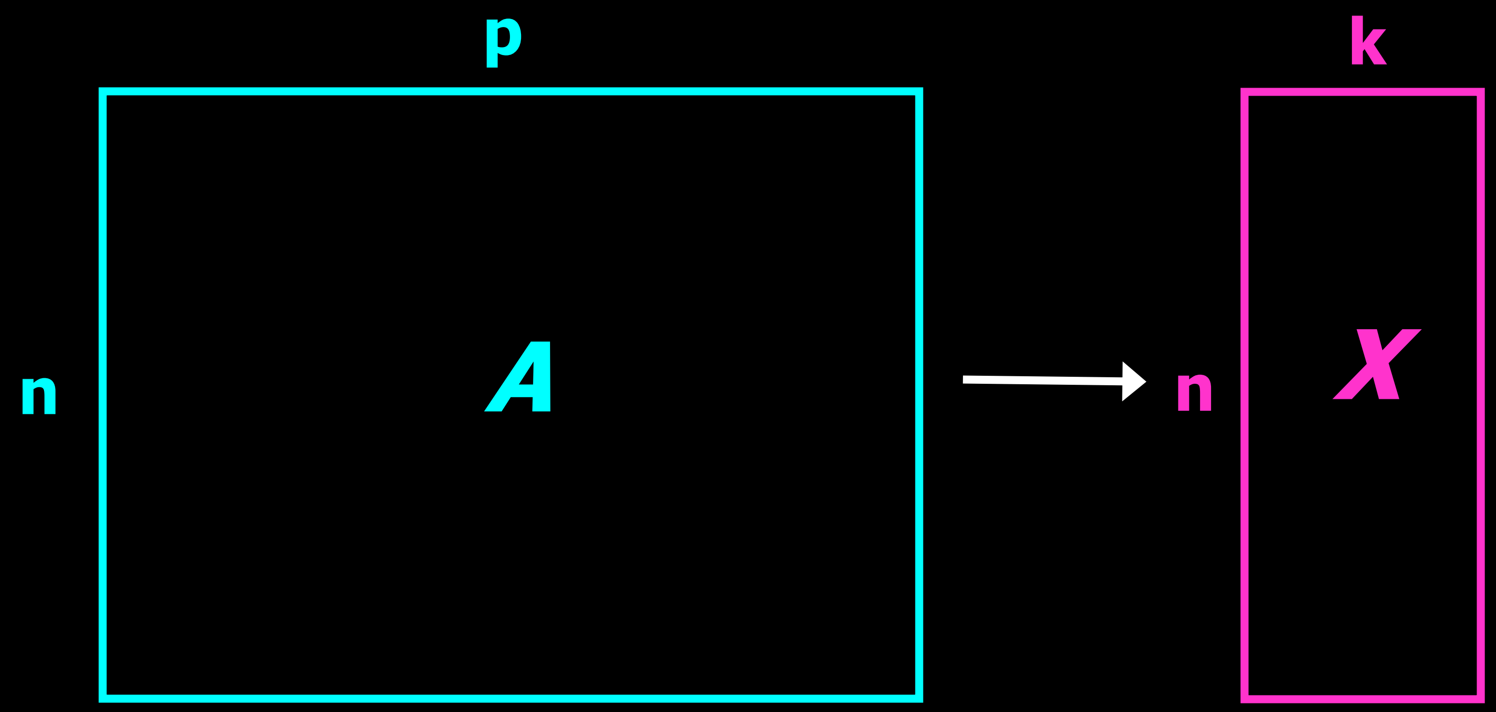
Principal Component Analysis (PCA)

Analiza componentelor principale

Reducerea dimensionalității

- rezumarea datelor cu multe (p) variabile printr-un set mai mic de (k) variabilee derivate (sintetice, compozite)



Reducera Dimensionalitatii

- Variația „reziduală” este informația din A care nu este păstrată în X
- act de echilibru între
  - claritatea reprezentării, ușurința înțelegerii
  - suprasimplificare: pierderea informațiilor importante sau relevante.

Principal Component Analysis (PCA)

- probabil cea mai utilizată și cunoscută metodă „standard” de analiză multivariată
- inventată de Pearson (1901) și Hotelling (1933)

Principal Component Analysis (PCA)

- Se da o matrice de date de n observații si p variabile, potențial corelate și o rezumă prin axe necorelate (componente principale sau axe principale) care sunt combinații liniare ale celor p variabile originale
- primele k componente vor explica cea mai mare parte din variația dintre variabile

Explicatii Geometrice pt PCA

- observațiile sunt reprezentate ca un nor de n puncte într-un spațiu multidimensional cu o axă pentru fiecare dintre cele p variabile
- **centroidul** punctelor este definit de media fiecărei variabile
- **variația** fiecărei variabile este abaterea medie pătrată ale celor n valori ale sale în jurul valorii medii:

$$V_i = \frac{1}{n-1} \sum_{m=1}^{n} \left(X_{im} - \overline{X}_i \right)^2$$

Explicatii Geometrice pt PCA

- **gradul în care variabilele sunt corelate liniar este reprezentat de covarianța lor.**

$$C_{ij} = \frac{1}{n-1}\sum_{m=1}^{n}\left(X_{im} - \overline{X}_i\right)\left(X_{jm} - \overline{X}_j\right)$$

