Singularni razcep - 3. del

9. Kovariančna in korelacijska matrika

Najprej osvežimo osnovne pojme iz opisne statistike. Ponavljamo poskus in beležimo podatke na merilnih napravah. Dobimo **tabelo podatkov**.

Če imamo eno merilno napravo, ima tabela podatkov en stolpec podatkov.

ponovitev	podatek
1.	x_1
2.	x_2
$\vdots$	$\vdots$
n-ta	x_n

Za ta stolpec izračunamo povprečje

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n}.$$

Če nas zanima razpršenost podatkov okrog povprečja, potem najprej izračunamo odmike podatkov od povprečja, se pravi števila

$$x_i' = x_i - \bar{x}.$$

Dobimo **centralizirano tabelo podatkov**

ponovitev	centralizirani podatek
1.	$x_1' = x_1 - \bar{x}$
2.	$x_2' = x_2 - \bar{x}$
$\vdots$	$\vdots$
n-ta	$x_n' = x_n - \bar{x}$

Povprečen odmik od povprečja ni dobra mera za razpršenost, saj velja

$$\frac{x_1' + \ldots + x_n'}{n} = \frac{x_1 + \ldots + x_n - n\bar{x}}{n} = \bar{x} - \bar{x} = 0.$$

Boljša mera za razpršenost podatkov je kvadratična sredina odmikov

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i')^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Temu številu pravimo **standardna deviacija** stolpca podatkov.

Če vse elemente v centralizirani tabeli podatkov delimo z s dobimo **standardizirano tabelo podatkov**

ponovitev	standardizirani podatek
1.	$x_1'' = x_1'/s = (x_1 - \bar{x})/s$
2.	$x_2'' = x_2'/s = (x_2 - \bar{x})/s$
$\vdots$	$\vdots$
n-ta	$x_n'' = x_n'/s = (x_n - \bar{x})/s$

Povprečje standardizirane tabele podatkov je 0, njena standardna deviacija pa je 1, saj je $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i'')^2} = \sqrt{\frac{1}{ns^2}\sum_{i=1}^{n}(x_i')^2} = \sqrt{\frac{s^2}{s^2}} = 1.$

Če imamo dve merilni napravi, ima tabela podatkov dva stolpca podatkov.

ponovitev poskusa	podatek prve merilne naprave	podatek druge merilne naprave
1.	x_1	y_1
2.	x_2	y_2
$\vdots$	$\vdots$	$\vdots$
n-ta	x_n	y_n

Seveda lahko obdelamo vsako stolpec posebej kot pri eni merilni napravi.
Če stolpca centraliziramo, dobimo **centralizirano tabelo podatkov**.
Če ju standardiziramo, dobimo **standardizirano tabelo podatkov**.

Še bolj nas zanima koliko so podatki v enem stolpcu odvisni od podatkov v drugem stolpcu. Stopnjo odvisnosti merimo s **kovarianco**

$$K = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i' y_i'$$

Če so podatki prve merilne naprave neodvisni od podatkov druge merilne naprave, je kovarianca enaka nič. Obratno ne velja.

Opomba: Pojem kovariance je tesno povezan s pojmom regresijske premice. Spomnimo se, da je to premica, ki se najbolje prilega točkam $(x_1, y_1), \ldots, (x_n, y_n)$. Izpeljali smo, da je njen smerni koeficient enak

$$k = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}.$$

Pokažimo, da je števec tega izraza ravno kovarianca. Velja namreč

$$\begin{aligned} K &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n}\bar{x}y_i - \frac{1}{n}\sum_{i=1}^{n} x_i\bar{y} + \frac{1}{n}\sum_{i=1}^{n}\bar{x}\bar{y} \\ &= \overline{xy} - \bar{x}\bar{y} - \bar{x}\bar{y} + \bar{x}\bar{y} \\ &= \overline{xy} - \bar{x}\bar{y} \end{aligned}$$

Skoraj enak račun nam pokaže, da je imenovalec enak s_1^2, kjer je s_1 standardna deviacija prvega stolpca. Velja torej $k = \frac{K}{s_1^2}$.

