Calcul Numeric

Laboratorul 3.

Descompunerea Valorilor Singulare

1 Descompunerea valorilor singulare

Pentru orice matrice $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ există matricile ortogonale $\boldsymbol{U} \in \mathbb{R}^{m \times m}$ și $\boldsymbol{V} \in \mathbb{R}^{n \times n}$ astfel încât

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\top} \tag{1}$$

unde $\boldsymbol{S} \in \mathbb{R}^{m \times n}$ este o matrice diagonală având r elemente nenule, unde r reprezintă rangul matriici $\boldsymbol{A}$. Elementele nenule σ_i ale lui $\boldsymbol{S}$ sunt ordonate descrescător $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r > 0$. Acestea se numesc valorile singulare ale matricii $\boldsymbol{A}$. Coloanele lui $\boldsymbol{U}$ reprezintă vectorii singulari la stângă, în timp ce coloanele lui $\boldsymbol{v}^{\top}$ vectorii singulari la dreapta.

Valorile singulare ale lui $\boldsymbol{A}$ sunt rădăcinile pătrate ale valorilor proprii ale matricii $\boldsymbol{A}^{\top}\boldsymbol{A}$.

Descompunerea valorilor singulare reprezintă o schimbare de bază. Privind matricea $\boldsymbol{A}$ ca o transformare, spunem că ea transformă spațiul $\mathbb{R}^m$ în spațiul $\mathbb{R}^n$. De aceea, o reprezentare utilă a matricii presupune găsirea unei perechi de baze pentru cele două spații. Matricile $\boldsymbol{U}$ și $\boldsymbol{V}$ constituie o alegere potrivită pentru aceaste baze, iar reprezentarea lui $\boldsymbol{A}$ în raport cu acestea este o matrice diagonală.

1.1 Rangul unei matrici

Noțiunea de rang al unei matrici reprezintă numărul de coloane liniar independente. În practică, pentru a calcula rangul se folosește observația că acesta este egal cu numărul de valori singulare nenule.

Prezența zgomotului în date poate masca rangul real al unei matrici. Fie $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ o matrice de rang nemaximal r, cu $r < min(m, n)$. Prin introducerea de variații mici, zgomotul perturbă liniaritatea coloanelor din $\boldsymbol{A}$, făcând ca aceasta să aibă rang maxim. Prezența zgomotului poate fi evaluată din valorile singulare ale matricii. Valorile $\sigma_i < \epsilon$ pot fi considerate neglijabile, ele corespunzând zgomotului. Valoarea lui ϵ poate fi aleasă empiric.

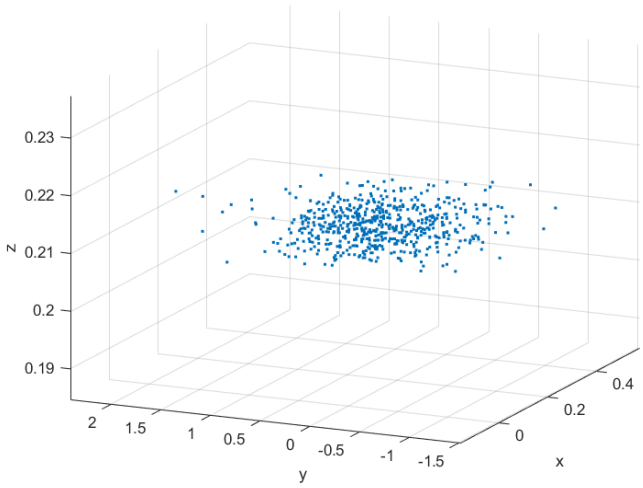

Figura 1: Dimensiunea datelor poate fi redusă.

Observația de mai sus duce la o altă aplicație a SVD: aproximarea de rang mic a unei matrici. Aceasta poate fi văzută ca o compresie a informației din matricea $\boldsymbol{A}$ având rang r și presupune aproximarea matricii prin trunchierea valorilor singulare. Numim $\boldsymbol{A_k}$ o aproximare de rang k a lui $\boldsymbol{A}$, cu $k < r$

$$\boldsymbol{A_k} = \sum_{j=1}^{k} \sigma_j \boldsymbol{u}_j \boldsymbol{v}_j^\top = \boldsymbol{U}[:, 1:k]\boldsymbol{S}[1:k, 1:k]\boldsymbol{V}[:, 1:k]^\top \tag{2}$$

Matricea $\boldsymbol{A_k}$ va conține doar informația *importantă* din $\boldsymbol{A}$.

2 Analiza componentelor principale

Secțiunea aceasta prezintă motivația ultimei afirmații de mai sus. Semnalele întâlnite în practică pot avea dimensiune mare, însă o parte din informația conținută în ele poate fi redundantă sau nesemnificativă. Aceasta înseamnă că ele pot fi aproximate cu semnale de dimensiune mai mică, fără a pierde informație relevantă (pentru aplicația în cauză). O astfel de reprezentare are și avantajul unui cost de calcul mai mic în rezolvarea problemelor cu setul respectiv de date.

Figura 1 prezintă o astfel de situație. Cu toate că datele au trei dimensiuni, se observă că ele pot fi aproximate în două dimensiuni, deoarece variația pe axa z este nesemnificativă (în raport cu variația pe celelalte două axe).

