Adatbányászat

Dimenziócsökkentő eljárások

Szegedi Tudományegyetem

Dimenziócsökkentés szerepe

- Az adatpontok reprezentálására használt dimenziók számának csökkentésével spórolhatunk
- Többdimenziós adatpontok vizualizálása (pl. 2 vagy 3D-ben)
- Zajcsökkentés, jellemzőszelekció
- Ötlet: próbáljuk meg alacsonyabb dimenzióban ábrázolni a pontjainkat
- Az alacsonyabb dimenziós reprezentáció kapcsán miben mérjük az elégedettségünket? $\rightarrow$ PCA, LDA, SVD, ...

Főkomponens analízis (Principal Component Analysis)

- Transzformáljuk úgy alacsonyabb dimenzióba a több/sokdimenziós adatunkat, hogy a benne rejlő variancia minél kisebb részét veszítsük csak el
- Feltevés: az eredeti m-dimenziós adatpontok egy bizonyos m'-dimenziós altéren (vagy legalábbis annak közelében) helyezkednek el $\rightarrow$ az adatpontokat az altér tengelyeire transzformálva jól reprezentálható az eredeti adat
- Mi lehet ez a m'-dimenziós altér?
- $\sum_{i=1}^{n} \|(x_i - x_i')\|^2$ rekonstrukciós hibát szeretnénk minimálisnak látni, ahol x_i' az eredeti x_i pont egy közelítése

Kovariancia

Emlékeztető

- Y és Z véletlen változók együttmozgását számszerűsíti
- $cov(Y, Z) = \mathbb{E}[(Y - \mu_Y)(Z - \mu_Z)]$
  - Ahol $\mu_Y = \frac{1}{n}\sum_{i=1}^{n} y_i$ és $\mu_Z = \frac{1}{n}\sum_{i=1}^{n} z_i$

- Az X adatmátrixunk i, j oszlopaira $(X_{:,i}, X_{:,j})$ tekinthetünk véletlen változóként

Szóródási,- és kovarianciamátrixok

- Szóródási mátrix: $S = \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\mathsf{T}$
- Torzított (torzítatlan) kovarianciamátrix: $\Sigma = \frac{1}{n} S$ ($\Sigma = \frac{1}{n-1} S$)

$$\Sigma = \begin{bmatrix} cov(X_{:,1}, X_{:,1}) & cov(X_{:,1}, X_{:,2}) & \dots & cov(X_{:,1}, X_{:,m}) \\ cov(X_{:,2}, X_{:,1}) & cov(X_{:,2}, X_{:,2}) & \dots & cov(X_{:,2}, X_{:,m}) \\ \vdots & \ddots & cov(X_{:,i}, X_{:,j}) & \vdots \\ cov(X_{:,m}, X_{:,1}) & \dots & \dots & cov(X_{:,m}, X_{:,m}) \end{bmatrix}$$

- $\Sigma_{i,j}$ tehát az i. és j. jellemzők kovarianciája ($cov(X_{:,i}, X_{:,j})$)
- Milyen elemek állnak a főátlóban?

Szóródási,- és kovarianciamátrixok tulajdonságai

- Állítás: S és Σ mátrixok szimmetrikusak és pozitív definitek

Bizonyítás.

$$S = \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\mathsf{T} = \left(\sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\mathsf{T} \right)^\mathsf{T} = S^\mathsf{T}$$

$$\boldsymbol{a}^\mathsf{T} S \boldsymbol{a} = \sum_{i=1}^{n} (\boldsymbol{a}^\mathsf{T}(\boldsymbol{x}_i - \boldsymbol{\mu}))((\boldsymbol{x}_i - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{a}) = \sum_{i=1}^{n} (\boldsymbol{a}^\mathsf{T}(\boldsymbol{x}_i - \boldsymbol{\mu}))^2 \geq 0 \qquad \square$$