Covarianta dintre variabilele *i* and *j*

Suma peste cele *n* observatii

Valoarea variabilei *i* in observatia *m*

Media variabilei *i*

Valoarea variabilei *j* in observatia *m*

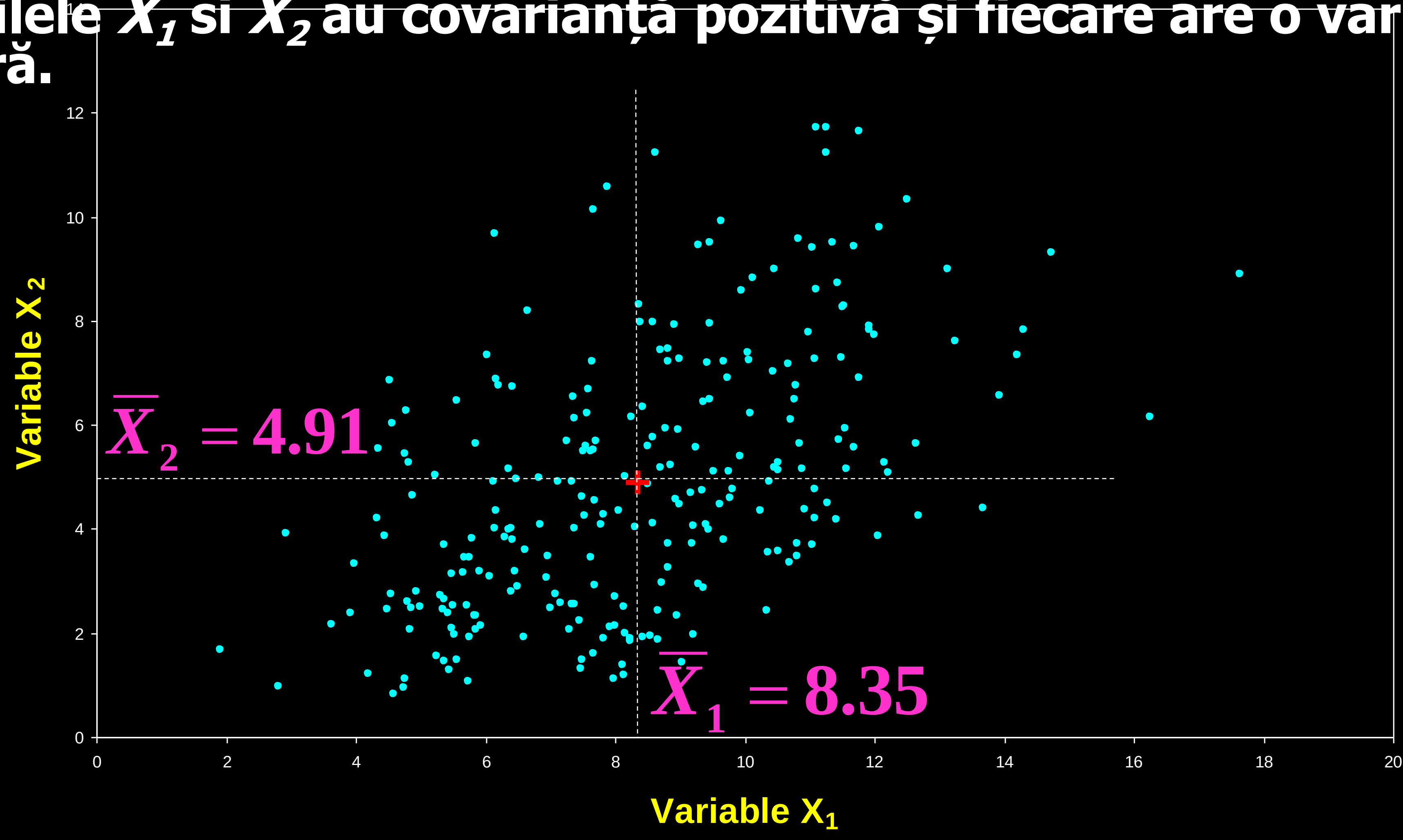
Media variabilei *j*

Explicatii Geometrice pt PCA

- obiectivul PCA este de a roti rigid axele acestui spațiu p-dimensional către noi poziții (axe principale) care au următoarele proprietăți:
  - ordonat astfel încât axa principală 1 să aibă cea mai mare varianță, axa 2 are următoarea cea mai mare varianță, …., iar axa p are cea mai mică varianță
  - covarianța dintre fiecare pereche de axe principale este zero (axele principale sunt necorelate)