S pomočjo kovariance in standardnih deviacij obeh stolpcev lahko definiramo **kovariančno matriko**

$$C = \begin{bmatrix} s_1^2 & K \\ K & s_2^2 \end{bmatrix}.$$

Če vstavimo definicije s_1, s_2 in K, dobimo

$$\begin{aligned} C &= \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n (x_i')^2 & \frac{1}{n}\sum_{i=1}^n x_i' y_i' \\ \frac{1}{n}\sum_{i=1}^n x_i' y_i' & \frac{1}{n}\sum_{i=1}^n (y_i')^2 \end{bmatrix} \\ &= \frac{1}{n}\begin{bmatrix} x_1' & \dots & x_n' \\ y_1' & \dots & y_n' \end{bmatrix} \begin{bmatrix} x_1' & y_1' \\ \vdots & \vdots \\ x_n' & y_n' \end{bmatrix} \\ &= \frac{1}{n}(X')^T X' \end{aligned}$$

kjer je X' matrika centraliziranih podatkov.

Opomba: Matriko X' dobimo tako, da v matriki rezultatov X, od vsakega elementa odštejemo povprečje stolpca v katerem se nahaja.

Opazimo, da se kovarianca vedno nahaja med $-s_1 s_2$ in $s_1 s_2$. To sledi iz Cauchy-Schwartzove neenakosti

$$|\sum_{i=1}^{n} x_i' y_i'| \leq \sqrt{\sum_{i=1}^{n} (x_i')^2} \sqrt{\sum_{i=1}^{n} (y_i')^2}.$$

če obe strani delimo z $n = (\sqrt{n})^2$. Zato pridemo na idejo, da bi definirali

$$r = \frac{K}{s_1 s_2}$$

Temu izrazu pravimo **korelacija** (ali korelacijski koeficient) obeh stolpcev v matriki podatkov. S standardiziranimi podatki se izraža takole

$$r = \frac{1}{n s_1 s_2} \sum_{i=1}^{n} x_i' y_i' = \frac{1}{n} \sum_{i=1}^{n} x_i'' y_i''$$

Torej je korelacija enaka nič natanko tedaj, ko sta stolpca v standardizirani tabeli podatkov ortogonalna.

Definirajmo še **korelacijsko matriko**

$$R = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$$

Opazimo, da velja

$$\begin{aligned} R &= \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n}(x_i'')^2 & \frac{1}{n}\sum_{i=1}^{n} x_i'' y_i'' \\ \frac{1}{n}\sum_{i=1}^{n} x_i'' y_i'' & \frac{1}{n}\sum_{i=1}^{n}(y_i'')^2 \end{bmatrix} \\ &= \frac{1}{n}\begin{bmatrix} x_1'' & \dots & x_n'' \\ y_1'' & \dots & y_n'' \end{bmatrix} \begin{bmatrix} x_1'' & y_1'' \\ \vdots & \vdots \\ x_n'' & y_n'' \end{bmatrix} \\ &= \frac{1}{n}(X'')^T X'' \end{aligned}$$

kjer je X'' matrika standardiziranih podatkov.

Opomba: Matriko X'' dobimo tako, da v matriki centraliziranih podatkov X' vsak stolpec delimo z njegovo standardno deviacijo.

Posplošimo sedaj definicije iz dveh na p merilnih naprav. Naj bo $x_{i,j}$ podatek j-te merilne naprave pri i-ti ponovitvi poskusa.

