Cursul 14

Modele liniare pentru date. Reducerea dimensiunii datelor folosind PCA

Odată cu dezvoltarea tehnologiei calculatoarelor şi a tehnologiei internet, marile companii, institutele de cercetare, organizaţiile mondiale de statistică, reuşesc să înregistreze un volum imens de date, care nu pot fi studiate şi devin inutile, pe măsură ce dimensiunea şi complexitatea acestora creşte. Pentru a putea extrage informaţie şi cunoaştere din date, a început să se acorde atenţie deosebită problemei reducerii dimensiunii acestora, în domenii ca data mining, machine learning, computer vision, regăsirea informaţiei stocate (information retrieval).

Scopul reducerii dimensiunii datelor este de a obţine o reprezentare mai compactă a acestora, cu o pierdere limitată de informaţie. Algoritmii tradiţionali de reducere a dimensiunii datelor se bazează pe modelul vectorial al acestora. Conform acestui model, datele numerice, ce cuantifică m atribute (proprietăţi) ale unor entităţi, sunt reprezentate de un vector din $\mathbb{R}^m$. Un eşantion constând din n entităţi este observat sau măsurat şi colecţia celor n vectori, notaţi $d_1, d_2, \ldots, d_n$, ce înregistrează valorile de observaţie definesc o matrice, D, matricea datelor:

$$D = [d_1|d_2|\ldots|d_n], \quad d_j \in \mathbb{R}^m$$

În limbaj statistic, se consideră că $d_1, d_2, \ldots, d_n$ constituie un eşantion de volum n, de observaţii asupra a n variabile, m–dimensionale:

$$X^1, X^2, \ldots, X^n$$

Un model liniar al acestor date este un subspaţiu vectorial $S \subset \mathbb{R}^m$, în sensul că vectorii din S înglobează caracteristicile principale ale variabilelor observate, excluzând informaţia redundantă şi zgomotul din datele înregistrate. Complexitatea modelului este caracterizată de dimensiunea subspaţiului respectiv. Având datele $d_1, d_2, \ldots, d_n$, se determină un model S, prin *data fitting*, adică modelul se potriveşte cel mai bine, într-un sens precizat, datelor.

14.1 Problematica PCA

Fie $D = (x_i^j)$, $i = \overline{1,m}, j = \overline{1,n}$, $n >> m$, o matrice de date. Coloanele matricii sunt constituite dintr-un eşantion de n observaţii asupra unor variabile m–dimensionale,

$X^1, X^2, \ldots, X^n$, ale căror coordonate cuantifică m trăsături (caracteristici) ale indivizilor populaţiei investigate. Volumul eşantionului este cu mult mai mare decât numărul caracteristicilor investigate.

$$D = \begin{array}{c} \\ \text{caract. } 1 \to \\ \text{caract. } 2 \to \\ \vdots \\ \text{caract } m \to \end{array} \begin{bmatrix} X^1 & X^2 & \ldots & X^n \\ \downarrow & \downarrow & \ldots & \downarrow \\ x_1^1 & x_1^2 & \ldots & x_1^n \\ x_2^1 & x_2^2 & \ldots & x_2^n \\ \vdots & \vdots & \ldots & \vdots \\ x_m^1 & x_m^2 & \ldots & x_m^n \end{bmatrix} \begin{array}{l} \\ \to \hat{m}_1 = \dfrac{1}{n}(x_1^1 + x_1^2 + \ldots + x_1^n) \\ \to \hat{m}_2 = \dfrac{1}{n}(x_2^1 + x_2^2 + \ldots + x_2^n) \\ \vdots \\ \to \hat{m}_m = \dfrac{1}{n}(x_m^1 + x_m^2 + \ldots + x_m^n) \end{array}$$

Scopul analizei componentelor principale (cunoscută în literatura de specialitate ca PCA=*Principal Component Analysis*) este de a determina direcţia pe care proiecţia datelor centrate (cu centrul geometric în originea axelor de cootdonate) are cea mai mare dispersie. Proiecţia pe un subspaţiu generat de direcţii, ce satisfac proprietatea de maxim a dispersiei în lungul lor, conduce la filtrarea zgomotului din date şi relevarea structurii ascunse.

Având dat norul de n puncte din $\mathbb{R}^m$, îi determinăm centrul, adică centrul de greutate al acestuia, notat

$$\hat{m} = (\hat{m}_1, \hat{m}_2, \ldots, \hat{m}_m)^T$$

Coordonata $\hat{m}_i$, reprezintă media caracteristicii i, observate (media aritmetică a elementelor de pe linia i a matricii datelor):

$$\hat{m}_i = \frac{1}{n}(x_i^1 + x_i^2 + \cdots + x_i^n), i = \overline{1, m}$$

Notând cu $e = (1, 2, \ldots, 1)^T \in \mathbb{R}^n$ vectorul ce are toate coordonatele egale cu 1, centrul de greutate al norului de puncte se obţine matricial astfel.

