diagonalna macierzą dodatnich wartości własnych λ_i. Zauważmy, że wariancja zmiennej losowej $y_i = \gamma_{(i)}^T(\mathbf{x} - \mathbf{m})$ to

$$\mathrm{var}(y_i) = \gamma_{(i)}^T \mathbf{S} \gamma_{(i)} = \gamma_{(i)}^T \left(\sum_{j=1}^{p} \lambda_j \gamma_{(j)} \gamma_{(j)}^T \right) \gamma_{(i)} = \lambda_i$$

oraz, że

$$\mathrm{cov}(y_i, y_j) = \mathrm{E}\left[\gamma_{(i)}^T (\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^T \gamma_{(j)} \right] = \gamma_{(i)}^T \boldsymbol{S} \gamma(j) = 0$$

Dowód (cd)

Teraz musimy wykazać, że pierwsza składowa główna spełnia równość $y_1 = \gamma_{(1)}^T(\mathbf{x} - \mathbf{m})$, gdzie $\gamma_{(1)}$ jest pierwszym wektorem własnym macierzy $\mathbf{S}$.

Dowód (cd)

Teraz musimy wykazać, że pierwsza składowa główna spełnia równość $y_1 = \gamma_{(1)}^T(\mathbf{x} - \mathbf{m})$, gdzie $\gamma_{(1)}$ jest pierwszym wektorem własnym macierzy $\mathbf{S}$. Rozważmy dowolną kombinację liniową $\mathbf{x}$, tzn. $\mathbf{a}^T\mathbf{x}$ ($\|a\| = 1$). Wektor $\mathbf{a}$ można zapisać w bazie tworzonej przez wektory własne jako:

Dowód (cd)

Teraz musimy wykazać, że pierwsza składowa główna spełnia równość $y_1 = \gamma_{(1)}^T(\mathbf{x} - \mathbf{m})$, gdzie $\gamma_{(1)}$ jest pierwszym wektorem własnym macierzy $\mathbf{S}$. Rozważmy dowolną kombinację liniową $\mathbf{x}$, tzn. $\mathbf{a}^T\mathbf{x}$ ($\|a\| = 1$). Wektor $\mathbf{a}$ można zapisać w bazie tworzonej przez wektory własne jako:

$$\mathbf{a} = c_1\gamma_{(1)} + \ldots + c_p\gamma_{(p)},$$

Dowód (cd)

Teraz musimy wykazać, że pierwsza składowa główna spełnia równość $y_1 = \gamma_{(1)}^T(\mathbf{x} - \mathbf{m})$, gdzie $\gamma_{(1)}$ jest pierwszym wektorem własnym macierzy $\mathbf{S}$. Rozważmy dowolną kombinację liniową $\mathbf{x}$, tzn. $\mathbf{a}^T\mathbf{x}$ ($\|a\| = 1$). Wektor $\mathbf{a}$ można zapisać w bazie tworzonej przez wektory własne jako:

$$\mathbf{a} = c_1\gamma_{(1)} + \ldots + c_p\gamma_{(p)},$$

gdzie $c_1, \ldots, c_p \in \mathbb{R}$ i

$$\sum_{i=1}^{p} c_i^2 = 1$$

Stąd

Dowód (cd)

Teraz musimy wykazać, że pierwsza składowa główna spełnia równość $y_1 = \gamma_{(1)}^T(\mathbf{x} - \mathbf{m})$, gdzie $\gamma_{(1)}$ jest pierwszym wektorem własnym macierzy $\mathbf{S}$. Rozważmy dowolną kombinację liniową $\mathbf{x}$, tzn. $\mathbf{a}^T\mathbf{x}$ ($\|a\| = 1$). Wektor $\mathbf{a}$ można zapisać w bazie tworzonej przez wektory własne jako:

$$\mathbf{a} = c_1\gamma_{(1)} + \ldots + c_p\gamma_{(p)},$$

gdzie $c_1, \ldots, c_p \in \mathbb{R}$ i

$$\sum_{i=1}^{p} c_i^2 = 1$$

Stąd

$$\mathrm{var}(\mathbf{a}^T\mathbf{x}) = \mathbf{a}^T\mathbf{S}\mathbf{a} = \mathbf{a}^T\left(\sum_{i=1}^{p}\lambda_i\gamma_{(i)}\gamma_{(i)}^T\right)\mathbf{a} = \sum_{i=1}^{p}\lambda_i c_i^2$$

Dowód (cd)

Teraz musimy wykazać, że pierwsza składowa główna spełnia równość $y_1 = \gamma_{(1)}^T(\mathbf{x} - \mathbf{m})$, gdzie $\gamma_{(1)}$ jest pierwszym wektorem własnym macierzy $\mathbf{S}$. Rozważmy dowolną kombinację liniową $\mathbf{x}$, tzn. $\mathbf{a}^T\mathbf{x}$ ($\|a\| = 1$). Wektor $\mathbf{a}$ można zapisać w bazie tworzonej przez wektory własne jako:

$$\mathbf{a} = c_1\gamma_{(1)} + \ldots + c_p\gamma_{(p)},$$

gdzie $c_1, \ldots, c_p \in \mathbb{R}$ i

$$\sum_{i=1}^{p} c_i^2 = 1$$

Stąd

$$\mathrm{var}(\mathbf{a}^T\mathbf{x}) = \mathbf{a}^T\mathbf{S}\mathbf{a} = \mathbf{a}^T\left(\sum_{i=1}^{p}\lambda_i\gamma_{(i)}\gamma_{(i)}^T\right)\mathbf{a} = \sum_{i=1}^{p}\lambda_i c_i^2$$

Ale, ponieważ zachodzi $\lambda_1 \geq \ldots \geq \lambda_p > 0$ i $\sum_{j=k}^{p} c_j^2 = 1$, to wariancja osiąga wartość maksymalną, równą λ_1, gdy

$$c_1 = 1 \text{ i } c_2 = \ldots = c_p = 0$$

Dowód (cd)