	1. merilnik	2. merilnik	...	p-ti merilnik
1. ponovitev	$x_{1,1}$	$x_{1,2}$	...	$x_{1,p}$
2. ponovitev	$x_{2,1}$	$x_{2,2}$	...	$x_{2,p}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
n-ta ponovitev	$x_{n,1}$	$x_{n,2}$	...	$x_{n,p}$
povprečje	$\bar{x}_1$	$\bar{x}_2$	...	$\bar{x}_p$
stand. deviacija	s_1	s_2	...	s_p

Za vsak stolpec v tabeli smo izračunali povprečje in standardno deviacijo

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{i,j}, \qquad s_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{i,j} - \bar{x}_j)^2}$$

Matriki $X = [x_{i,j}]$ pravimo **matrika podatkov**. Njena velikost je $n \times p$, kjer je n število ponovitev poskusa, p pa število merilnih naprav.

Matriki $X' = [x'_{i,j}]$, kjer je

$$x'_{i,j} = x_{i,j} - \bar{x}_j$$

odmik od povprečja, bomo rekli **centralizirana matrika podatkov**.

Matriki $X'' = [x''_{i,j}]$, kjer je

$$x''_{i,j} = x'_{i,j}/s_j = (x_{i,j} - \bar{x}_j)/s_j$$

standardiziran odmik, bomo rekli **standardizirana matrika podatkov**.

Kovarianca j-tega in k-tega stolpca matrike X je enaka

$$C_{j,k} = \frac{1}{n}\sum_{i=1}^{n}(x_{i,j} - \bar{x}_j)(x_{i,k} - \bar{x}_k) = \frac{1}{n}\sum_{i=1}^{n} x'_{i,j}x'_{i,k}$$

Opazimo, da je $C_{j,j} = s_j^2$, kjer je s_j standardna deviacija j-tega stolpca matrike X. **Kovariančna matrika** $C = [C_{j,k}]$ zadošča

$$C = \frac{1}{n}(X')^T X'$$

Korelacija med j-tim in k-tim stolpcem matrike X je enaka

$$r_{j,k} = C_{j,k}/(s_j s_k)$$

Velja $-1 \leq r_{j,k} \leq 1$ in $r_{j,j} = 1$. **Korelacijska matrika** $R = [r_{j,k}]$ zadošča

$$R = \frac{1}{n}(X'')^T X''$$

10. Metoda glavnih komponent (PCA)

Naj bo X' centralizirana matrika podatkov za nek poskus.
(Pozor, nekateri raje delajo s standardizirano matriko podatkov X''.)

Radi bi zmanjšali število stolpcev matrike X' (iz p na k), ne da bi pri tem izgubili veliko informacij. Ideja je, da si zapomnimo samo take linearne kombinacije stolpcev matrike X', ki imajo največjo standarno deviacijo.

Spomnimo se, da vsako linearno kombinacijo stolpcev X' lahko zapišemo kot $X'w$ za nek vektor w. Ponavljamo naslednji postopek:

- Poišči tak normiran vektor w_1, da ima vektor $y_1 := X'w_1$ največjo možno standardno deviacijo.
- Med vsemi normiranimi vektorji, ki so ortogonalni na w_1, poišči tak vektor w_2, da ima $y_2 := X'w_2$ največjo možno standardno deviacijo.
- Med vsemi normiranimi vektorji, ki so ortogonalni na w_1 in w_2, poišči tak vektor w_3, da ima $y_3 := X'w_3$ največjo možno stand. deviacijo.
- ...

Končamo, ko se X' ujema z $y_1 w_1^T + \ldots + y_k w_k^T$ do predpisane natančnosti.

Opomba: Vektorjem $y_1, \ldots, y_k$ bomo rekli **glavne komponente** matrike X', vektorjem $w_1, \ldots, w_k$ pa **glavne osi** matrike X'.
Pozor, terminologija se zelo razlikuje od avtorja do avtorja.

Trditev 1

Za vektorje $w_1, w_2, w_3, \ldots$ lahko vzamemo kar lastne vektorje kovariančne matrike C, ki ustrezajo prvi, drugi, tretji, ... največji lastni vrednosti.