Exemplu 2D de PCA

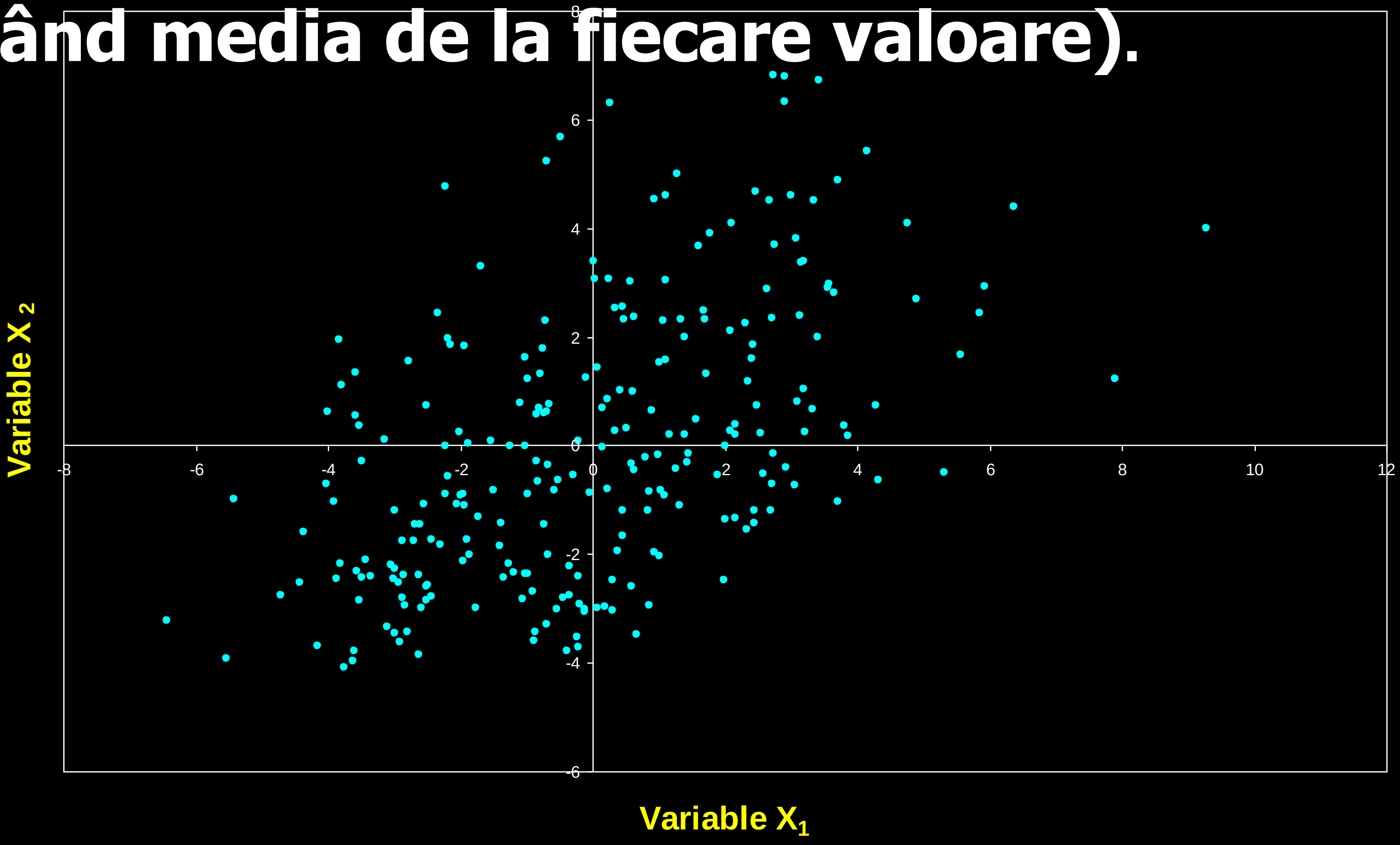
- **variabilele X_1 si X_2 au covarianță pozitivă și fiecare are o variație similară.**