Teraz musimy wykazać, że pierwsza składowa główna spełnia równość $y_1 = \gamma_{(1)}^T(\mathbf{x} - \mathbf{m})$, gdzie $\gamma_{(1)}$ jest pierwszym wektorem własnym macierzy $\mathbf{S}$. Rozważmy dowolną kombinację liniową $\mathbf{x}$, tzn. $\mathbf{a}^T\mathbf{x}$ ($\|a\| = 1$). Wektor $\mathbf{a}$ można zapisać w bazie tworzonej przez wektory własne jako:

$$\mathbf{a} = c_1\gamma_{(1)} + \ldots + c_p\gamma_{(p)},$$

gdzie $c_1, \ldots, c_p \in \mathbb{R}$ i

$$\sum_{i=1}^{p} c_i^2 = 1$$

Stąd

$$\mathrm{var}(\mathbf{a}^T\mathbf{x}) = \mathbf{a}^T\mathbf{S}\mathbf{a} = \mathbf{a}^T\left(\sum_{i=1}^{p}\lambda_i\gamma_{(i)}\gamma_{(i)}^T\right)\mathbf{a} = \sum_{i=1}^{p}\lambda_i c_i^2$$

Ale, ponieważ zachodzi $\lambda_1 \geq \ldots \geq \lambda_p > 0$ i $\sum_{j=k}^{p} c_j^2 = 1$, to wariancja osiąga wartość maksymalną, równą λ_1, gdy

$$c_1 = 1 \text{ i } c_2 = \ldots = c_p = 0$$

Czyli pierwsza składowa główna jest dana przez pierwszy wektor własny $\mathbf{S}$.

Dowód (cd)

Musimy jeszcze wykazać, że k-ta składowa główna ($k \geq 2$) spełnia równość $y_k = \gamma_{(k)}^T(\mathbf{x} - \mathbf{m})$, gdzie $\gamma_{(k)}$ jest k-tym wektorem własnym macierzy $\mathbf{S}$.

Dowód (cd)

Musimy jeszcze wykazać, że k-ta składowa główna ($k \geq 2$) spełnia równość $y_k = \gamma_{(k)}^T(\mathbf{x} - \mathbf{m})$, gdzie $\gamma_{(k)}$ jest k-tym wektorem własnym macierzy $\mathbf{S}$. Rozważmy dowolną kombinację liniową $\mathbf{a}^T\mathbf{x}$, niesorelowaną z $\gamma_{(i)}^T\mathbf{x}$, $i < k$. Mamy:

Dowód (cd)

Musimy jeszcze wykazać, że k-ta składowa główna ($k \geq 2$) spełnia równość $y_k = \gamma_{(k)}^T(\mathbf{x} - \mathbf{m})$, gdzie $\gamma_{(k)}$ jest k-tym wektorem własnym macierzy $\mathbf{S}$. Rozważmy dowolną kombinację liniową $\mathbf{a}^T\mathbf{x}$, niesorelowaną z $\gamma_{(i)}^T\mathbf{x}$, $i < k$. Mamy:

$$\mathrm{E}\left[\mathbf{a}^T(\mathbf{x} - \mathbf{m})\gamma_{(i)}^T(\mathbf{x} - \mathbf{m})\right] = \mathbf{a}^T\mathbf{S}\gamma_{(i)} = \mathbf{a}^t\left(\sum_{i=1}^{p}\lambda_i\gamma_{(i)}\gamma_{(i)}^T\right)\gamma_{(i)} = \lambda_i c_i =$$

Dowód (cd)

Musimy jeszcze wykazać, że k-ta składowa główna ($k \geq 2$) spełnia równość $y_k = \gamma_{(k)}^T(\mathbf{x} - \mathbf{m})$, gdzie $\gamma_{(k)}$ jest k-tym wektorem własnym macierzy $\mathbf{S}$. Rozważmy dowolną kombinację liniową $\mathbf{a}^T\mathbf{x}$, niesorelowaną z $\gamma_{(i)}^T\mathbf{x}$, $i < k$. Mamy:

$$\mathrm{E}\left[\mathbf{a}^T(\mathbf{x} - \mathbf{m})\gamma_{(i)}^T(\mathbf{x} - \mathbf{m})\right] = \mathbf{a}^T\mathbf{S}\gamma_{(i)} = \mathbf{a}^t\left(\sum_{i=1}^{p}\lambda_i\gamma_{(i)}\gamma_{(i)}^T\right)\gamma_{(i)} = \lambda_i c_i = 0,$$

Dowód (cd)

Musimy jeszcze wykazać, że k-ta składowa główna ($k \geq 2$) spełnia równość $y_k = \gamma_{(k)}^T(\mathbf{x} - \mathbf{m})$, gdzie $\gamma_{(k)}$ jest k-tym wektorem własnym macierzy $\mathbf{S}$. Rozważmy dowolną kombinację liniową $\mathbf{a}^T\mathbf{x}$, niesorelowaną z $\gamma_{(i)}^T\mathbf{x}$, $i < k$. Mamy:

$$\mathrm{E}\left[\mathbf{a}^T(\mathbf{x} - \mathbf{m})\gamma_{(i)}^T(\mathbf{x} - \mathbf{m})\right] = \mathbf{a}^T\mathbf{S}\gamma_{(i)} = \mathbf{a}^t\left(\sum_{i=1}^{p} \lambda_i \gamma_{(i)}\gamma_{(i)}^T\right)\gamma_{(i)} = \lambda_i c_i = 0,$$

Ale, wszystkie wartości własne $\lambda_i > 0$, czyli $c_i = 0$ dla $i < k$. Stąd

$$\mathrm{var}(\mathbf{a}^T\mathbf{x}) = \sum_{j=k}^{p} \lambda_j c_j^2$$

Dowód (cd)