- Következmény: S és Σ sajátértékeire $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m \geq 0$
- Az adat varianciáját legjobban megőrző m'-dimenziós projekciót úgy kapjuk, ha az adatpontokat S (vagy Σ) m' legnagyobb sajátértékéhez tartozó sajátvektorából képzett mátrixszal szorozzuk (lásd: tábla)

Lagrange szorzók

- Korlátozott/feltételes (nemlineáris) optimalizációs feladat

$$f(\boldsymbol{x}) \to min/max$$
$$\text{f.h. } g_i(\boldsymbol{x}) = 0 \forall i \in \{1, \ldots, n\}$$

- Lagrange függvény:

$$L(\boldsymbol{x}, \lambda) = \mathrm{f(x)} - \sum_{\mathrm{i=1}}^{\mathrm{n}} \lambda_{\mathrm{i}} \mathrm{g_i(x)}$$

- Karush-Kuhn-Tucker (KKT) feltételek: az optimalitás **szükséges** feltételei

$$\nabla L(\boldsymbol{x}, \lambda) = 0 \tag{1}$$
$$\lambda_i g_i(\boldsymbol{x}) = 0 \forall i \in \{1, \ldots, n\} \tag{2}$$
$$\lambda_i \geq 0 \tag{3}$$

Gyakorlati megfontolások

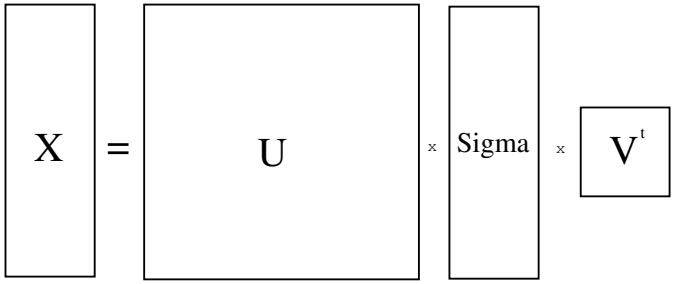
- Az adatpontokat érdemes egységes skálára hozni: normalizáció
  - min-max normalizáció: $x_{i,j} = \frac{x_{i,j} - min(x_{*,j})}{max(x_{*,j}) - min(x_{*,j})}$
  - standardizálás: $x_{i,j} = \frac{x_{i,j} - \mu_j}{\sigma_j}$
- A redukált dimenziószámot (m') minek válasszuk?

  $Hint: \left(\sum_{i=1}^{m} \lambda_i = \sum_{i=1}^{m} s_i^2 \right)$
  - $$m' = \arg \min_{1 \leq k \leq m} \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{m} \lambda_i} \geq t \text{ küszöbérték}$$
  - $$m' = \arg \max_{1 \leq i \leq m} \left(\lambda_i > \frac{1}{m} \sum_{j=1}^{m} \lambda_j \right)$$
  - $$m' = \arg \max_{1 \leq i \leq m-1} (\lambda_i - \lambda_{i+1})$$

PCA összegzés

- Centralizáljuk és normalizáljuk az X adatmátrixot
- Számoljuk ki a (centralizált és normalizált) X adatmátrixot jellemző kovariancia/szóródási-mátrixot
- Számoljuk ki a mátrix sajátértékeit
- Az m' legnagyobb sajátértékhez tartozó sajátvektorból képezzük P projekciós mátrixot
- $X' = XP$ adja meg a transzformált adatmátrixot
- $X'P^{\mathsf{T}}$ alakban kaphatjuk vissza az eredeti adatpontok egy közelítését
- PCA tutorial

Szinguláris érték felbontás



- $X = U\Sigma V^{\mathsf{T}} = \sum_{i=1}^{rang(X)} \sigma_i u_i v_i^{\mathsf{T}}$
- $\|X\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}^2} = \sqrt{\sum_{i=1}^{rang(X)} \sigma_i^2}$