Analiza componentelor principale (Principal Component Analysis - PCA) este un algoritm de reducere a dimensiunii semnalelor ce folosește principiul variației maxime. Ne dorim să reținem din date doar trăsăturile semnificative. Prin urmare vom căuta direcțiile pe care varianța este mare și le vom elimina

pe cele în care datele au varianță mică. PCA găsește aceste direcții, numite componente principale. Prima astfel de direcție este cea care minimizează media pătratelor distanțelor de la puncte la acea dreaptă. Următoarea direcție este aleasă extrăgând in matricea inițială contribuția primei direcții și calculul noii drepte care minimizează media pătratelor distanțelor.

Această procedură este echivalentă cu calculul descompunerii valorilor singulare ale matricii $\boldsymbol{A}$, deci cu descompunerea valorilor proprii ale lui $\boldsymbol{A}^\top \boldsymbol{A}$. Componentele principale sunt vectorii proprii ai $\boldsymbol{A}^\top \boldsymbol{A}$. Pentru reducerea dimensiunii este suficientă deci alegerea unui număr p de vectori proprii.

Pentru a calcula PCA este necesară o etapă preliminară de centrare a datelor. Această operație presupune scăderea din fiecare coloană a mediei respectivei coloane, astfel încât fiecare nouă coloană are medie 0. O altă etapă de preprocesare, utilă atunci când variațiile coloanelor diferă semnificativ (spre exemplu datorită faptului că reprezintă trăsături ale semnalului ce corespund unor mărimi fizice diferite ca scală) este operația de standardizare. Pentru aceasta, fiecare coloană se împarte la deviația standard ale elementelor coloanei.

3 Ghid Python

Pentru a rezolva exercițiile din laboratorul de astăzi, aveți nevoie de bibliotecile `numpy`, `scipy` și `pandas`, modulele `random.sample`, `sklearn.decomposition.PCA` `matplotlib.image` și `matplotlib.pyplot`.

Pentru a calcula rangul unei matrici folosiți `numpy.linalg.matrix_rank(A)`.

Pentru a selecta aleator un număr de k elemente dintr-un șir crescător de numere $[1 : n]$, puteți folosi `random.sample(range(n),k)`.

Pentru a crea un zgomot Gaussian de medie m și dispersie d, folosiți `numpy.random.normal(m,d,size)`, unde *size* reprezintă dimensiunea vectorului/matricii.

Pentru a calcula descopunerea valorilor singulare pentru o matrice, utilizați `U, S, V = numpy.linalg.svd(A)`.
Funcția generează matricea $\boldsymbol{V}$ deja transpusă.

Pentru a încărca o imagine utilizați `matplotlib.image.imread(nume_imagine)`.
Pentru a afișa o imagine salvată în matricea $\boldsymbol{A}$, folosiți
`matplotlibpyplot.imshow(A)`.

Pentru a obține o matrice diagonală dintr-un vector, `numpy.diag(A)`.

Pentru a realiza reducerea dimensională a unui set de date $\boldsymbol{X}$ cu algoritmul PCA, utilizați următoarele instrucțiuni, după ce ați încărcat modulul `sklearn.decomposition.PCA`

```
pca = PCA(n_components=comp)
    components = pca.fit_transform(X)
```

alegând un număr *comp* de componente relevant pentru dimensiunea semnalelor. Variabila *components* va conține, pe coloane, direcțiile principale ale $\boldsymbol{X}$.

Pentru a importa un fișier `.csv` în format `pandas` dintr-o adresă url, utilizați `data = pandas.read_csv('https://example.com/')`.

4 Exerciții

1. În Secțiunea 1.1 se menționează faptul că prezența zgomotului poate *ascunde* rangul real al unei matrici. Testați aplicabilitatea SVD în determinarea rangului.
   (a) Creați o matrice aleatoare $\boldsymbol{A} \in \mathbb{R}^{m \times r}$, unde $m = 10$ și $r = 4$. Calculați rangul matricii.
   (b) Adăugați lui $\boldsymbol{A}$ un număr de 4 coloane liniar dependente astfel încât noua matrice va avea dimensiunea $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, cu $n = 8$. Pentru a crea noile coloane, luați o combinație liniară de $c = 3$ coloane deja existente în $\boldsymbol{A}$. Ați obținut astfel o matrice de rang nemaximal. Verificați, calculând din nou rangul matricii.
   (c) Adăugați la matricea obținută anterior un zgomot Gaussian de medie 0 și dispersie 0.2. Calculați rangul noii matrici.
   (d) Calculați descompunerea valorilor singulare pentru matricea de mai sus și afișați valorile singulare. Câte din aceste valori sunt neglijabile? Puteți deduce rangul matricii neafectate de zgomot din aceste valori? Comentați.
2. Comprimați o imagine folosind SVD.
   (a) Încărcați o imagine oarecare (.bmp, .jpg etc) într-o matrice și calculați SVD.
   (b) Alegeți un rang k pentru care vreți să obțineți aproximarea matricii de mai sus, astfel încât $k < min(m, n)$, unde m și n reprezintă dimensiunile matricii. Obțineți aproximarea de rang k a matricii, trunchiind valorile singulare.
   (c) Vizualizați noua imagine.
   (d) Repetați pentru $2 - 3$ valori ale lui k alese în funcție de dimensiunea imaginii, pentru a vedea gradul de compresie în fiecare caz.
3. Utilizați PCA pentru a reduce dimensiunea semnalelor din baza de date `iris`, disponibilă la adresa `https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data`. Alegeți numărul de componente principale $n_{components} = 2$ și vizualizați noul set de date.