Opomba: Lastni vektorji kovariančne matrike

$$C = \frac{1}{n}(X')^T X'$$

se ujemajo z lastnimi vektorji matrike

$$(X')^T X'$$

ti pa se ujemajo s stolpci matrike Q_2 v singularnem razcepu

$$X' = Q_1 D Q_2^T$$

Dokaz. Naj bodo $\lambda_1 \geq \ldots \geq \lambda_p$ lastne vrednosti matrike C in naj bodo $v_1, \ldots, v_p$ pripadajoči lastni vektorji. Poskrbimo, da so ortonormirani.

Vsak normiran vektor $w \in \mathbb{R}^p$ lahko zapišemo kot $w = \beta_1 v_1 + \ldots + \beta_p v_p$ kjer je $\beta_1^2 + \ldots + \beta_p^2 = 1$. Standardna deviacija vektorja $X'w$ je enaka

$$\begin{aligned}
\sigma_{X'w} &= \sqrt{\frac{1}{n}(X'w)^T(X'w)} \\
&= \sqrt{w^T C w} \\
&= \sqrt{(\beta_1 v_1 + \ldots + \beta_p v_p)^T(\beta_1 \lambda_1 v_1 + \ldots + \beta_p \lambda_p v_p)} \\
&= \sqrt{\beta_1^2 \lambda_1 + \ldots + \beta_p^2 \lambda_p} \\
&= \sqrt{(1 - \beta_2^2 - \ldots - \beta_p^2)\lambda_1 + \beta_2^2 \lambda_2 + \ldots + \beta_p^2 \lambda_p} \\
&= \sqrt{\lambda_1 - (\beta_2^2(\lambda_1 - \lambda_2) + \ldots + \beta_p^2(\lambda_1 - \lambda_p))} \\
&\leq \sqrt{\lambda_1}
\end{aligned}$$

Enačaj je očitno dosežen pri $w = v_1$. Torej lahko vzamemo $w_1 = v_1$.

Vsak normiran vektor $w \in \mathbb{R}^p$, ki je ortogonalen na v_1 lahko zapišemo kot

$$w = \gamma_2 v_2 + \ldots + \gamma_p v_p,$$

kjer je $\gamma_2^2 + \ldots + \gamma_p^2 = 1$. Standardna deviacija vektorja $X'w$ je enaka

$$\begin{aligned}
\sigma_{X'w} &= \sqrt{w^T C w} \\
&= \sqrt{\gamma_2^2 \lambda_2 + \ldots + \gamma_p^2 \lambda_p} \\
&= \sqrt{(1 - \gamma_3^2 - \ldots - \gamma_p^2)\lambda_2 + \gamma_3^2 \lambda_3 + \ldots + \gamma_p^2 \lambda_p} \\
&= \sqrt{\lambda_2 - (\gamma_3^2(\lambda_2 - \lambda_3) + \ldots + \gamma_p^2(\lambda_2 - \lambda_p))} \\
&\leq \sqrt{\lambda_2}
\end{aligned}$$

Enačaj je dosežen pri $w = v_2$, zato lahko vzamemo $w_2 = v_2$.

Normiran vektor $w \in \mathbb{R}^p$, ki je ortogonalen na v_1 in v_2 lahko zapišemo kot $w = \delta_3 v_3 + \ldots + \delta_p v_p$, kjer je $\delta_3^2 + \ldots + \delta_p^2 = 1$. Podobno kot zgoraj je $\sigma_{X'w} \leq \sqrt{\lambda_3}$ in enačaj je dosežen pri $w = v_3$. To lahko nadaljujemo. □

Trditev 3

Če je $i \neq j$, potem je kovarianca med stolpcema y_i in y_j enaka nič.

Dokaz: Ker sta stolpca y_i in y_j linearni kombinaciji stolpcev matrike X', sta njuni povprečji enaki nič. Torej je njuna kovarianca enaka

$$\begin{aligned} K &= \frac{1}{n}\langle y_i, y_j \rangle \\ &= \frac{1}{n} y_j^T y_i \\ &= \frac{1}{n}(X' w_j)^T (X' w_i) \\ &= w_j^T C w_i \\ &= w_j^T (\lambda_i w_i) \\ &= \lambda_i \langle w_i, w_j \rangle \\ &= 0 \end{aligned}$$

Opomba: Različne glavne komponente X' so torej nekorelirane.