Musimy jeszcze wykazać, że k-ta składowa główna ($k \geq 2$) spełnia równość $y_k = \gamma_{(k)}^T(\mathbf{x} - \mathbf{m})$, gdzie $\gamma_{(k)}$ jest k-tym wektorem własnym macierzy $\mathbf{S}$. Rozważmy dowolną kombinację liniową $\mathbf{a}^T\mathbf{x}$, niesorelowaną z $\gamma_{(i)}^T\mathbf{x}$, $i < k$. Mamy:

$$\mathrm{E}\left[\mathbf{a}^T(\mathbf{x} - \mathbf{m})\gamma_{(i)}^T(\mathbf{x} - \mathbf{m})\right] = \mathbf{a}^T\mathbf{S}\gamma_{(i)} = \mathbf{a}^t\left(\sum_{i=1}^{p} \lambda_i \gamma_{(i)}\gamma_{(i)}^T\right)\gamma_{(i)} = \lambda_i c_i = 0,$$

Ale, wszystkie wartości własne $\lambda_i > 0$, czyli $c_i = 0$ dla $i < k$. Stąd

$$\mathrm{var}(\mathbf{a}^T\mathbf{x}) = \sum_{j=k}^{p} \lambda_j c_j^2$$

Stąd

Dowód (cd)

Musimy jeszcze wykazać, że k-ta składowa główna ($k \geq 2$) spełnia równość $y_k = \gamma_{(k)}^T(\mathbf{x} - \mathbf{m})$, gdzie $\gamma_{(k)}$ jest k-tym wektorem własnym macierzy $\mathbf{S}$. Rozważmy dowolną kombinację liniową $\mathbf{a}^T\mathbf{x}$, niesorelowaną z $\gamma_{(i)}^T\mathbf{x}$, $i < k$. Mamy:

$$\mathrm{E}\left[\mathbf{a}^T(\mathbf{x} - \mathbf{m})\gamma_{(i)}^T(\mathbf{x} - \mathbf{m})\right] = \mathbf{a}^T\mathbf{S}\gamma_{(i)} = \mathbf{a}^t\left(\sum_{i=1}^{p}\lambda_i\gamma_{(i)}\gamma_{(i)}^T\right)\gamma_{(i)} = \lambda_i c_i = 0,$$

Ale, wszystkie wartości własne $\lambda_i > 0$, czyli $c_i = 0$ dla $i < k$. Stąd

$$\mathrm{var}(\mathbf{a}^T\mathbf{x}) = \sum_{j=k}^{p}\lambda_j c_j^2$$

Stąd

$$\mathrm{var}(\mathbf{a}^T\mathbf{x}) = \mathbf{a}^T\mathbf{S}\mathbf{a} = \mathbf{a}^T\left(\sum_{i=1}^{p}\lambda_i\gamma_{(i)}\gamma_{(i)}^T\right)\mathbf{a} = \sum_{i=1}^{p}\lambda_i c_i^2$$

Dowód (cd)

Musimy jeszcze wykazać, że k-ta składowa główna ($k \geq 2$) spełnia równość $y_k = \gamma_{(k)}^T(\mathbf{x} - \mathbf{m})$, gdzie $\gamma_{(k)}$ jest k-tym wektorem własnym macierzy $\mathbf{S}$. Rozważmy dowolną kombinację liniową $\mathbf{a}^T\mathbf{x}$, niesorelowaną z $\gamma_{(i)}^T\mathbf{x}$, $i < k$. Mamy:

$$\mathrm{E}\left[\mathbf{a}^T(\mathbf{x} - \mathbf{m})\gamma_{(i)}^T(\mathbf{x} - \mathbf{m})\right] = \mathbf{a}^T\mathbf{S}\gamma_{(i)} = \mathbf{a}^t\left(\sum_{i=1}^{p}\lambda_i\gamma_{(i)}\gamma_{(i)}^T\right)\gamma_{(i)} = \lambda_i c_i = 0,$$

Ale, wszystkie wartości własne $\lambda_i > 0$, czyli $c_i = 0$ dla $i < k$. Stąd

$$\mathrm{var}(\mathbf{a}^T\mathbf{x}) = \sum_{j=k}^{p}\lambda_j c_j^2$$

Stąd

$$\mathrm{var}(\mathbf{a}^T\mathbf{x}) = \mathbf{a}^T\mathbf{S}\mathbf{a} = \mathbf{a}^T\left(\sum_{i=1}^{p}\lambda_i\gamma_{(i)}\gamma_{(i)}^T\right)\mathbf{a} = \sum_{i=1}^{p}\lambda_i c_i^2$$

Ponieważ zachodzi $\lambda_1 \geq \ldots \geq \lambda_p > 0$ i $\sum_{j=k}^{p} c_j^2 = 1$ i $\mathrm{var}(y_j) = \lambda_j$, to otrzymujemy tezę twierdzenia. **CBDN.**

Wszystkie p składowe główne tworzą wektor $\mathbf{y}$ postaci

$$\mathbf{y} = \Gamma^T(\mathbf{x} - \mathbf{m}) \tag{1}$$

Wszystkie p składowe główne tworzą wektor $\mathbf{y}$ postaci

$$\mathbf{y} = \boldsymbol{\Gamma}^T(\mathbf{x} - \mathbf{m}) \tag{1}$$

czyli jest to przekształcenie wektora losowego $\mathbf{x}$ w wektor $\mathbf{y}$, polegające kolejno na

1. przesunięciu wektora $\mathbf{x}$ o jego wartość oczekiwaną (scentrowania wektora),
2. liniowym przekształceniu scentrowanego wektora za pomocą macierzy ortogonalnej $\boldsymbol{\Gamma}^T$ (geometrycznie jest obrócenie oryginalnego układu współprzędnych o pewien kąt)

Wszystkie p składowe główne tworzą wektor **y** postaci

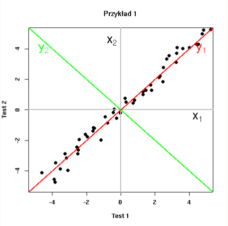
$$\mathbf{y} = \boldsymbol{\Gamma}^T(\mathbf{x} - \mathbf{m}) \tag{1}$$