- X alacsony(abb) rangú közelítése: $\tilde{X} = U\tilde{\Sigma}V^{\mathsf{T}}$
- $\tilde{X}$ előállítása során csupán Σ első $m' < m$ szinguláris értékére hagyatkozzunk
- Ha a közelítés eltérését Frobenius-normában mérjük, akkor X megadható legpontosabb m'-dimenziós becslése $\tilde{X}$

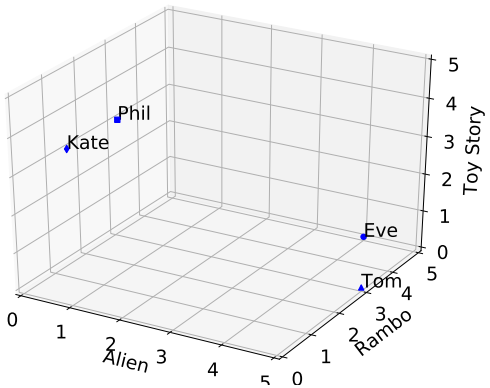
SVD példa

$$\begin{bmatrix} 5 & 3 & 0 \\ 4 & 5 & 0 \\ 1 & 0 & 4 \\ 2 & 0 & 5 \end{bmatrix} =$$

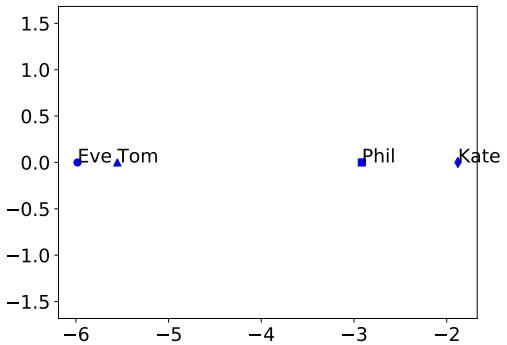
$$\begin{bmatrix} -0.63 & 0.22 & 0.73 \\ -0.67 & 0.33 & -0.65 \\ -0.21 & -0.58 & -0.19 \\ -0.33 & -0.72 & 0.08 \end{bmatrix} \begin{bmatrix} 8.87 & 0 & 0 \\ 0 & 6.33 & 0 \\ 0 & 0 & 1.52 \end{bmatrix} \begin{bmatrix} -0.75 & -0.59 & -0.28 \\ 0.06 & 0.36 & -0.93 \\ 0.65 & -0.72 & -0.24 \end{bmatrix}$$

SVD fölhasználása

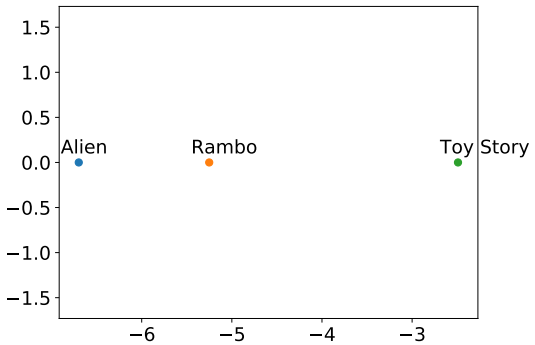
- A szinguláris vektorokkal egy *látens reprezentációt* nyerhetünk
- $X = U\Sigma V^\intercal$ alapján $XV = U\Sigma$, valamint $U^\intercal X = \Sigma V^\intercal$
- Tetszőleges $\mathrm{x} \notin X$ pontot elhelyezhetünk a látens térben $\mathrm{x}^\intercal V$ módon



(a) Erederti felhasználó–termék értékelések 3D-ben



(b) Felhasználók 1 rangú látens reprezentációja



(c) Termékek 1 rangú látens reprezentációja

SVD és a sajátérték dekompozíció viszonya

Emlékeztető

Egy szimmetrikus A mátrix felbontható $A = X\Lambda X^{-1}$ alakban, ahol $X = [\boldsymbol{x}_1 \boldsymbol{x}_2 \dots \boldsymbol{x}_m]$ az A (ortogonális) sajátvekroraiból képzett mátrix, $\Lambda = diag([\lambda_1 \lambda_2 \dots \lambda_m])$ pedig az egyes sajátvektorokhoz tartozó sajátértékeből álló diagonális mátrix. Miért is?