V nadaljevanje nas bo zanimalo, kolikšno napako naredimo, če matriko X' zamenjamo z matriko $y_1 w_1^T + \ldots + y_k w_k^T$. Napako bomo merili s Frobeniusovo matrično normo.

Uvedimo nekaj oznak. Za vsak $k = 1, \ldots, p$ naj bo

$$W_k = \begin{bmatrix} w_1 & \ldots & w_k \end{bmatrix} \quad \text{in} \quad Y_k = \begin{bmatrix} y_1 & \ldots & y_k \end{bmatrix} = X' W_k$$

Očitno je

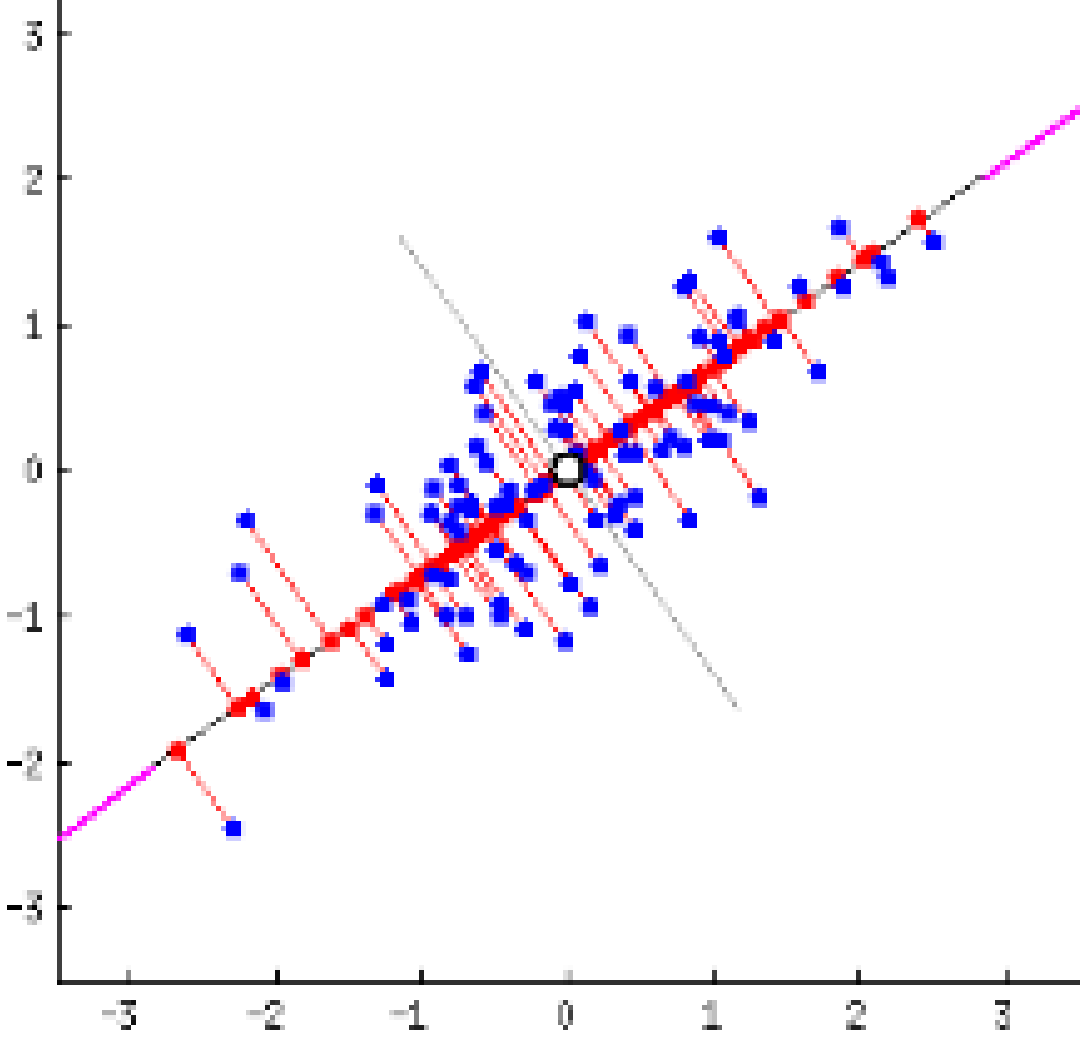
$$Y_k W_k^T = y_1 w_1^T + \ldots + y_k w_k^T$$