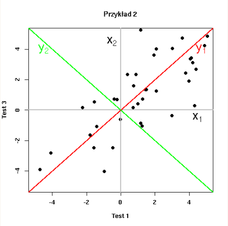
czyli jest to przekształcenie wektora losowego **x** w wektor **y**, polegające kolejno na

1. przesunięciu wektora **x** o jego wartość oczekiwaną (scentrowania wektora),
2. liniowym przekształceniu scentrowanego wektora za pomocą macierzy ortogonalnej $\boldsymbol{\Gamma}^T$ (geometrycznie jest obrócenie oryginalnego układu współprzędnych o pewien kąt)

Osie nowego układu współrzednych , wyznaczone przez wektory ładunków $\gamma_{(i)}$ są tak dobrane, aby maksymalizować wariancje rzutów oryginalnych wektorów losowych na te osie, pod warunkiem, że kolejne rzuty nie są skorelowane z wcześniejszymi.

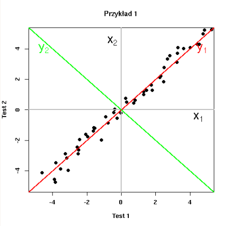
Przykład

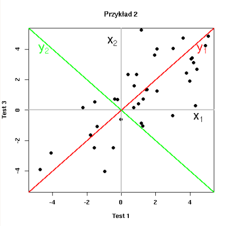




Prosty, dwuwymiarowy przykład: obserwacją była para punktów zdobytych przez studenta w dwóch testach — w pierwszej grupie był to test z pilotażu oraz test zbiorczy z jęz. polskiego i historii. W drugiej — ten sam test z pilotażu oraz zbiorczy z matematyki i fizyki.

Przykład

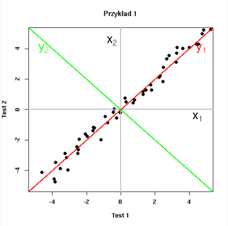


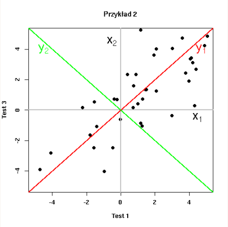


Prosty, dwuwymiarowy przykład: obserwacją była para punktów zdobytych przez studenta w dwóch testach — w pierwszej grupie był to test z pilotażu oraz test zbiorczy z jęz. polskiego i historii. W drugiej — ten sam test z pilotażu oraz zbiorczy z matematyki i fizyki.

- dla danych o dużym wymiarze ogromnego znaczenia nabiera możliwość zredukowania wymiaru, z małą utratą informacji,

Przykład

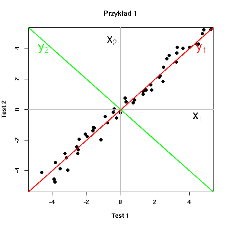


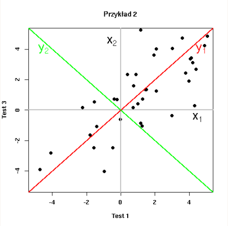


Prosty, dwuwymiarowy przykład: obserwacją była para punktów zdobytych przez studenta w dwóch testach — w pierwszej grupie był to test z pilotażu oraz test zbiorczy z jęz. polskiego i historii. W drugiej — ten sam test z pilotażu oraz zbiorczy z matematyki i fizyki.

- dla danych o dużym wymiarze ogromnego znaczenia nabiera możliwość zredukowania wymiaru, z małą utratą informacji,
- tutaj: na górnym rysunku widać, że zmienność w kierunku y_1 jest znaczenie większa niż w kierunku y_2 — dane dwuwymiarowe można swobodnie zastąpić jednowymiarowymi,

Przykład

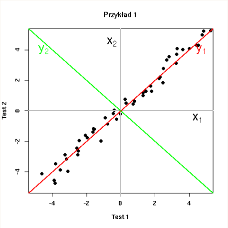


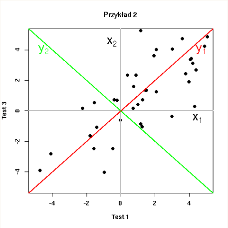


Prosty, dwuwymiarowy przykład: obserwacją była para punktów zdobytych przez studenta w dwóch testach — w pierwszej grupie był to test z pilotażu oraz test zbiorczy z jęz. polskiego i historii. W drugiej — ten sam test z pilotażu oraz zbiorczy z matematyki i fizyki.

- dla danych o dużym wymiarze ogromnego znaczenia nabiera możliwość zredukowania wymiaru, z małą utratą informacji,
- tutaj: na górnym rysunku widać, że zmienność w kierunku y_1 jest znaczenie większa niż w kierunku y_2 — dane dwuwymiarowe można swobodnie zastąpić jednowymiarowymi,
- taka redukcja w drugim przypadku jest nieuzasadniona

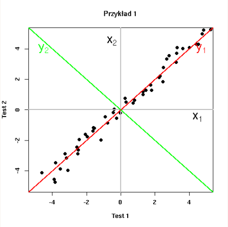
Przykład

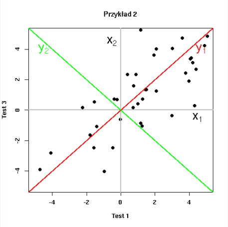




- drugi istotny aspekt, to ewentualna możliwość zinterpretwoania otrzymanych nowych zmiennych (**reifikacja**),

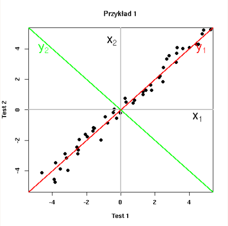
Przykład

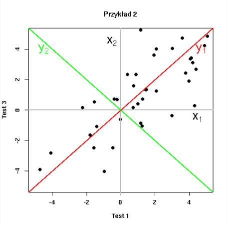




- drugi istotny aspekt, to ewentualna możliwość zinterpretwoania otrzymanych nowych zmiennych (**reifikacja**),
- składowa y_2 opisuje **kontrast** pomiędzy zmiennymi oryginalnymi (matematycznie: pewne ładunki składowej głównej są istotnie dodatnie, inne istotnie ujemne),