$$\hat{m} = \frac{1}{n} De$$

Analiza datelor este mult simplificată dacă acestea sunt centrate. Din punct de vedere geometric, datele centrate sunt datele iniţiale raportate la un nou sistem de coordonate. Şi anume, se consideră iniţial că norul de n puncte m–dimensionale:

$$X^1(x_1^1, x_2^1, \ldots, x_m^1), X^2(x_1^2, x_2^2, \ldots, x_m^2), \ldots, X^n(x_1^n, x_2^n, \ldots, x_m^n)$$

este raportat la sistemul de axe $Ox_1x_2 \ldots x_m$. Efectând o translaţie cu originea în centrul de greutate $\hat{m}$, $O'(\hat{m}_1, \hat{m}_2, \ldots, \hat{m}_m)$ şi notând noile axe cu $O'x_1'x_2' \ldots x_m'$, coordonatele unui punct din nor raportat la acest sistem sunt: $(x'^k_1, x'^k_2, \ldots, x'^k_m)$, cu $x'^k_i = x_i^k - \hat{m}_i$,

$i = \overline{1,m}$, $k = \overline{1,n}$. Astfel matricea datelor raportate la noul sistem de axe este:

$$D' = D - \frac{1}{n} Dee^T =$$

$$= D - \frac{1}{n} \begin{bmatrix} x_1^1 & x_1^2 & \dots & x_1^n \\ x_2^1 & x_2^2 & \dots & x_2^n \\ \vdots & \vdots & \dots & \vdots \\ x_m^1 & x_m^2 & \dots & x_m^n \end{bmatrix} \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & \dots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix} =$$

$$= D - \begin{bmatrix} \hat{m}_1 & \hat{m}_1 & \dots & \hat{m}_1 \\ \hat{m}_2 & \hat{m}_2 & \dots & \hat{m}_2 \\ \vdots & \vdots & \dots & \vdots \\ \hat{m}_m & \hat{m}_m & \dots & \hat{m}_m \end{bmatrix}$$

şi coloanele ei pot fi interpretate ca observaţii asupra variabilelor centrate

$$X'^1 = \begin{bmatrix} X_1^1 - \hat{m}_1 \\ X_2^1 - \hat{m}_2 \\ \vdots \\ X_m^1 - \hat{m}_m \end{bmatrix}, X'^2 = \begin{bmatrix} X_1^2 - \hat{m}_1 \\ X_2^2 - \hat{m}_2 \\ \vdots \\ X_m^2 - \hat{m}_m \end{bmatrix}, X'^n = \begin{bmatrix} X_1^n - \hat{m}_1 \\ X_2^n - \hat{m}_2 \\ \vdots \\ X_m^n - \hat{m}_m \end{bmatrix}$$

Centrul datelor D' este acum originea noului sistem de coordonate.

Exemplul 1. Pentru $n = 10$ pacienţi, care au suferit un preinfarct sunt analizate nivelul valorilor colesterolului, glicemiei şi a nivelului lipidelor (deci se înregistrează $m = 3$ atribute ale pacienţilor). Matricea datelor culese[1] este următoarea:

$$D = \begin{array}{l} \text{colest.} \rightarrow \\ \text{glicemie} \rightarrow \\ \text{lipide} \rightarrow \end{array} \begin{bmatrix} 280 & 313 & 250 & 345 & 210 & 245 & 320 & 290 & 350 & 270 \\ 125 & 180 & 115 & 200 & 95 & 100 & 150 & 97 & 160 & 120 \\ 540 & 400 & 600 & 780 & 640 & 615 & 800 & 520 & 744 & 700 \end{bmatrix}$$

Centrul norului constituit din cele 10 puncte este $\hat{m} = (287.3, 134.2, 633.9)^T$.

Datele centrate, adică elementele matricii $D' = D - \frac{1}{10} Dee^T$, sunt:

$$D' = \begin{bmatrix} -7.3 & 25.7 & -37.3 & 57.7 & -77.3 & -42.3 & 32.7 & 2.7 & 62.7 & -17.3 \\ -9.2 & 45.8 & -19.2 & 65.8 & -39.2 & -34.2 & 15.8 & -37.2 & 25.8 & -14.2 \\ -93.9 & -233.9 & -33.9 & 146.1 & 6.1 & -18.9 & 166.1 & -113.9 & 110.1 & 66.1 \end{bmatrix}$$

Vizualizarea norului de puncte relativ la "sistemul natural" de axe al datelor şi respectiv la sistemul obţinut după centrare este dată în Fig. 14.1:

Pentru a evidenţia corelatiile dintre două câte două caracteristici, se asociază matricii de date D, respectiv matricii datelor centrate, matricea de covarianţă. Pentru

[1] Sunt date inventate de mine.

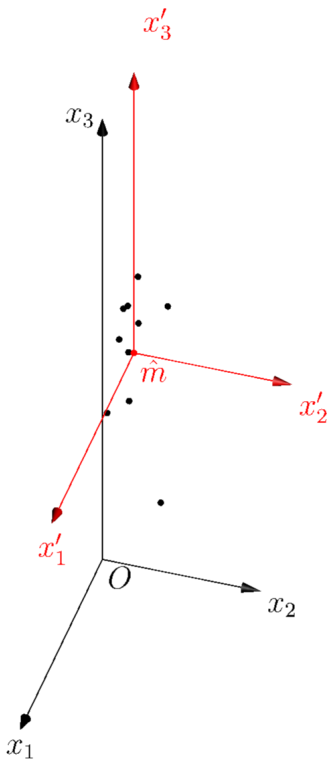

Fig.14.1: Datele raportate la sistemul de axe corespunzător măsurătorilor şi relativ la sistemul de axe translaltat în centrul norului.