$$V_1 = 6.67 \qquad V_2 = 6.24 \qquad C_{1,2} = 3.42$$

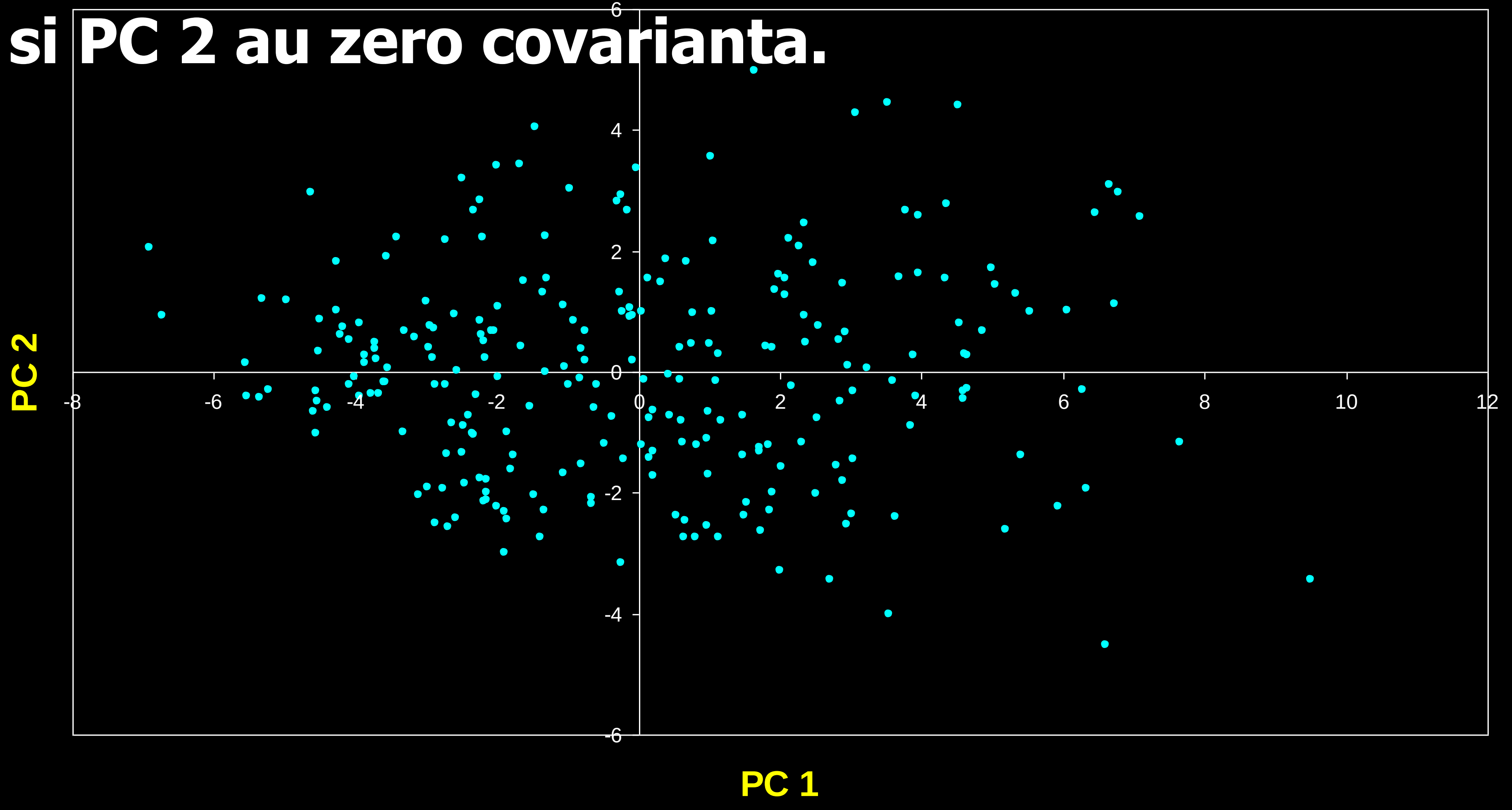
Centrarea variabilelor

- fiecare variabilă este ajustată la o medie de zero (scăzând media de la fiecare valoare).



Calcularea Componentelor Principale

- PC 1 are cea mai mare varianta (9.88)
- PC 2 are varianta de 3.03
- PC 1 si PC 2 au zero covarianta.



Măsura de disimilare folosită în PCA este distanța euclidiană

- PCA folosește distanța euclidiană calculată din variabilele p ca măsură a disimilarității dintre n obiecte

- PCA obține cea mai bună reprezentare k-dimensională (k<p) a distanțelor euclidiene între obiecte