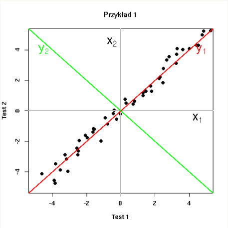
Przykład

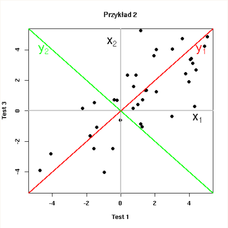




- drugi istotny aspekt, to ewentualna możliwość zinterpretwoania otrzymanych nowych zmiennych (**reifikacja**),
- składowa y_2 opisuje **kontrast** pomiędzy zmiennymi oryginalnymi (matematycznie: pewne ładunki składowej głównej są istotnie dodatnie, inne istotnie ujemne),
- tutaj: mały kontrast w pierwszym przypadku, duży w drugim

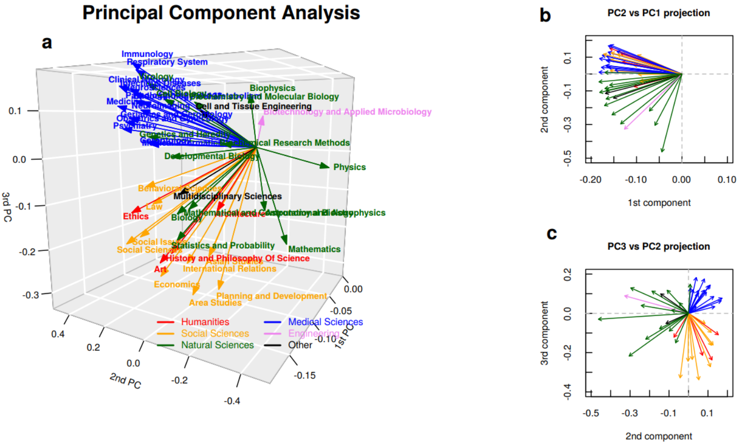
Przykład





- drugi istotny aspekt, to ewentualna możliwość zinterpretwoania otrzymanych nowych zmiennych (**reifikacja**),
- składowa y_2 opisuje **kontrast** pomiędzy zmiennymi oryginalnymi (matematycznie: pewne ładunki składowej głównej są istotnie dodatnie, inne istotnie ujemne),
- tutaj: mały kontrast w pierwszym przypadku, duży w drugim
- można się pokusić o zinterpretowanie y_2 z Przykładu 2 jako źdolność myślenia abstrakcyjnego"(niekoniecznie sprzyjająca pilotażowi), natomiast y_1 z Przykładu 1 jako "wobraźni"(sprzyjająca zarówno dobremu lataniu jak i humanistyce)

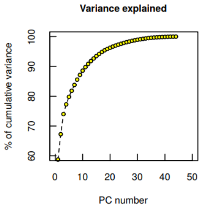
Przykład - współpraca naukowców



Przykład - współpraca naukowców

e

d



A w jaki sposób określić **możliwość** dokonania redukcji wymiaru?

A w jaki sposób określić **możliwość** dokonania redukcji wymiaru?
Na szczęście kryterium jest dość proste, dzieki temu, że dla macierzy symetrycznej jej ślad jest równy sumie wartości własnych tej macierzy. Zatem suma wszytskich wartości własnych macierzy kowariancji $\mathbf{S}$ jest równa wariancji poszczególnych współprzędnych wektora $\mathbf{x}$. Czyli wielkość

A w jaki sposób określić **możliwość** dokonania redukcji wymiaru?
Na szczęście kryterium jest dość proste, dzieki temu, że dla macierzy symetrycznej jej ślad jest równy sumie wartości własnych tej macierzy. Zatem suma wszytskich wartości własnych macierzy kowariancji **S** jest równa wariancji poszczególnych współprzędnych wektora **x**. Czyli wielkość

$$\frac{\lambda_1 + \ldots + \lambda_k}{\lambda_1 + \ldots + \lambda_p} 100\%$$

A w jaki sposób określić **możliwość** dokonania redukcji wymiaru?
Na szczęście kryterium jest dość proste, dzieki temu, że dla macierzy symetrycznej jej ślad jest równy sumie wartości własnych tej macierzy. Zatem suma wszytskich wartości własnych macierzy kowariancji **S** jest równa wariancji poszczególnych współprzędnych wektora **x**. Czyli wielkość

$$\frac{\lambda_1 + \ldots + \lambda_k}{\lambda_1 + \ldots + \lambda_p} 100\%$$

$k \leq p$ wyraża **procent zmienności** wektora losowego **x** wyjasniony przez k pierwszych składowych głównych, gdy przez zmienność całkowitą rozumie się sumę wariancji.

Niech d_{ij}, $i, j, = 1, \ldots, n$ będą odległosciami euklidesowymi między obserwacjami $\mathbf{x}_i$ i $\mathbf{x}_j$ w przestrzeni $\mathbb{R}^p$. Zdanie polega na znalezieniu takiej podprzestrzeni $\mathbb{R}^r$ o wymiarze r, by odległości euklidesowe $\hat{d}_{ij}$ między rzutami obserwacji na tę podprzestrzeń minimalizowały sumę