a întelege definiţia elementelor acestei matrici, reamintim că dacă $x_1, x_2, \ldots, x_n$ sunt observaţii asupra variabilei aleatoare X şi $y_1, y_2, \ldots, y_n$ sunt observaţii asupra variabilei Y, atunci un estimator centrat al covarianţîei celor două variabile este:

$$\hat{\text{cov}} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}), \quad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Interpretând observaţiile de pe aceeaşi linie i a matricii datelor, ca fiind făcute asupra variabilei aleatoare ce măsoară o carateristică i a indivizilor, atunci covariantă a două linii i şi j dă intensitatea legăturii dintre cele două caracteristici, observate la indivizii,$1, 2, \ldots, n$.

Astfel definim matricea de covarianţă a datelor iniţiale, D, ca fiind matricea din $\mathbb{R}^{m\times m}$ ce are elementele C_{ij}, estimatorii covarianţelor (intensităţii legăturii) dintre caracteristicile i şi j ale celor n indivizi observaţi:

$$C_{ij} = \widehat{\text{cov}}_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_i^k - \hat{m}_i)(x_j^k - \hat{m}_j)$$

În cazul datelor centrate, vectorul mediilor este vectorul nul şi deci elementele matricii de

covarianţă sunt:

$$C_{ij} = \widehat{\text{cov}}_{ij} = \frac{1}{n-1}\sum_{k=1}^{n} x_i^k \cdot x_j^k = \frac{1}{n-1} \cdot \text{produsul scalar al liniei } i \text{ cu linia } j \text{ din } D'$$

Astfel matricea de covarianţă asociată datelor centrate este:

$$C = \frac{1}{n-1} D' D'^T$$

Matricea covarianţelor celor trei parametri măsuraţi pacienţilor este:

$$C = \begin{bmatrix} 2056 & 1390 & 1731 \\ 1390 & 1343 & 1065 \\ 1731 & 1065 & 15941 \end{bmatrix}$$

C_{11} reprezintă dispersia experimentală a colesterolului. Deci $\sqrt{C_{11}}$ reprezintă abaterea standard a valorilor observate a colesterolului de la media $\hat{m}_1 = 287.3$, etc.

În continuare presupunem că D este matricea datelor centrate. Asociem matricii de covarianţă a datelor, matricea împrăştierii, $S = (n-1)C$.

Din cursul 12 Algebră, rezultă că matricea $S = DD^T$ este o matrice simetrică şi (semi)pozitiv definită. Matricea de covarianţă diferă doar prin factorul $1/(n-1)$ de S, deci este şi ea simetrică şi semipozitiv definită. Rezultă astfel că C are m valori proprii mai mari sau egale cu zero.

Putem acum formula ce problemă ridică şi rezolvă analiza PCA a datelor centrate:

Dintre toate dreptele din spaţiul $\mathbb{R}^m$, ce trec prin centrul de greutate al norului de puncte, pe care dispersia (împrăstierea) proiecţiilor ortogonale a punctelor din nor, este maximă?

În Fig.14.2 este vizualizat un nor de puncte din $\mathbb{R}^2$, o dreaptă ce trece prin centrul norului, $\hat{m}$, şi proiecţiile ortogonale ale punctelor pe dreaptă.

Propoziţia 14.1.1 *Fie u o direcţie arbitrară în $\mathbb{R}^m$, $\|u\| = 1$ şi $z_1, z_2, \ldots, z_n$ proiecţiile ortogonale ale punctelor $X^1, X^2, \ldots, X^n \in \mathbb{R}^m$ pe direcţia u,*

$$z_i =< X^i, u >, i = \overline{1,n}$$

Dispersia (împrăstierea) datelor proiecţie pe dreapta de direcţie u este atunci

$$s^2 =< Cu, u >,$$

unde C este matricea de covarianţă a datelor.

Demonstraţie: Mai întâi o precizare. Din punct de vedere algebric, proiecţia unui vector v pe un versor u este: $pr_u(v) =< v, u > u$, adică vectorul coliniar cu u şi factorul de coliniaritate $< v, u >$ (vezi cursul aferent de la algebră). Considerând pe u ca unitatea de măsură pe axa de direcţie şi sens u, înseamnă că $< v, u >$ indică "câte unităţi de măsură

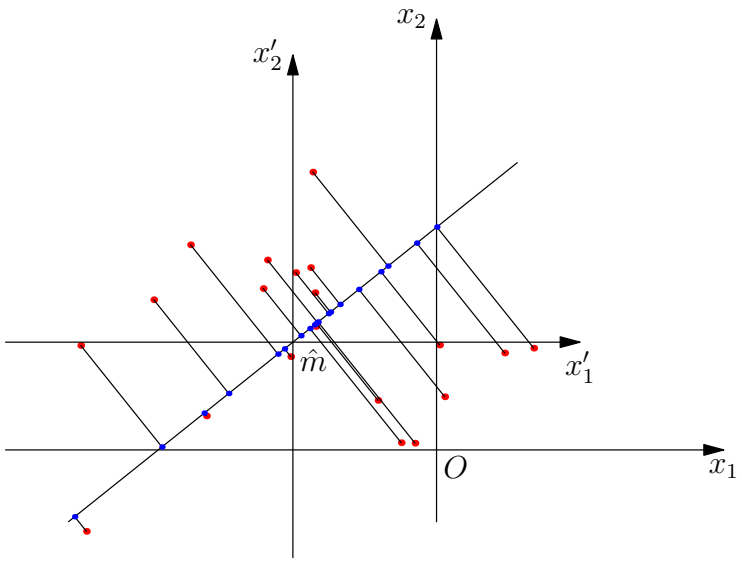

Fig.14.2: Proiecţia ortogonală a unor date 2D pe o dreaptă ce trece prin centrul norului de puncte.