- Tetszőleges $n \times m$-es X mátrix előáll $U\Sigma V^\intercal$ szorzatként, ahol
  - $U_{n\times n} = [\boldsymbol{u}_1 \boldsymbol{u}_2 \dots \boldsymbol{u}_n]$ $XX^\intercal$ ortonormált sajátvektoraiból álló mátrix
  - $\Sigma_{n\times m} = diag(\sqrt{\lambda}_1, \sqrt{\lambda}_2, \dots, \sqrt{\lambda}_m)$
  - $V_{m\times m} = [\boldsymbol{v}_1 \boldsymbol{v}_2 \dots \boldsymbol{v}_m]$ $X^\intercal X$ ortonormált sajátvektoraiból álló mátrix. Miért?
- Ortogonális mátrix: $M^\intercal M = I$ (szemléletesen: olyan forgatás, ami a vektorok hosszát nem változtatja meg) Miért?

SVD és a Frobenius-norma viszonya

- Tfh. $M = P \times Q \times R$, azaz $m_{ij} = \sum_k \sum_l p_{ik} q_{kl} r_{lj}$

$$M = \begin{bmatrix} -58 & 87 \\ -32 & 48 \\ -28 & 42 \end{bmatrix} = \begin{bmatrix} 6 & 3 & 1 \\ 2 & 4 & 2 \\ 3 & 2 & 0 \end{bmatrix} \begin{bmatrix} 4 \\ 1 \\ 2 \end{bmatrix} \begin{bmatrix} -2 & 3 \end{bmatrix} \Rightarrow m_{32} = 3*4*3+2*1*3+0$$

- Ekkor $\|M\|_F^2 = \sum_i \sum_j (m_{ij})^2 = \sum_i \sum_j \left(\sum_k \sum_l p_{ik} q_{kl} r_{lj} \right)^2$
- Továbbá $\left(\sum_k \sum_l p_{ik} q_{kl} r_{lj} \right)^2 = \sum_k \sum_l \sum_m \sum_n p_{ik} q_{kl} r_{lj} p_{in} q_{nm} r_{mj}$
- Ahonnan $\|M\|_F^2 = \sum_i \sum_j \sum_k \sum_l \sum_m \sum_n p_{ik} q_{kl} r_{lj} p_{in} q_{nm} r_{mj}$
- Amennyiben P, Q, R mátrixok az SVD felbontásból jönnek, $\|M\|_F^2 = \sum_{i,j,k,n} p_{ik} q_{kk} r_{kj} p_{in} q_{nn} r_{nj} = \sum_{j,k} q_{kk} r_{kj} q_{kk} r_{kj} = \sum_k (q_{kk})^2$. Miért?
- X mátrix $\tilde{X} = U\tilde{\Sigma}V^\intercal$ mátrixszal történő közelítésének hibája: $\|X - \tilde{X}\|_F^2 = \|U(\Sigma - \tilde{\Sigma})V^\intercal\|_F^2 = \sum_k (\sigma_{kk} - \tilde{\sigma}_{kk})^2$

CUR

- SVD hátránya, hogy egy jellemzően ritka mátrixot sűrű mátrixokat (U és V) is tartalmazó szorzatként állítja elő
- Egy alternatíva az eredeti mátrix CUR formában történő előállítása
  - Csak az U mátrix adódik sűrűnek, és az SVD felbontás egy közelítését adja
  - C és R mátrixok az eredeti X mátrix soraiból, illetve oszlopaiból származnak, így azok megőrzik X ritkaságát
  - Míg az SVD egyedi felbontást eredményez, addig a CUR nem