Niech d_{ij}, $i, j, = 1, \ldots, n$ będą odległosciami euklidesowymi między obserwacjami $\mathbf{x}_i$ i $\mathbf{x}_j$ w przestrzeni $\mathbb{R}^p$. Zdanie polega na znalezieniu takiej podprzestrzeni $\mathbb{R}^r$ o wymiarze r, by odległości euklidesowe $\hat{d}_{ij}$ między rzutami obserwacji na tę podprzestrzeń minimalizowały sumę

$$V = \sum_{i=1}^{n} \sum_{j=1}^{n} \left(d_{ij}^2 - \hat{d}_{ij}^2 \right)$$

Niech d_{ij}, $i, j, = 1, \ldots, n$ będą odległosciami euklidesowymi między obserwacjami $\mathbf{x}_i$ i $\mathbf{x}_j$ w przestrzeni $\mathbb{R}^p$. Zdanie polega na znalezieniu takiej podprzestrzeni $\mathbb{R}^r$ o wymiarze r, by odległości euklidesowe $\hat{d}_{ij}$ między rzutami obserwacji na tę podprzestrzeń minimalizowały sumę

$$V = \sum_{i=1}^{n} \sum_{j=1}^{n} \left(d_{ij}^2 - \hat{d}_{ij}^2 \right)$$

Okazuje się, że przestrzeń określana przez r pierwszych składowych głównych jest rozwiązaniem zadania — czyli podana podprzestrzeń najlepiej **odtwarza** oryginalną konfigurację obserwacji.

Niech d_{ij}, $i, j, = 1, \ldots, n$ będą odległosciami euklidesowymi między obserwacjami $\mathbf{x}_i$ i $\mathbf{x}_j$ w przestrzeni $\mathbb{R}^p$. Zdanie polega na znalezieniu takiej podprzestrzeni $\mathbb{R}^r$ o wymiarze r, by odległości euklidesowe $\hat{d}_{ij}$ między rzutami obserwacji na tę podprzestrzeń minimalizowały sumę

$$V = \sum_{i=1}^{n} \sum_{j=1}^{n} \left(d_{ij}^2 - \hat{d}_{ij}^2 \right)$$

Okazuje się, że przestrzeń określana przez r pierwszych składowych głównych jest rozwiązaniem zadania — czyli podana podprzestrzeń najlepiej **odtwarza** oryginalną konfigurację obserwacji.

Odtworzenie konfiguracji punktów w przestrzeni o wymiarze mniejszym od oryginalnego ma wielkie znaczenie, biorąc pod uwagę postęp wizualizacji danych dwu- i trzywymiarowych.

Podobnie istotne jest uzyskanie przedstawienia danych *niekoniecznie* ilościowych w przestrzeni euklidesowej o małym wymiarze. Taką możliwość stwarza zastąpienie miary odległości miarami odmienności i skorzystanie z jednego z algorytmów **skalowania wielowymiarowego**.

Podobnie istotne jest uzyskanie przedstawienia danych *niekoniecznie* ilościowych w przestrzeni euklidesowej o małym wymiarze. Taką możliwość stwarza zastąpienie miary odległości miarami odmienności i skorzystanie z jednego z algorytmów **skalowania wielowymiarowego**.

Takie skalowanie jest szczególnie istotne, gdy **macierz odmienności** jest **wyjściowym** zbiorem danych, jakim dysponujemy. W niektórych badaniach w ogóle nie mamy do czynienia z wektorem obserwacji, a tylko z odmiennościami między obiektami.
Przykład - badanie bliskości brzmienia głosek. Bliskość brzmienia głosek (np. *s* i *z* w jęz. polskim) można zmierzyć wypowiadając raz jedną, raz drugą głoskę w obecności kolejnych osób i wyliczając ułamek wzięcia jednej głoski za drugą. Przeprowadziwszy takie badanie dla różnych par głosek, uzyskuje się macierz podobieństwa

Zadanie...

Czy mając macierz odmienności $\mathbf{d}_{ij}$, możliwe jest znalezienie przestrzeni $\mathbb{R}^s$ oraz takiej konfiguracji punktów w tej przestrzeni, że odległości euklidesowe pomiędzy nimi $\hat{d}_{ij}$ dokładnie odtwarzają macierz $\mathbf{d}_{ij}$?

Zadanie...

Czy mając macierz odmienności $\mathbf{d}_{ij}$, możliwe jest znalezienie przestrzeni $\mathbb{R}^s$ oraz takiej konfiguracji punktów w tej przestrzeni, że odległości euklidesowe pomiędzy nimi $\hat{d}_{ij}$ dokładnie odtwarzają macierz $\mathbf{d}_{ij}$?

... i dalej

Jeżeli tak, to, czy mając przestrzeń $\mathbb{R}^s$ o podanej własności, można dla każdej liczby naturalnej $u < s$ wyznaczyć przestrzeń $\mathbb{R}^u$ oraz taką konfigurację w tej przestrzeni, że odległości euklidesowe miedzy nimi minimalizują wskaźnik V?

Rozwiązanie

Rozwiązanie

- zakładamy, że symetryczna macierz odmienności $\mathbf{d}_{ij}$ o wymiarze (n, n) spełnia nierówność trójkąta,

Rozwiązanie

- zakładamy, że symetryczna macierz odmienności $\mathbf{d}_{ij}$ o wymiarze (n, n) spełnia nierówność trójkąta,
- tworzymy macierz Γ tego samego wymiaru o elementach $\gamma_{ij} = -\frac{1}{2} d_{ij}^2$,

Rozwiązanie

- zakładamy, że symetryczna macierz odmienności $\mathbf{d}_{ij}$ o wymiarze (n, n) spełnia nierówność trójkąta,
- tworzymy macierz $\mathbf{\Gamma}$ tego samego wymiaru o elementach $\gamma_{ij} = -\frac{1}{2}d_{ij}^2$,
- od każdego elementu γ_{ij} odejmujemy średnią wartość elementów i-tego wiersza i j-tej kolumny macierzy $\mathbf{\Gamma}$ oraz dodajemy średnią wartość wszystkich elemntów macierzy $\mathbf{\Gamma}$,