are vectorul proiecţie". În continuare folosim doar aceste valori ale proiecttţiilor, adică doar $< v, u >$, nu vectorul $< v, u > u$.

Dispersia valorilor $z_1, z_2, \ldots, z_n$ este:

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(z_j - \overline{z})^2$$

unde $\overline{z}$ este media lor:

$$\overline{z} = \frac{1}{n}(z_1 + z_2 + \cdots + z_n) = \frac{1}{n}\sum_{j=1}^{n} < X^j, u > = < \frac{1}{n}\sum_{j=1}^{n} X^j, u > = < \hat{m}, u >$$

Prin urmare media valorilor proiecţie este proiecţia ortogonală, pe versorul u, a vectorului medie $\hat{m}$ asociat datelor. Dar dacă datele sunt centrate, atunci $\hat{m} = \theta \in \mathbb{R}^m$ şi deci $\overline{z} = \theta \in \text{span}(u)$. Înlocuind în expresia dispersiei obţinem:

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{j=1}^{n}(z_j - \underbrace{\overline{z}}_{=0})^2 = \frac{1}{n-1}\sum_{j=1}^{n}(< X^j, u >)^2 = \\ &= \frac{1}{n-1} < D^T u, D^T u > = \\ &= \frac{1}{n-1} < DD^T u, u > = < Cu, u > \end{aligned}$$

□

Observaţia 14.1.1 *Dispersia proiecţiilor datelor pe un vector propriu unitar, al matricii covarianţă, este valoarea proprie corepunzătoare:*

$$s^2 =< Cu, u >=< \lambda u, u >= \lambda \|u\|^2 = \lambda$$

Determinarea direcţiilor critice u, $\|u\| = 1$, pentru funcţia $f(u) = s^2(u) =< Cu, u >$(direcţiile critice sunt direcţiile pentru care diferenţiala funcţiei se anulează), revine la a determina punctele critice pentru problema de maxim cu restricţia $\|u\| = 1$:

$$g(u) =< Cu, u > -\lambda(\|u\|^2 - 1), \quad \lambda \in \mathbb{R}$$

Se demonstrează că aceste direcţii sunt direcţiile u, $\|u\| = 1$, cu proprietatea că:

$$(C - \lambda I_m)u = 0,$$

ceea ce este echivalent cu $Cu = \lambda\, u$, adică u este vector propriu unitar, corespunzător valorii proprii λ.

Dacă valorile proprii ale matricii de covarianţă sunt:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$$

şi $u_1, u_2, \ldots, u_m$ o bază ortonormată în $\mathbb{R}^m$ formată din vectori proprii, $Cu_i = \lambda_i u_i$, $i = \overline{1, m}$, atunci dispersia maximă este atinsă proiectând datele pe vectorul propriu unitar, u_1, corespunzător valorii maxime, λ_1.

Rezultă astfel că dispersiile $s^2_{u_i}$, ale proiecţiilor ortogonale ale datelor pe vectorii bazei $(u_1, u_2, \ldots, u_m)$ sunt ordonate astfel:

$$s^2_{u_1} \geq s^2_{u_2} \geq \cdots \geq s^2_{u_m}$$

14.2 Calculul numeric al PCA

codul `MATLAB` pentru calculul PCA al unei matrici de date:

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function [Y,PC,Lambda] = pcaCov(D)
% Functia pcaCov calculeaza PCA folosind covarianta datelor
% D - mxn este matricea datelor
% m=nr de  caracteristici, n=volum esantion
% Y - matricea datelor proiectie
% PC - este matricea ce are pe coloane, directiile principale
% Lambda - mx1 matricea dispersiilor pe directiile principale
[m,n] = size(D);
mestim = mean(D,2);% mestim este vectorul mediilor
D = D - repmat(mestim,1,n); %datele centrate
%repmat(mestim,1,n) este matricea:
```

```
%[mestim(1) mestim(1).... mestim(1);
% mestim(2) mestim(2).... mestim(2);...
% mestim(m) mestim(m).... mestim(m)]

C = 1 / (n-1) * D * D'; %C matricea de cov. a datelor centrate
[PC, U] = eig(C);%Coloanele lui PC sunt vectori proprii,
                 %val proprii pe diagonala lui U
Lambda = diag(U);% Lambda= vector ce contine valorile proprii
% sortam valorile proprii in ordine descrescatoare
[W, Ind] = sort(-Lambda);
 Lambda = Lambda(Ind);
 PC = PC(:,Ind);
% proiectam datele pe directiile principale
Y = PC'* D;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

Liniile:

```
mestim = mean(D,2);
 D = D - repmat(mestim,1,n);
```

calculează mai simplu şi mai rapid ceea ce ar realiza liniile:

```
e=ones(n,1);
D = D - D*e*e'/n; %datele centrate
```

ce implementează întocmai formulele teoretice de calcul, prezentate mai sus.