CUR dekompozíció – C és R előállítása

- Visszatevésessel válasszunk ki k oszlopot az eredeti mátrixból
  - Előfordulhatnak ismétlődő oszlopok C-ben
  - Egy oszlop kiválasztásának p_i valószínűsége legyen a benne található elemek négyzetösszegével arányos
  - A kiválasztott oszlopot szorozzuk be $1/\sqrt{kp_i}$-vel (kp_i az i oszlop kiválasztásainak várható mennyisége)
- R mátrix előállítása csak abban tér el, hogy ott sorokkal dolgozunk

CUR dekompozíció – U előállítása

- $U = C^\dagger X R^\dagger$, ahol † a pszeudoinverzet jelöli, hiszen ekkor $CUR = C(C^\dagger X R^\dagger)R = (CC^\dagger)X(R^\dagger R) \approx X$
  - A pszeudoinverz a „hagyományos" mátrixinverz kiterjesztése nem feltétlen négyzetes és/vagy invertálható mátrixokra
    - $MM^\dagger M = M$
    - Négyzetes (és invertálható) M esetén $M^{-1} = M$
  - SVD vs. pszeudoinverz: $M = U\Sigma V^\intercal$ megadható, így $M^\dagger = (U\Sigma V^\intercal)^\dagger = V\Sigma^\dagger U^\intercal$
    - Diagonális mátrixot (pszeudo)invertálni könnyű

$$\begin{bmatrix} 5 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -0.2 \\ 0 & 0 & 0 \end{bmatrix}^\dagger = \begin{bmatrix} 0.2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -5 & 0 \end{bmatrix}$$

    - Elég transzponálni, és a nemnulla elemek reciprokát venni

CUR dekompozíció – példa

	Alien	Rambo	Toy Story	P
Tom	5	3	0	$\frac{34}{121}$
Eve	4	5	0	$\frac{41}{121}$
Kate	1	0	4	$\frac{17}{121}$
Phil	2	0	5	$\frac{29}{121}$
P	$\frac{46}{121}$	$\frac{34}{121}$	$\frac{41}{121}$	

$$C = \begin{bmatrix} 5.734 & 0 \\ 4.587 & 0 \\ 1.147 & 4.859 \\ 2.294 & 6.074 \end{bmatrix}$$

$$U = \begin{bmatrix} 0.134 & -0.001 \\ -0.047 & 0.113 \end{bmatrix}$$

$$R = \begin{bmatrix} 6.670 & 5.363 & 0 \\ 2.889 & 0 & 7.222 \end{bmatrix}$$

Lineáris Diszkriminancia Analízis (Linear Discriminant Analysis)

- Transzformáljuk úgy alacsonyabb dimenzióba a több/sokdimenziós címkézett adatunkat, hogy az új térben az eltérő osztályú pontok minél kevésbé keveredjenek
- Mi legyen w vetítés iránya?
- $\tilde{\boldsymbol{\mu}}_c = \boldsymbol{w}^\mathsf{T}\boldsymbol{\mu}_c \Rightarrow |\tilde{\boldsymbol{\mu}}_1 - \tilde{\boldsymbol{\mu}}_2| = |\boldsymbol{w}^\mathsf{T}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)|$
- $$\tilde{s}_c^2 = \sum_{\{(\boldsymbol{x}_i, y_i) | y_i = c\}} (\boldsymbol{w}^\mathsf{T}\boldsymbol{x} - \tilde{\boldsymbol{\mu}}_c)^2$$

$$\boldsymbol{w}^* = \arg\max_{\boldsymbol{w}} J(\boldsymbol{w}) = \arg\max_{\boldsymbol{w}} \frac{|\tilde{\boldsymbol{\mu}}_1 - \tilde{\boldsymbol{\mu}}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2} \tag{1}$$