Rozwiązanie

- zakładamy, że symetryczna macierz odmienności $\mathbf{d}_{ij}$ o wymiarze (n, n) spełnia nierówność trójkąta,
- tworzymy macierz $\mathbf{\Gamma}$ tego samego wymiaru o elementach $\gamma_{ij} = -\frac{1}{2}d_{ij}^2$,
- od każdego elementu γ_{ij} odejmujemy średnią wartość elementów i-tego wiersza i j-tej kolumny macierzy $\mathbf{\Gamma}$ oraz dodajemy średnią wartość wszystkich elemntów macierzy $\mathbf{\Gamma}$,
- otrzymaną macierz oznaczamy jako $\mathbf{\Phi}$

$$\mathbf{\Phi} = \left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\mathbf{\Gamma}\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right),$$

Rozwiązanie

- zakładamy, że symetryczna macierz odmienności $\mathbf{d}_{ij}$ o wymiarze (n, n) spełnia nierówność trójkąta,
- tworzymy macierz $\mathbf{\Gamma}$ tego samego wymiaru o elementach $\gamma_{ij} = -\frac{1}{2}d_{ij}^2$,
- od każdego elementu γ_{ij} odejmujemy średnią wartość elementów i-tego wiersza i j-tej kolumny macierzy $\mathbf{\Gamma}$ oraz dodajemy średnią wartość wszystkich elemntów macierzy $\mathbf{\Gamma}$,
- otrzymaną macierz oznaczamy jako $\mathbf{\Phi}$

$$\mathbf{\Phi} = \left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\mathbf{\Gamma}\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right),$$

gdzie $\mathbf{I}$ jest macierzą jednostkową, a $\mathbf{1}$ jest wektorem jedynek.

Rozwiązanie

- zakładamy, że symetryczna macierz odmienności $\mathbf{d}_{ij}$ o wymiarze (n, n) spełnia nierówność trójkąta,
- tworzymy macierz $\mathbf{\Gamma}$ tego samego wymiaru o elementach $\gamma_{ij} = -\frac{1}{2}d_{ij}^2$,
- od każdego elementu γ_{ij} odejmujemy średnią wartość elementów i-tego wiersza i j-tej kolumny macierzy $\mathbf{\Gamma}$ oraz dodajemy średnią wartość wszystkich elemntów macierzy $\mathbf{\Gamma}$,
- otrzymaną macierz oznaczamy jako $\mathbf{\Phi}$

$$\mathbf{\Phi} = \left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\mathbf{\Gamma}\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right),$$

  gdzie $\mathbf{I}$ jest macierzą jednostkową, a $\mathbf{1}$ jest wektorem jedynek.

Przy przyjetych założeniach, jeżeli ponadto macierz $\mathbf{\Phi}$ jest nieujemnie określona jest to rozwiązanie pierwszej części zadania. Macierz $\mathbf{\Phi}$ jest rzędu co najwyżej $n-1$, wymiar s przestrzeni euklidesowej, w której można wskazać konfigurację punktów odtwarzających macierz $\mathbf{d}_{ij}$ jest równy rzędowi macierzy $\mathbf{\Phi}$.

Rozwiązanie

Algorytm uzyskiwania tych punktów jest następujący:

Algorytm uzyskiwania tych punktów jest następujący:

- znajdź wartości własne $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n = 0$ oraz odpowiadające im wektory własne $\mathbf{v}_i$ macierzy $\mathbf{\Phi}$,

Algorytm uzyskiwania tych punktów jest następujący:

- znajdź wartości własne $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n = 0$ oraz odpowiadające im wektory własne $\mathbf{v}_i$ macierzy $\mathbf{\Phi}$,
- tak przeskaluj wektory własne, by spełniony był warunek $\mathbf{v}_i^T \mathbf{v}_i = \lambda_i$; współrzędne n punktów wzdłuż i-tej osi w przestrzeni euklidesowej R^s dane są przez kolejne elementy wektora $\mathbf{v}_i$, wymiar s przestrzeni jest równy liczbie niezerowych wartości własnych λ_i

Podany algorytm nosi nazwę **skalowania klasycznego** lub **analizy współrzędnych głównych**.

Algorytm uzyskiwania tych punktów jest następujący:

- znajdź wartości własne $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n = 0$ oraz odpowiadające im wektory własne $\mathbf{v}_i$ macierzy $\boldsymbol{\Phi}$,
- tak przeskaluj wektory własne, by spełniony był warunek $\mathbf{v}_i^T \mathbf{v}_i = \lambda_i$; współrzędne n punktów wzdłuż i-tej osi w przestrzeni euklidesowej R^s dane są przez kolejne elementy wektora $\mathbf{v}_i$, wymiar s przestrzeni jest równy liczbie niezerowych wartości własnych λ_i

Podany algorytm nosi nazwę **skalowania klasycznego** lub **analizy współrzędnych głównych**.

Dzięki tej samej konstrukcji można odpowiedzieć na drugie pytanie: najlepsza (w sensie wskaźnika V) u-wymiarowa reprezentacja punktów o macierzy odmienności $\mathbf{d}_{ij}$, $u < s$ dana jest przez u pierwszych wektorów własnych macierzy $\boldsymbol{\Phi}$, przy czym $V = 2n(\lambda_{u+1} + \ldots + \lambda_n)$.

Algorytm uzyskiwania tych punktów jest następujący:

- znajdź wartości własne $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n = 0$ oraz odpowiadające im wektory własne $\mathbf{v}_i$ macierzy $\mathbf{\Phi}$,
- tak przeskaluj wektory własne, by spełniony był warunek $\mathbf{v}_i^T \mathbf{v}_i = \lambda_i$; współrzędne n punktów wzdłuż i-tej osi w przestrzeni euklidesowej R^s dane są przez kolejne elementy wektora $\mathbf{v}_i$, wymiar s przestrzeni jest równy liczbie niezerowych wartości własnych λ_i

Podany algorytm nosi nazwę **skalowania klasycznego** lub **analizy współrzędnych głównych**.

Dzięki tej samej konstrukcji można odpowiedzieć na drugie pytanie: najlepsza (w sensie wskaźnika V) u-wymiarowa reprezentacja punktów o macierzy odmienności $\mathbf{d}_{ij}$, $u < s$ dana jest przez u pierwszych wektorów własnych macierzy $\mathbf{\Phi}$, przy czym $V = 2n(\lambda_{u+1} + \ldots + \lambda_n)$.