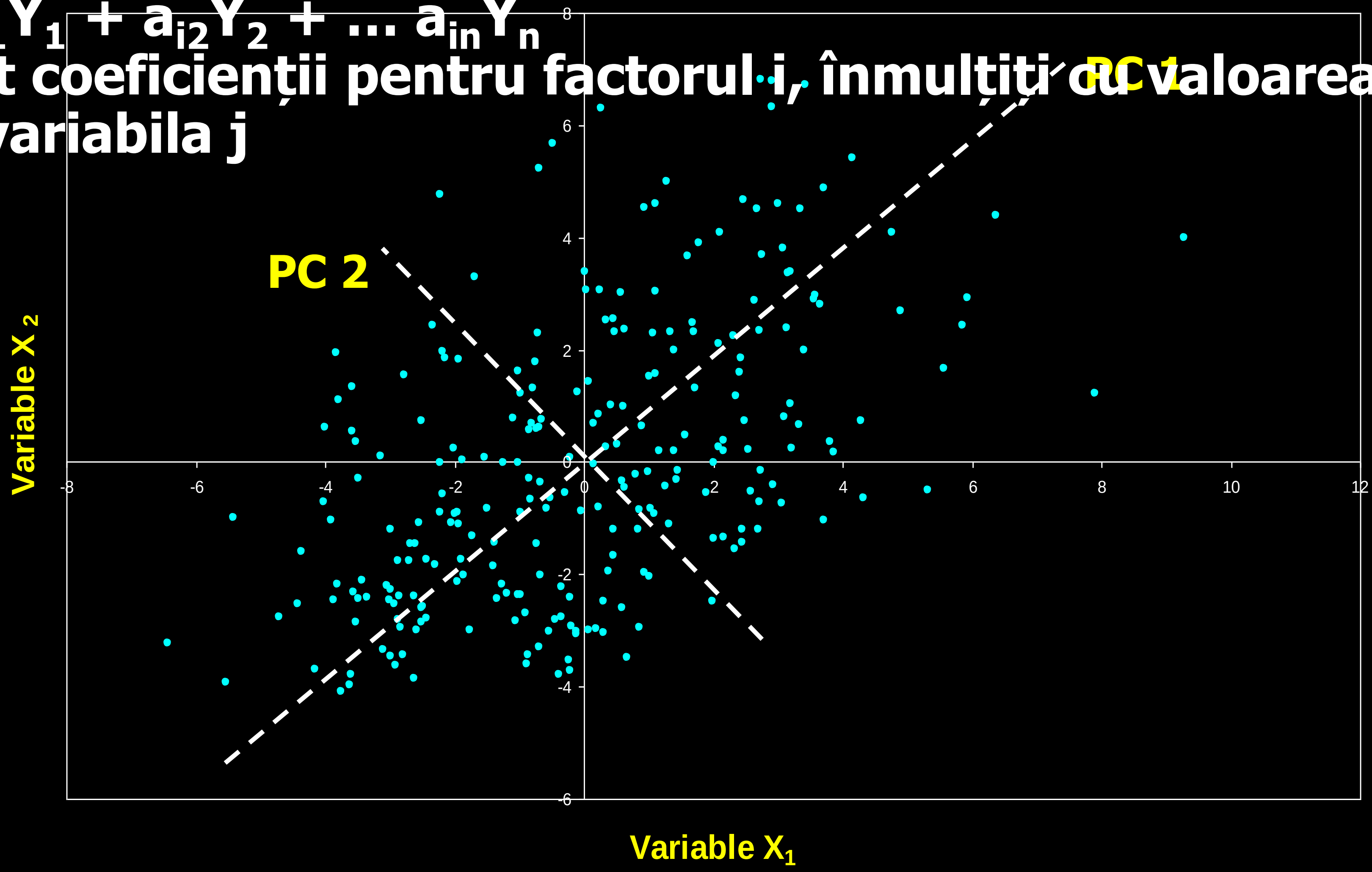
Generalizare la p dimensiuni

- În practică nimeni nu folosește PCA cu doar 2 variabile
- Algoritmul pentru găsirea axelor principale se generalizează la p variabile
- PC 1 este direcția de varianță maximă în norul p-dimensional de puncte
- PC 2 se orientează spre următoarea varianță cea mai mare, cu constrângerea de a avea covarianță zero cu PC 1.