LDA – Belső,- és külső szóródási mátrixok

- c osztályú transzformált pontok belső szóródási mátrixa:

$$S_c = \sum_{\{(\boldsymbol{x}_i, y_i) | y_i = c\}} (\boldsymbol{x}_i - \boldsymbol{\mu}_c)(\boldsymbol{x}_i - \boldsymbol{\mu}_c)^\mathsf{T}$$

- Aggregált belső szóródási mátrix: $S_W = S_1 + S_2$
- c osztályba eső transzformált pontok szóródása:

$$\tilde{s}_c^2 = \sum_{\{(\boldsymbol{x}_i, y_i) | y_i = c\}} (\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i - \boldsymbol{w}^\mathsf{T}\boldsymbol{\mu}_c)^2 =$$
$$= \sum_{\{(\boldsymbol{x}_i, y_i) | y_i = c\}} \boldsymbol{w}^\mathsf{T}(\boldsymbol{x}_i - \boldsymbol{\mu}_c)(\boldsymbol{x}_i - \boldsymbol{\mu}_c)^\mathsf{T}\boldsymbol{w} = \boldsymbol{w}^\mathsf{T} S_c \boldsymbol{w}$$

- Eltérő osztályú transzformált pontok közötti szóródási mátrix: $S_B = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\mathsf{T}$
- Eltérő osztályú transzformált pontok közötti szóródás:

$$(\tilde{\boldsymbol{\mu}}_1 - \tilde{\boldsymbol{\mu}}_2)^2 = (\boldsymbol{w}^\mathsf{T}\boldsymbol{\mu}_1 - \boldsymbol{w}^\mathsf{T}\boldsymbol{\mu}_2)^2 = \boldsymbol{w}^\mathsf{T} S_B \boldsymbol{w}$$