14.3 Calculul PCA folosind SVD

Direcţiile principale şi dispersiile în lungul lor, se pot calcula exploatând descompunerea SVD a unei matrici obţinută din scalarea matricii datelor, evitând astfel calculul covarianţei matricii datelor centrate şi sortarea valorilor proprii ale acesteia.

Deoarece matricea de covarianţă este o matrice simetrică şi pozitiv definită, rezultă că descompunerea ei SVD coincide cu descompunerea ortogonală şi anume valorile sale singulare coincid cu valorile proprii.

Descompunerea SVD a unei matrici mai simple decât matricea de covarianţă permite de asemenea calculul PCA. Şi anume, asociem matricii datelor centrate, matricea

$$D_s = \frac{1}{\sqrt{n-1}} D^T$$

Astfel $D_s^T D_s = \frac{1}{n-1} D D^T = C$ şi deci valorile proprii ale matricii $D_s^T D_s$ coincid cu valorile proprii ale matricii C, iar valorile singulare ale matricii D_s sunt rădăcinile pătrate

ale valorilor proprii ale matricii C, ordonate descrescător. Vectorii singulari drepţi ai matricii $D_s^T D_s = C$ sunt astfel direcţiile principale. Dispersiile pe direcţiile principale sunt pătratele valorilor singulare ale matricii D_s. Avem astfel următorul cod MATLAB ce calculează direcţiile principale şi dispersiile pe direcţiile principale:

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function [Y,PC,Lambda] = pcaSVD(D)
% Functia pcaSVD calculeaza PCA folosind descompunerea SVD
% a matricii datelor transpusa si scalata
% D - mxn este matricea datelor
% m=nr de  caracteristici, n=volum esantion
% Y - matricea datelor proiectie
% PC - este matricea ce are pe coloane, directiile princ.
% Lambda - matricea mx1 a dispersiilor  pe directiile princ.
[m,n] = size(D);
mestim = mean(D,2);
D = D - repmat(mestim,1,n); %datele centrate
% calculeaza matricea de Ds
Ds =  D'/sqrt(n-1);
% determina descompunerea SVD a matricii Ds
[U,Sig, PC] = svd(Ds);
% extragem valorile singulare de pe diagonala lui S
Sig = diag(Sig);
Lambda=Sig.*Sig; %calculeaza dispersiile pe PC
% proiectam datele pe directiile principale
Y = PC'* D;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

PCA dă rezulta bune când datele sunt n observaţii asupra unui vector aleator de m coordonate, normal distribuit, pentru că distribuţia normală este o distribuţie perfect caracterizată de vectorul medie şi matricea de covarianţă a coordonatelor vectorului aleator.

În continuare testăm PCA pe o matrice de date ce constă din n observaţii asupra unui vector aleator de 3 coordonate (cele trei coordonate modelează 3 caracteristici ale indivizilor unei populaţii), având distribuţia normală de medie $m = (-1, 2, 2.5)^T$ şi matrice de covarianţă S:

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%               testPCA.m
clear;
%(1,2,3) sunt trei caracteristici  normal
%distribuite,  N(medii, S)
medii=[-1; 2; 2.5];
%Covarianta celor trei caracteristici
S= [2.5800    2.0400   -3.6770;
```

```
    2.0400    4.6800   -0.3420;
   -3.6770   -0.3420   10.4686];
n=600;
D=zeros(3, n);
%simuleaza vect. aleator normal distrib, gener. n observatii
%ce se inregistreaza pe coloanele matricii D
C=chol(S);% descompunerea Choleski a matricii  S
for i=1:n
    z=randn(1,3);
    D(:,i)=medii+C'*z';
end
[m,n]=size(D);
figure(1);
 plot3(D(1,:), D(2,:),D(3,:),'r.');
 axis equal;
[Y,PC,Lambda] = pcaCov(D);
[Y1,PC1, Lambda1]=pcaSVD(D);
figure(2);
plot3(Y(1,:), Y(2,:), Y(3,:), 'r.');hold on;
axis equal;
figure(3)
%proiectam datele pe subspatiul generat de
%primele doua directii principale
plot(Y(1,:), Y(2,:),'g.'); axis equal;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

14.4 Calculul numeric iterativ al componentelor principale

Dacă matricea datelor are dimensiuni foarte mari, oricare din metodele prezentate mai sus este costisitoare din punctul de vedere al timpului de execuţie şi consumului de memorie. În acest caz se aplică o metodă iterativă de calcul a direcţiilor principale şi dispersiilor în lungul lor, exploatând faptul ca descompunerea SVD a matricii de covarianţă coincide cu descompunerea ei ortogonală, ca matrice simetrică. Deci, dacă $C \in \mathbb{R}^{m\times m}$ este matricea de covarianţă a caracteristicilor, asociată unei matrici de date centrate şi $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ sunt valorile sale proprii, cu vectorii proprii ortonormaţi corespunzători, $u_1, u_2, \ldots, u_m$, rezultă că C se exprimă ca şi combinaţie liniară a m matrici de rang 1:

$$C = \lambda_1 u_1 u_1^T + \lambda_2 u_2 u_2^T + + \cdots + \lambda_m u_m u_m^T$$

Ideea de calcul iterativ a valorilor şi vectorilor proprii este să se calculeze prin metoda puterii λ_1 şi u_1 şi apoi să se aplice aceeaşi metodă pentru matricea $C - \lambda_1 u_1 u_1^T$, etc.