Iskana napaka je torej

$$\|X' - Y_k W_k^T\|$$

Trditev 4 (Geometrijski pomen $Y_k W_k^T$ in $\|X' - Y_k W_k^T\|$)

Če i-to vrstico matrike X' ortogonalno projeciramo na podprostor $\mathrm{Lin}\{w_1^T, \ldots, w_k^T\}$ v $\mathbb{R}^p$, potem dobimo ravno i-to vrstico matrike $Y_k W_k^T$. Vsota kvadratov oddaljenosti vrstic matrike X' od njihovih ortogonalnih projekcij na podprostor $\mathrm{Lin}\{w_1^T, \ldots, w_k^T\}$ je torej enaka $\|X' - Y_k W_k^T\|^2$.



Vsaka koordinatna os ustreza eni merilni napravi. Na skici je $p = 2$.
Vsaka modra točka ustreza eni vrstici matrike X', torej eni ponovitvi poskusa.
Modra premica ustreza podprostoru $\mathrm{Lin}\{w_1^T, \ldots, w_k^T\}$. Na skici je $k = 1$.
Vsaka rdeča točka ustreza eni vrstici matrike $Y_k W_k^T$.

Dokaz. Vrstice matrik X' in $Y_k W_k^T$ ter vektorji $w_1^T, \ldots, w_k^T$ so vrstični vektorji dolžine p. Ker so $w_1, \ldots, w_k$ ortonormirani, so tudi $w_1^T, \ldots, w_k^T$ ortonormirani. Naj bo $x^{(i)}$ i-ta vrstica matrike X'. Njena ortogonalna projekcija na podprostor $\mathrm{Lin}\{w_1^T, \ldots, w_k^T\}$ je enaka $\sum_{j=1}^{k} \langle x^{(i)}, w_j^T \rangle w_j^T$.

Najprej opazimo, da je skalarni produkt $\langle x^{(i)}, w_j^T \rangle$ enak matričnemu produktu $x^{(i)} w_j$. Poleg tega je

$$\sum_{j=1}^{k} \langle x^{(i)}, w_j^T \rangle w_j^T = \begin{bmatrix} x^{(i)} w_1 & \ldots & x^{(i)} w_k \end{bmatrix} \begin{bmatrix} w_1^T \\ \vdots \\ w_k^T \end{bmatrix} = x^{(i)} W_k W_k^T.$$

Tudi i-ta vrstica produkta $X'(W_k W_k^T)$ je enaka $x^{(i)} W_k W_k^T$.

Drugi del trditve sledi iz prvega dela in naslednjega računa. Za poljubni matriki A in B iste velikosti $n \times p$ velja

$$\|A - B\|^2 = \sum_{i=1}^{n} \sum_{j=1}^{p} (a_{i,j} - b_{i,j})^2 = \sum_{i=1}^{n} \|a^{(i)} - b^{(i)}\|^2. \quad \square$$

Trditev 5

Če so $\lambda_1 \geq \ldots \geq \lambda_p$ lastne vrednosti kovariančne matrike $C = \frac{1}{n}(X')^T X'$, potem je $\|X' - Y_k W_k^T\|^2 = n(\lambda_{k+1} + \ldots + \lambda_p)$.