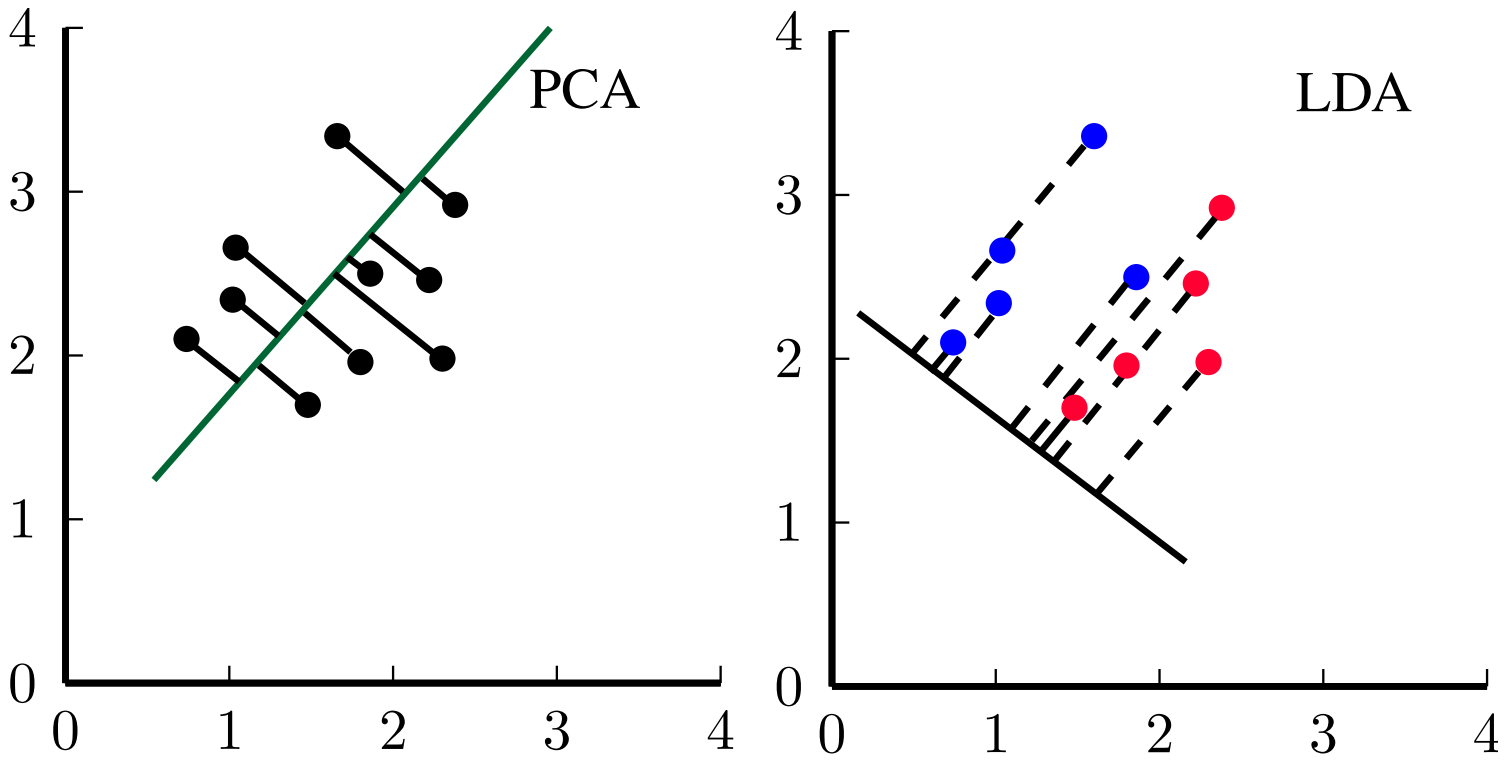
LDA

- Azaz az eredeti (1)-ben foglalt célunk átfogalmazható $\boldsymbol{w}^* = \arg\max_{\boldsymbol{w}} J(\boldsymbol{w}) = \arg\max_{\boldsymbol{w}} \frac{\boldsymbol{w}^\intercal S_B \boldsymbol{w}}{\boldsymbol{w}^\intercal S_W \boldsymbol{w}}$ alakra
  - $\frac{\boldsymbol{w}^\intercal S_B \boldsymbol{w}}{\boldsymbol{w}^\intercal S_W \boldsymbol{w}}$ az ún. általánosított Rayleigh-hányados
- $J(\mathrm{w})$ maximális amennyiben $\nabla \frac{\boldsymbol{w}^\intercal S_B \boldsymbol{w}}{\boldsymbol{w}^\intercal S_W \boldsymbol{w}} = 0 \Leftrightarrow S_B \boldsymbol{w} = \lambda S_W \boldsymbol{w} \Leftrightarrow S_W^{-1} S_B \boldsymbol{w} = \lambda \boldsymbol{w} \Leftrightarrow \boldsymbol{w} = S_W^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$

Emlékeztetőül

- $\left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x)g(x) - f(x)g'(x)}{g^2(x)}$
- $\nabla_{\boldsymbol{x}} \boldsymbol{x}^\intercal A \boldsymbol{x} = 2A\boldsymbol{x}$, amennyiben $A = A^\intercal$
- $\boldsymbol{x}\boldsymbol{x}^\intercal y = \left(\sum_{i=1}^{n} x_i y_i\right) \boldsymbol{x}$ (vagyis egy $\boldsymbol{x}$ irányába mutató vektor)