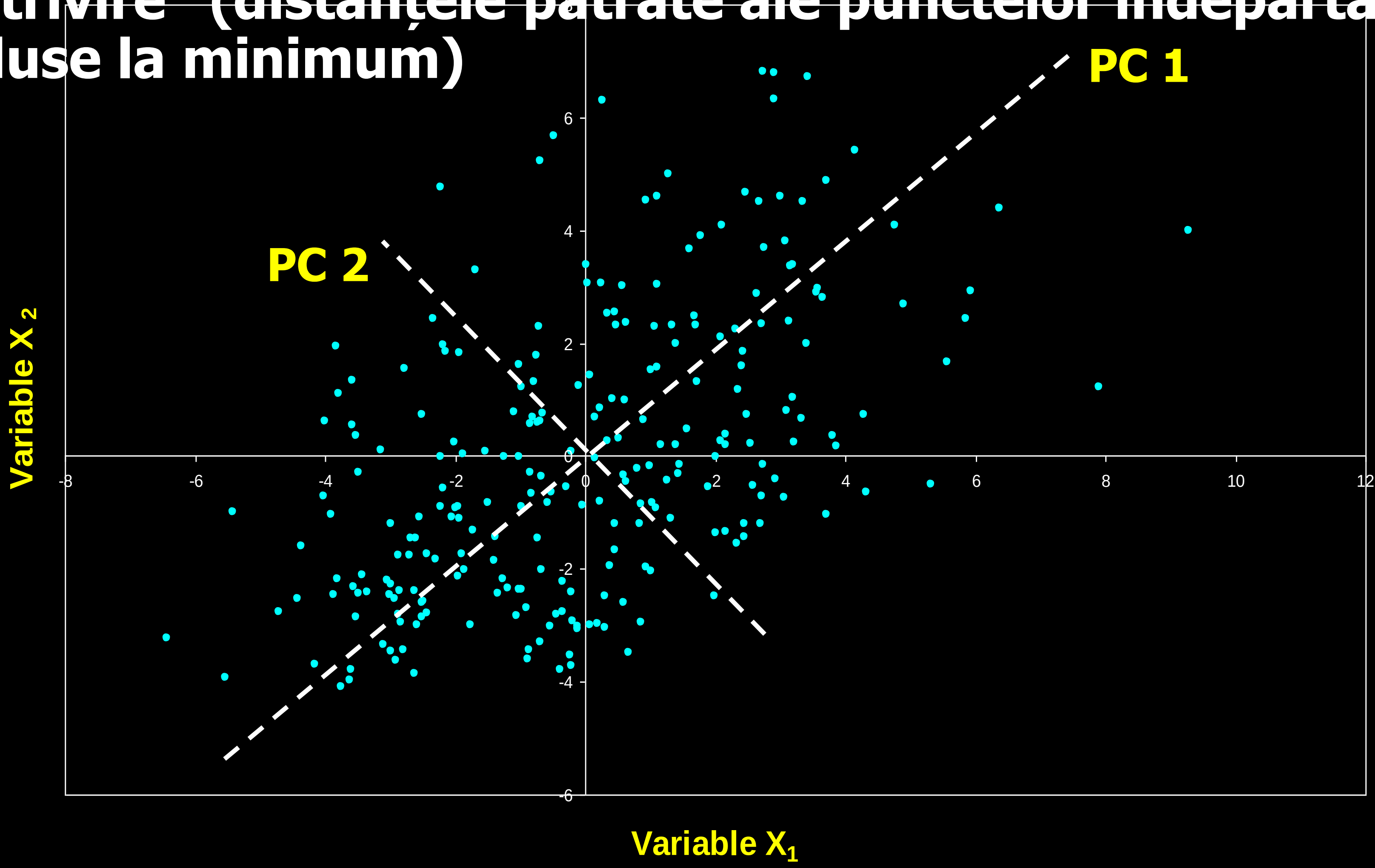
Generalizare la p dimensiuni

- PC 3 se orientează spre următoarea variantă cea mai mare, cu constrângerea că are covarianță zero atât cu PC 1 cât și cu PC 2
- și așa mai departe … până la componenta principala p

- **fiecare axă principală este o combinație liniară a celor două variabile originale**
- **$PC_j = a_{i1}Y_1 + a_{i2}Y_2 + \dots a_{in}Y_n$**
- **a_{ij}'s sunt coeficienții pentru factorul i, înmulțiți cu valoarea măsurată pentru variabila j**



- Axele PC sunt o rotație rigidă a variabilelor originale
- PC 1 este simultan direcția de varianță maximă și o „linie de cea mai bună potrivire” (distanțele pătrate ale punctelor îndepărtate de PC 1 sunt reduse la minimum)



Generalizare la p dimensiuni

- dacă luăm primele componente principale k, ele definesc „hiperplanul de cea mai bună potrivire” k-ul la nor
- din variația totală a celor p variabile:
  - PC-urile de la 1 la k reprezintă proporția maximă posibilă a acelei variații care poate fi afișată în dimensiuni k

Covarianță vs Corelație

- utilizarea covarianțelor dintre variabile are sens numai dacă sunt măsurate în aceleași unități
- chiar și atunci, variabilele cu variații mari vor domina componentele principale
- aceste probleme sunt în general evitate prin standardizarea fiecărei variabile la variația unitară și media zero

$$X'_{im} = \frac{\left(X_{im} - \overline{X}_i\right)}{SD_i}$$

Media variabilei *i*

Deviatia standard a a variabilei *i*

Covarianță vs Corelație

- covarianțele dintre variabilele standardizate sunt corelații
- după standardizare, fiecare variabilă are o variație de 1.000
- corelațiile pot fi, de asemenea, calculate din variații și covarianțe:

Corelatie intre variabilele *i* and *j*

$$r_{ij} = \frac{C_{ij}}{\sqrt{V_i V_j}}$$

Covarianta intre variabilele *i* and *j*

Varianta variabilei *i*

Varianta variabilei *j*

Algebra PCA

- prima etapă este calcularea matricei de variații și covarianțe (sau corelații) între fiecare pereche de variabile p
- matrice pătratică, simetrică
- diagonalele sunt variantele, in rest covariantele.

	X_1	X_2
X_1	6.6707	3.4170
X_2	3.4170	6.2384

Variance-covariance Matrix

	X_1	X_2
X_1	1.0000	0.5297
X_2	0.5297	1.0000

Correlation Matrix

Algebra PCA

- în notație matricială

$$\mathbf{S} = \mathbf{X}'\mathbf{X}$$

- unde X este matricea de date n x p, cu fiecare variabilă centrată (de asemenea, standardizată dacă se utilizează corelații).