Podobieństwo pomiędzy skalowaniem wielowymiarowym a analizą składowych głównych staje się równoważnością, gdy dane macierz $\mathbf{d}_{ij}$ jest macierzą odległości euklidesowych.

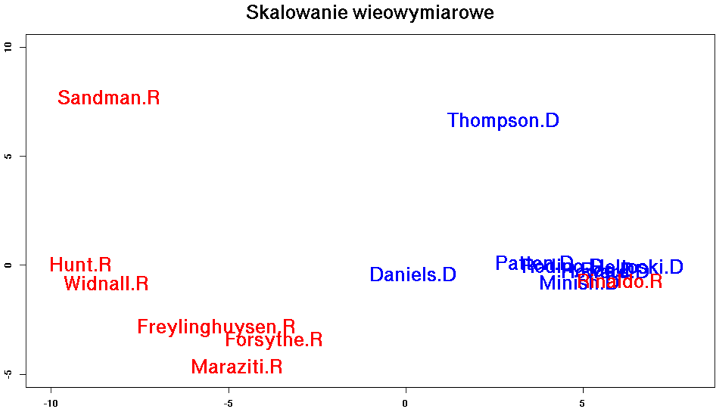
Przykład - głosowania w Kongresie USA
Dane zawierają macierz rozbieżności w głosowaniach dla 15 kongresmenów, dotycząca 19 głosowań (3 możliwe wyniki głosowania: za, przeciw, wstrzymanie się od głosu). Odległość jest wyznaczona jako liczba głosowań, w których głosowali różnie. Demokraci są zaznaczeni na niebiesko, republikanie na czerwono.

Przykład - głosowania w Kongresie USA

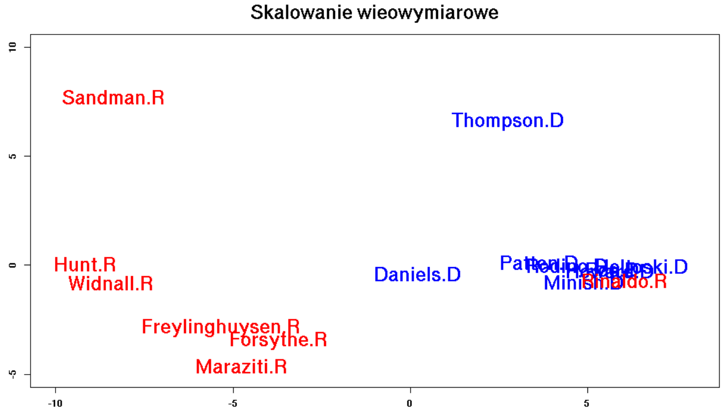
Dane zawierają macierz rozbieżności w głosowaniach dla 15 kongresmenów, dotycząca 19 głosowań (3 możliwe wyniki głosowania: za, przeciw, wstrzymanie się od głosu). Odległość jest wyznaczona jako liczba głosowań, w których głosowali różnie. Demokraci są zaznaczeni na niebiesko, republikanie na czerwono.

	H	S	H
Hunt	0	8	15
Sandman	8	0	17
Howard	15	17	0
Thompson	15	12	9
Freylinghuysen	10	13	16
Forsythe	9	13	12
Widnall	7	12	15
Roe	15	16	5
Heltoski	16	17	5
Rodino	14	15	6
Minish	15	16	5
Rinaldo	16	17	4
Maraziti	7	13	11
Daniels	11	12	10
Patten	13	16	7

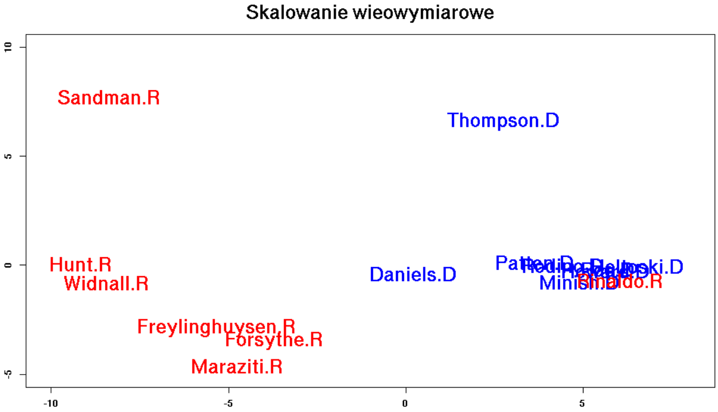
Przykład - głosowania w Kongresie USA



Przykład - głosowania w Kongresie USA

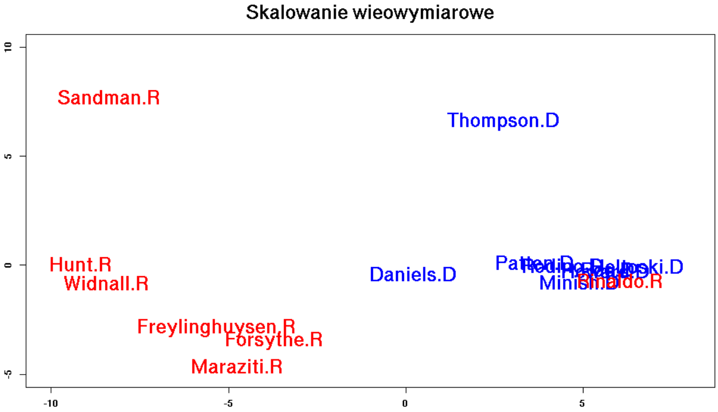


- Sandman i Thompson są daleko od centrów swoich partii,



- Sandman i Thompson są daleko od centrów swoich partii,
- republikanin Rinaldo jest bardzo blisko demokratów,

Przykład - głosowania w Kongresie USA



- Sandman i Thompson są daleko od centrów swoich partii,
- republikanin Rinaldo jest bardzo blisko demokratów,