Metoda iterativă porneşte cu o setare arbitrară pentru un vector unitar $u^{(0)} \in \mathbb{R}^m$ şi se calculează recursiv:

$$\begin{aligned} v^{(n)} &= Cu^{(n-1)} \\ u^{(n)} &= \frac{v^{(n)}}{\|v^{(n)}\|} \\ \lambda^{(n)} &= \|v^{(n)}\| \end{aligned} \tag{14.1}$$

Se demonstrează că $\lim_{n\to\infty} u^{(n)} = u_1$ şi $\lim_{n\to\infty} \lambda^{(n)} = \lambda_1$. După ce primul vector propriu şi valoarea proprie corespunzătoare au fost calculaţi matricea de covariantă este înlocuită cu:

$$C \leftarrow C - \lambda_1 u_1 u_1^T$$

Codul MATLAB al acestui algoritm este următorul:

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function [PC,Lambda]=pcaPutere(C)
%Calculeaza directiile principale, folosind metoda puterii
m = size(C,1);
epsilon = 1e-5;
maxIter = 1500;

for i=1:m
%initializam vectorul principal si il normalizam
v = randn(m,1); u = v./sqrt(v'*v);
%monitorizarea convergentei
eroarea = 1e15;
iter = 1;
%bucla principala de calcul a unui vector propriu
while (eroarea > epsilon) | (iter < maxIter)
 v = C*u;
 unou=v./sqrt(v'*v);
eroarea = sum((unou - u).^2);
u = unou;
%seteaza valoarea proprie
Lambda(i) = sqrt(v'*v);
iter = iter + 1;
end
%salvam in coloanele matricii PC, vectorii proprii gasiti
PC(:,i) = unou;
%recalculeaza noua matrice de cov
C = C - Lambda(i)*unou*unou';
end
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

14.5 Calculul numeric iterativ al componentelor principale

Calculul valorilor proprii şi al bazei ortonormate corespunzătoare a matricii de covarianţă, se realizează cu ajutorul metodei puterii. Şi anume se ţine seama că matricea de covarianţa a datelor centrate este simetrică şi semipozitiv definită şi deci descompunerea ei ortogonală $C = Q\text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_m)Q^T$ coincide cu descompunerea sa SVD, $C = U\Sigma V^T$. Adică $U = Q = [u_1|u_2|\ldots|u_m]$, iar $V = U$. Rezultă astfel că C se exprimă ca şi combinaţie liniară a cel mult m matrici de rang 1:

$$C = \lambda_1 u_1 u_1^T + \lambda_2 u_2 u_2^T + \cdots + \lambda_m u_m u_m^T$$

Ideea de calcul iterativ a valorilor şi vectorilor proprii este să se calculeze λ_1 şi u_1 prin metoda puterii, adică valoarea proprie dominantă (cea mai mare) şi vectorul propriu unitar corespunzător şi apoi să se aplice aceeaşi metodă pentru matricea $C_1 = C - \lambda_1 u_1 u_1^T$, pentru a calcula λ_2 şi u_2, etc. Matricea $C_i = C_{i-1} - \lambda_i u_i u_i^T$ este folosită de metoda puterii pentru a calcula λ_{i+1}, u_{i+1}.

Metoda iterativă porneşte cu o setare arbitrară pentru un vector unitar $u^{(0)} \in \mathbb{R}^m$ şi se calculează recursiv:

$$\begin{aligned} v^{(n)} &= Cu^{(n-1)} \\ u^{(n)} &= \frac{v^{(n)}}{\|v^{(n)}\|} \\ \lambda^{(n)} &= \|v^{(n)}\| \end{aligned} \tag{14.2}$$

Se demonstrează că $\lim_{n\to\infty} u^{(n)} = u_1$ şi $\lim_{n\to\infty} \lambda^{(n)} = \lambda_1$. După ce primul vector propriu şi valoarea proprie corespunzătoare au fost calculaţi matricea de covariantă este înlocuită cu:

$$C \leftarrow C - \lambda_1 u_1 u_1^T$$

Pseudocodul de calcul a valorilor şi vectorilor proprii ortonormaţi corespunzători, ai matricii de covarianţă a datelor centrate este următorul:

```
PCA(C,m){
 epsilon = 1e-5;
 maxIter = 1500;
 for(i=0;i<m;i++)
 {
    seteaza u, vector unitar;
    //monitorizarea convergentei
      eroarea = 1e15;
      iter = 1;
    //bucla principala de calcul a unui vector propriu
      while (eroarea > epsilon) || (iter < maxIter){
              v = C*u;
              unou=v/sqrt(v^t*v);
```

```
            eroarea = ||unou-u||^2;
            u = unou;
            //seteaza valoarea proprie
            lambda[i] = sqrt(v^t*v);
            iter = iter + 1;
            }
      //salvam in coloana i, Q_i,  a unei matrici Q, vectorul propriu  gasit
      Q_i = unou;
      //recalculeaza noua matrice de cov
      C = C - lambda[i]*unou*unou^t;
}
```