	X_1	X_2
X_1	6.6707	3.4170
X_2	3.4170	6.2384

Variance-covariance Matrix

	X_1	X_2
X_1	1.0000	0.5297
X_2	0.5297	1.0000

Correlation Matrix

Manipulari Matriciale

- transpusa: schimba coloanele în rânduri sau rândurile în coloane

$$X = \begin{pmatrix} 10 & 0 & 4 \\ 7 & 1 & 2 \end{pmatrix} \qquad X' = \begin{pmatrix} 10 & 7 \\ 0 & 1 \\ 4 & 2 \end{pmatrix}$$

- înmulțirea matricilor
  - trebuie să fie același număr de coloane în matricea din stânga ca și numărul de rânduri din matricea din dreapta

Algebra PCA

- suma diagonalelor matricei varianță-covarianță se numește urmă (trace)
- reprezintă variația totală a datelor
- este distanța euclidiană pătrată medie între fiecare obiect și centroid în spațiul p-dimensional.

	X_1	X_2
X_1	6.6707	3.4170
X_2	3.4170	6.2384

Trace = 12.9091

	X_1	X_2
X_1	1.0000	0.5297
X_2	0.5297	1.0000

Trace = 2.0000

Algebra PCA

- **găsirea principalelor axe implică analiza valorilor proprii ale matricei S)**
- **valorile proprii (rădăcini latente) ale lui S sunt soluții (λ) la ecuația caracteristică:**

$$\left|\mathbf{S} - \lambda \mathbf{I}\right| = \mathbf{0}$$

Algebra PCA

- **Valorile proprii, λ_1, λ_2, ... λ_p sunt variațiile coordonatelor pe fiecare axă principală a componentelor**
- **suma tuturor valorilor proprii este egală cu urma lui S (suma variațiilor variabilelor inițiale).**

	X_1	X_2
X_1	6.6707	3.4170
X_2	3.4170	6.2384

Trace = 12.9091

$\lambda_1 = 9.8783$

$\lambda_2 = 3.0308$

Obs: $\lambda_1 + \lambda_2 = 12.9091$

Algebra PCA

- **Fiecare vector propriu (eigenvector este format din p valori care reprezintă „contribuția” fiecărei variabile la axa componentă principală**
- **Vectorii proprii nu sunt corelați (ortogonali)**
  - **produsul lor scalar e zero.**

Eigenvectors

	u_1	u_2
x_1	0.7291	-0.6844
x_2	0.6844	0.7291

$$0.7291*(-0.6844) + 0.6844*0.7291 = 0$$

Algebra PCA

- coordonatele fiecărui obiect i pe axa principală k, cunoscute sub numele de scoruri pt. PC k, sunt calculate ca

$$z_{ki} = u_{1k}x_{1i} + u_{2k}x_{2i} + \cdots + u_{pk}x_{pi}$$

- where Z is the *n* x *k* matrix of PC scores, X is the *n* x *p* centered data matrix and U is the *p* x *k* matrix of eigenvectors

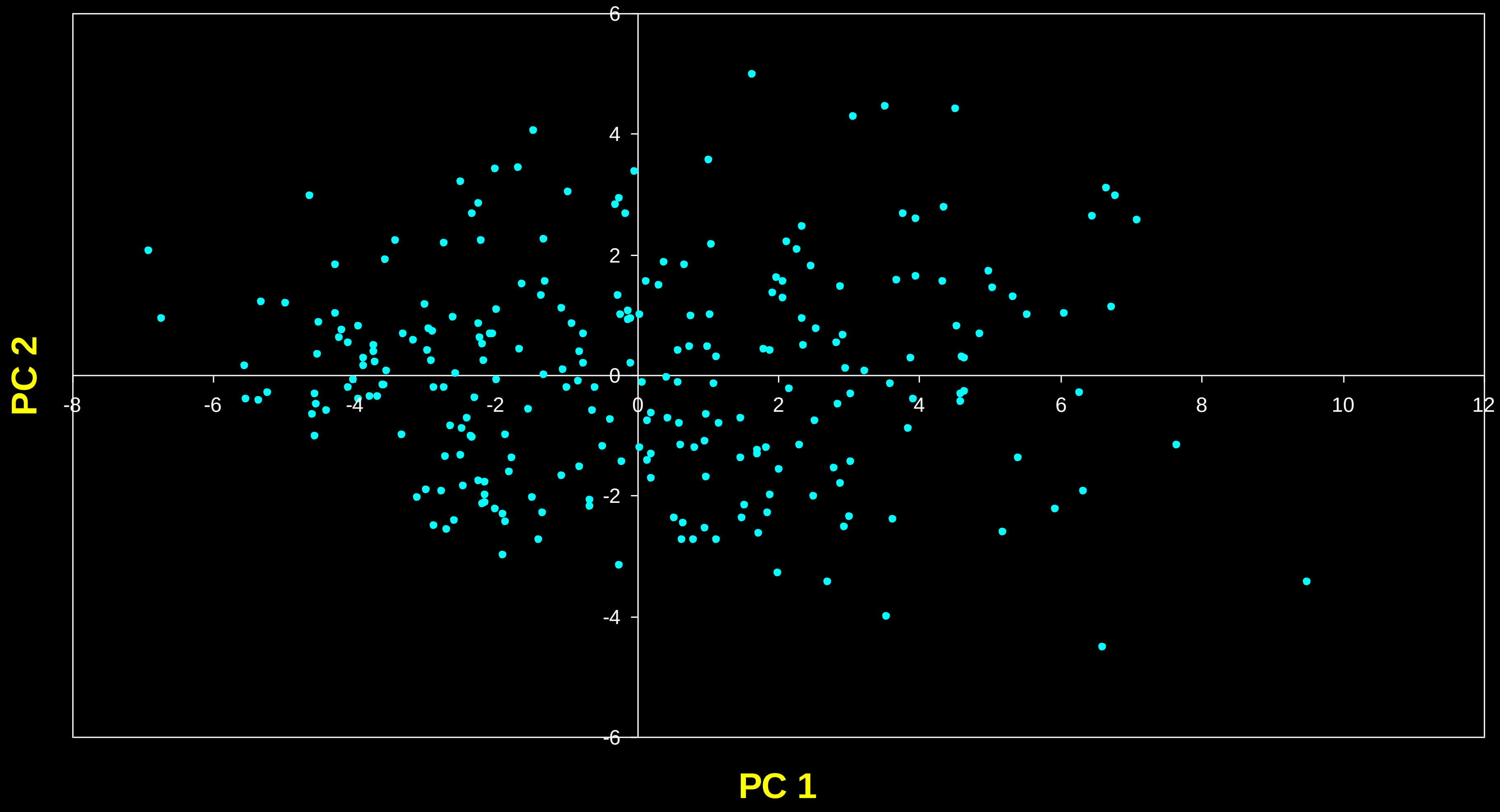
- unde Z este matricea n x k a scorurilor PC, X este matricea de date centrată n x p și U este matricea p x k a eigenvectorilor.