Dokaz. Izračunajmo razdaljo med X' in $Y_k W_k^T$ v Frobeniusovi normi. Naj bo $X' = UDW^T$ singularni razcep matrike X' in naj bo $W = \begin{bmatrix} W_k & Z \end{bmatrix}$ za nek Z. Iz $I_p = W^T W = W^T \begin{bmatrix} W_k & Z \end{bmatrix} = \begin{bmatrix} W^T W_k & W^T Z \end{bmatrix}$ sledi $W^T W_k = \begin{bmatrix} I_k \\ 0 \end{bmatrix}$. Naj bo D_k matrika, ki jo dobimo iz D tako, da vse diagonalne elemente razen prvih k zamenjamo z nič. Opazimo, da velja

$$Y_k W_k^T = X' W_k W_k^T = UDW^T W_k W_k^T = UD \begin{bmatrix} I_k \\ 0 \end{bmatrix} W_k^T =$$

$$= UD_k \begin{bmatrix} I_k \\ 0 \end{bmatrix} W_k^T = UD_k \begin{bmatrix} W_k^T \\ 0 \end{bmatrix} = UD_k \begin{bmatrix} W_k^T \\ Z^T \end{bmatrix} = UD_k W^T$$

Odtod sledi, da je v Frobeniusovi normi

$$\|X' - Y_k W_k^T\| = \|U(D - D_k)W^T\| = \|D - D_k\|.$$

Če so $\lambda_1 \geq \ldots \geq \lambda_p$ lastne vrednosti matrike $C = \frac{1}{n}(X')^T X'$, potem so $n\lambda_1 \geq \ldots \geq n\lambda_p$ lastne vrednosti matrike $(X')^T X'$. Odtod sledi, da so $\sqrt{n\lambda_1} \geq \ldots \geq \sqrt{n\lambda_p}$ diagonalni elementi matrike D. Torej je

$$\|D - D_k\|^2 = (\sqrt{n\lambda_{k+1}})^2 + \ldots + (\sqrt{n\lambda_p})^2 = n(\lambda_{k+1} + \ldots + \lambda_p). \quad \square$$

Naslednja posledica Trditve 5 nam pove kako izbrati k, da dosežemo predpisano natančnost.

Posledica

Če za nek k velja $\lambda_{k+1} + \ldots + \lambda_p \leq \frac{\varepsilon^2}{n}$, potem je $\|X' - Y_k W_k^T\| \leq \varepsilon$.
Če pa za nek k velja $\frac{\lambda_{k+1}+\ldots+\lambda_p}{\lambda_1+\ldots+\lambda_k+\lambda_{k+1}+\ldots+\lambda_p} \leq \varepsilon$, potem je $\frac{\|X'-Y_k W_k^T\|^2}{\|X'\|^2} \leq \varepsilon$.

Za konec si oglejmo povzetek algoritma za računanje glavnih komponent.

1. Preberi matriko podatkov X in predpisano relativno natančnost ε.
2. Zapomni si povprečja stolpcev matrike X; $\bar{x} := \begin{bmatrix} \bar{x}_1 & \dots & \bar{x}_p \end{bmatrix}$.
3. Centraliziraj matriko podatkov; $X' := X - e^T \bar{x}$, kjer $e = [1, \dots, 1]$.
4. Poišči singularni razcep $X' = UDW^T$, kjer za diagonalne elemente matrike D (=singularne vrednosti matrike X') velja $\sigma_1 \geq \dots \geq \sigma_p$.
5. Poišči tak k, da velja $\frac{\sigma_1^2 + \dots + \sigma_k^2}{\sigma_1^2 + \dots + \sigma_k^2 + \dots + \sigma_p^2} \geq 1 - \varepsilon^2$.
6. Zapomni si prvih k stolpcev matrike W ($= W_k =$ glavne osi X') in prvih k stolpcev matrike UD ($= Y_k =$ glavne komponente X')
7. Velja $\frac{\|X' - Y_k W_k^T\|}{\|X'\|} \leq \varepsilon$. Torej je $Y_k W_k^T$ dober približek za X'.
   Odtod sledi, da je $Y_k W_k^T + e^T \bar{x}$ dober približek za matriko X.