Dacă o parte dintre valorile proprii sunt foarte mici, să zicem $\epsilon \geq \lambda_{q+1}, \ldots, \lambda_m$, şi deci proiecţiile datelor pe aceste direcţii au dispersii aproape nule, informaţia relevantă conţinută în date se poate extrage proiectându-le pe subspaţiul vectorial generat de primii q vectori proprii, $S = \text{span}(u_1, u_2, \ldots, u_q)$. Coordonatele proiecţiei ortogonale a unui punct X^j din matricea de date, pe subspaţiul S se obţin efectuând produsul AX^j dintre matricea $A = [u_1|u_2|\ldots|u_q]^T \in \mathbb{R}^{q\times m}$ şi X^j (vezi cursul de algebră relativ la matricea proiecţiei ortogonale pe un subspaţiu). Notăm cu Y^j punctul proiecţie, $Y^j = P_S(X^j)$. Atunci matricea
$E = [AX^1|AX^2|\ldots|AX^n] = AD$ este matricea datelor obţinute prin proiecţie.

Proiecţia datelor pe subspaţiul S reprezintă practic operaţia de reducere a dimensiunii datelor, adică de reprezentarea acestora prin date dintr-un spaţiu de dimensiune mai redusă. Direcţiile $u_1, u_2, \ldots, u_q$ se numesc direcţii principale ale datelor. O astfel de transformare a datelor se numeşte transformarea Karhunen-Loéve.

Reducerea dimensiunii conduce la pierdere de informaţie. De aceea alegerea dimensiunii q de reprezentare a datelor se face în aşa fel încât să se minimizeze eroarea. În funcţie de contextul în care se abordează problema reducerii dimensiunii se alege q astfel încât:

$$\sum_{i=1}^{q} \lambda_i / \sum_{i=1}^{m} \lambda_i > \text{prag},$$

unde pragul poate fi $0.9, 0.95$, etc.

Propoziţia 14.5.1 *Datele rezultate prin transformarea Karhunen-Loéve a unor date centrate, D, sunt necorelate, adică au matricea de covarianţă, egală cu o matrice diagonală.*

Demonstraţie: Fie $T = [u_1|u_2|\ldots|u_m]$ o bază ortornormată formată din vectori proprii ai matricii de covarianţă $C = \dfrac{1}{n-1}DD^T$. $Cu_i = \lambda_i u_i$, $i = \overline{1,m}$ şi $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$, iar $B = [u_1|u_2|\ldots|u_q]$. Rezultă atunci că $T^tCT = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_m)$.

Datele rezulate prin transformarea Karhunen-Loéve sunt coloanele matricii $E = B^TD$, şi sunt la fel centrate. Matricea lor de covarianţă este atunci: $EE^T = (B^TD)(B^TD)^T = B^T(DD^T)B = B^TCB = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_q)$. Faptul că covarianţa a două câte două caracteristici (coordonate) distincte în noul spatiu de reprezentare este 0, rezultă că variabilele Y^j sunt necorelate. □

PCA dă rezulta bune când datele sunt n observaţii asupra unui vector aleator de m coordonate, normal distribuit, pentru că distribuţia normală este o distribuţie perfect caracterizată de vectorul medie şi matricea de covarianţă a coordonatelor vectorului aleator.

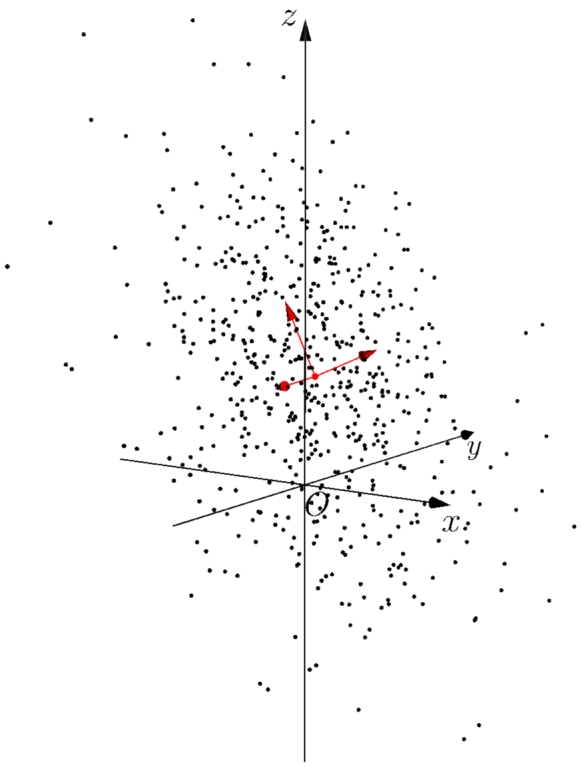

Fig.14.3: Nor de puncte 3D, centrul norului şi directiile principale

14.6 Aplicaţii ale analizei PCA

O primă aplicaţie a reprezentării reduse a datelor centrate, prin proiecţie ortogonală pe subspaţiul generat de primele două sau 3 direcţii principale, respectiv pe două câte două direcţii consecutive (u_i, u_{i+1}), $i = \overline{1,m}$ şi $u_{m+1} = u_1$, este în vizualizarea datelor. Datele multidimensionale se pot vizualiza prin proiecţie pe subspaţiii 2D, cel mult 3D. Vizualizarea proiecţiilor conduce la o imagine a distribuţiei acestora în jurul axelor principale.