LDA vs. PCA

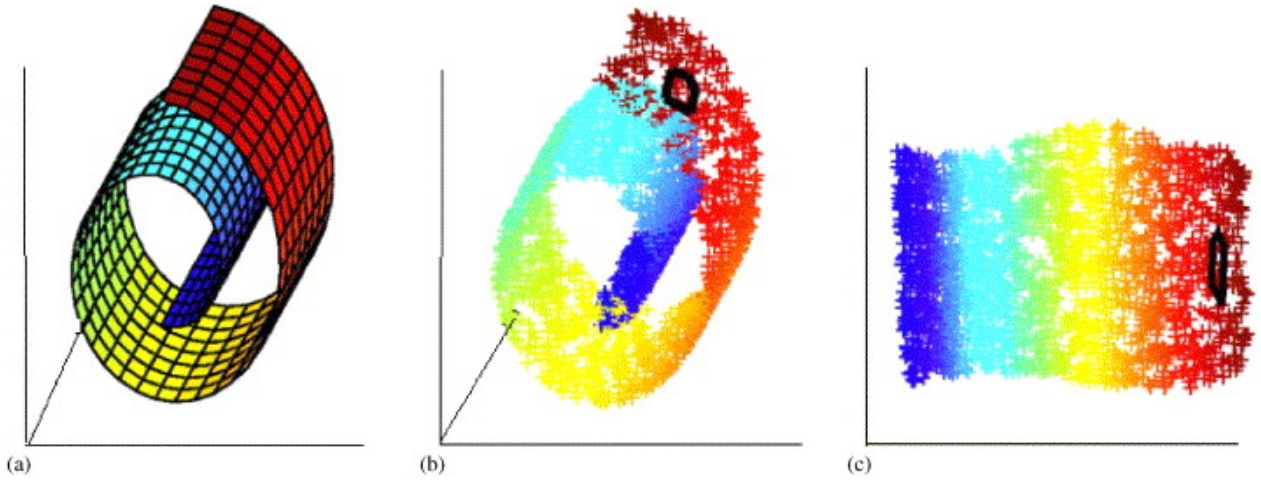


Multidimenziós skálázás (Multi-Dimensional Scaling)

- Cél: adott Δ pontpáronkénti költségeket/távolságokat tartalmazó mátrixhoz keressük meg az adatpontok azon pozícióit a térben, melyre $\|x_i - x_j\| \approx \delta_{ij}$
- Több/sokdimenziós adatpontok transzformációja alacsonyabb dimenzióba úgy, hogy a pontok közötti távolságok minél inkább megőrződjenek

Lokális lineáris beágyazás (Locally Linear Embedding)

- A PCA, SVD és LDA mind lineáris függést feltételeztek
- Nemlineáris dimenziócsökkentő eljárás
- Trükk: Minden ponthoz határozzuk meg a legközelebbi szomszédait, és írjuk fel őket azok lineáris kombinációiként
- $J(W) = \sum_{i=1}^{n} \|\boldsymbol{x}_i - \sum_{j=1}^{n} W_{ij}\boldsymbol{x_j}\|^2$, úgy, hogy $\sum_{j=1}^{n} W_{ij} = 1$, és $w_{ij} > 0 \Leftrightarrow x_j \in neighbors(x_i)$



Kanonikus korreláció analízis

- A pontjaink koordinátái két kordinátarendszer szerint ismertek
- Feladat: a két koordinátarendszer pontjainak egy (csökkentett dimenziójú) koordinátarendszerbe való összegyúrása, hogy a transzformált jellemzők korrelációja maximális legyen
- $\rho = \frac{\mathbb{E}\left[xy\right]}{\sqrt{\mathbb{E}\left[x^2\right]\mathbb{E}\left[y^2\right]}} = \frac{\mathbb{E}[w_x^\intercal x y^\intercal w_y]}{\sqrt{\mathbb{E}[w_x^\intercal x x^\intercal w_x^\intercal]\,\mathbb{E}\left[w_y^\intercal y y^\intercal w_y^\intercal\right]}} = \frac{w_x^\intercal C_{xy} w_y}{\sqrt{w_x^\intercal C_{xx} w_x w_y^\intercal C_{yy} w_y}}$
- $\arg\max \rho$-t nem befolyásolja w_x vagy w_y hossza $\Rightarrow \arg\max \rho = \arg\max w_x^\intercal C_{xy} w_y$
- $\Sigma = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix} = \mathbb{E}\left[\binom{x}{y}\binom{x}{y}^\intercal\right]$