În Fig.14.3 este vizualizat un 3D nor, cu centrul $\hat{m} = (-0.99, 1.99, 2.5)$ şi direcţiile principale, calculate folosind metoda puterii. Se observă că din figura 3D nu realizăm cât de mare este dispersia pe fiecare direcţie principală. De aceea în Fig.14.5 am vizualizat proiecţiile datelor centrate pe subspaţiile generate de (u_1, u_2), (u_2, u_3) şi respectiv (u_3, u_1).

Analiza PCA se aplică cu succes în computer vision (un domeniu al CS ce dezvoltă bazele teoretice şi tehnologiile de extragere a informaţiei din imagini digitale) şi anume

în probleme de clasificare (clusterizare), comprimarea imaginilor, recunoaşterea feţelor umane.

Spre deosebire de metoda de comprimare prin trunchierea descompunerii SVD a imaginii în ansamblu, PCA se aplică unei matrici date, D, asociată unei magini `Imag`, matricea D având coloanele formate din vectorii de dimensiune $m = q^2$, constituiţi din concatenarea liniilor consecutive ale unor blocuri adiacente de imagine, de dimensiuni $q \times q$ pixeli.

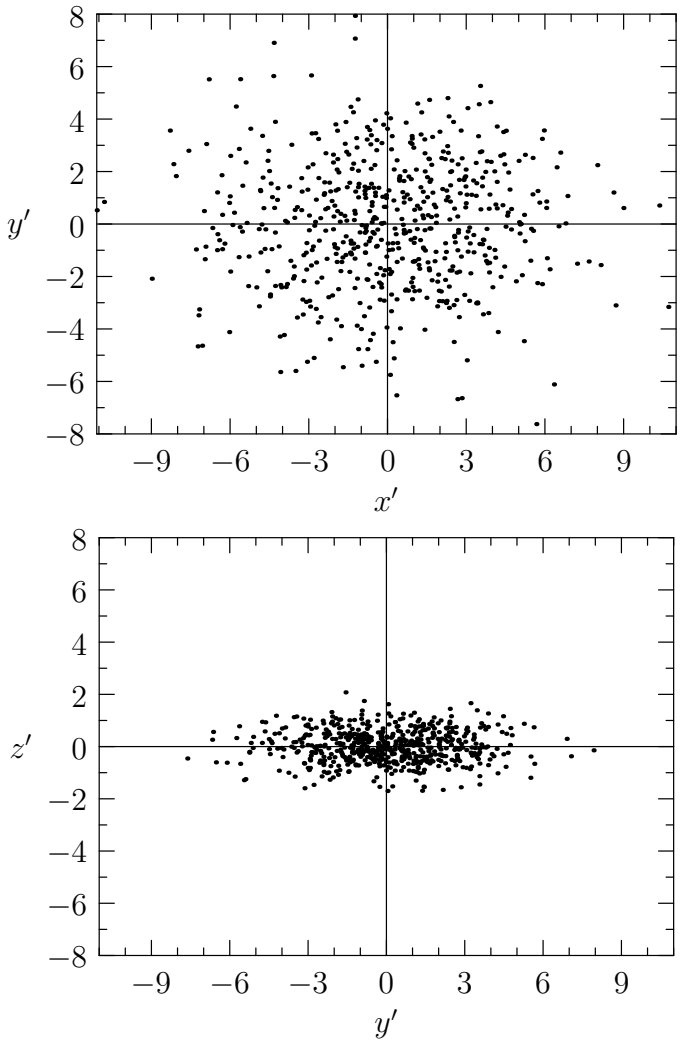

Fig.14.4: Vizualizarea proiecţiei ortogonale a norului 3D centrat pe planul generat de u_1 şi u_2, si u_2, u_3. Dispersiile pe cele trei direcţtii sunt respectiv, $\lambda_1 = 11.3146, \lambda_2 = 5.6080, \lambda_3 = 0.3887$.

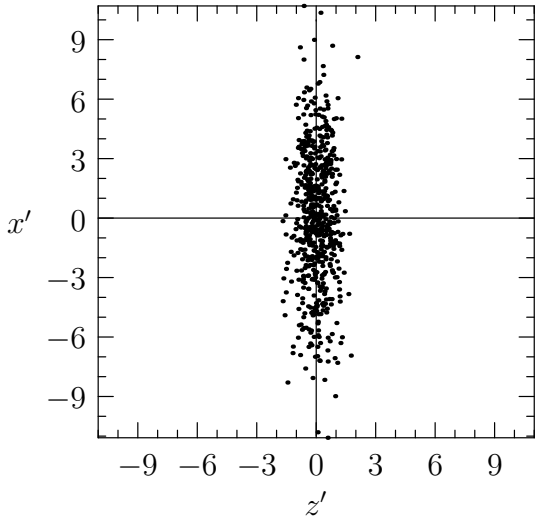

Fig.14.5: Vizualizarea proiecţiei ortogonale a norului 3D centrat pe planul generat de u_3, u_1, jos.