Algebra PCA

- variația scorurilor pe fiecare componentă este egală cu valoarea corespunzătoare a valorii proprii pentru acea axă
- valoarea proprie reprezintă variația afișată („explicată” sau „extrasă”) de axa k
- suma primelor k valori proprii este variația explicată prin ordonarea k-dimensională.

$\lambda_1 = 9.8783$ $\lambda_2 = 3.0308$ Trace = 12.9091

PC 1 "explică"
9.8783/12.9091 = 76.5% din variația totală



The Algebra of PCA

- Matricea de produse scalare calculată între axe principale are o formă simplă:
- toate valorile din afara diagonalei sunt zero (axele principale sunt necorelate)
- valorile diagonale sunt valorile proprii.

	PC_1	PC_2
PC_1	9.8783	0.0000
PC_2	0.0000	3.0308

Variance-covariance Matrix of the PC axes

Un exemplu mai provocator

- date din cercetările privind definirea habitatului în broasca Baw Baw pe cale de dispariție
- 16 variabile de mediu și structurale măsurate la fiecare din 124 de situri
- matricea de corelație folosită deoarece variabilele au unități diferite

Valori proprii

Axis	Eigenvalue	% de Variatie explicata	% cumulat de Variatie explicata
1	5.855	36.60	36.60
2	3.420	21.38	57.97
3	1.122	7.01	64.98
4	1.116	6.97	71.95
5	0.982	6.14	78.09
6	0.725	4.53	82.62
7	0.563	3.52	86.14
8	0.529	3.31	89.45
9	0.476	2.98	92.42
10	0.375	2.35	94.77

Interpretarea vectorilor proprii

- **Corelațiile dintre variabile și axele principale sunt cunoscute sub denumirea de sarcini**
- **fiecare element al vectorilor proprii reprezintă contribuția unei variabile date la o componentă**

	1	2	3
Altitude	0.3842	0.0659	-0.1177
pH	-0.1159	0.1696	-0.5578
Cond	-0.2729	-0.1200	0.3636
TempSurf	0.0538	-0.2800	0.2621
Relief	-0.0765	0.3855	-0.1462
maxERht	0.0248	0.4879	0.2426
avERht	0.0599	0.4568	0.2497
%ER	0.0789	0.4223	0.2278
%VEG	0.3305	-0.2087	-0.0276
%LIT	-0.3053	0.1226	0.1145
%LOG	-0.3144	0.0402	-0.1067
%W	-0.0886	-0.0654	-0.1171
H1Moss	0.1364	-0.1262	0.4761
DistSWH	-0.3787	0.0101	0.0042
DistSW	-0.3494	-0.1283	0.1166
DistMF	0.3899	0.0586	-0.0175

Câte axe sunt necesare?

- axa principală (k + 1) reprezintă mai multă variație decât ne-am aștepta?
- au fost propuse mai multe teste și reguli
- o „regulă generală” comună când PCA se bazează pe corelații este că axe cu valori proprii> 1 merită interpretate

Care sunt ipotezele PCA?

- relațiile dintre variabile sunt LINEARE
- nor de puncte în spațiul p-dimensional are dimensiuni liniare care pot fi rezumate eficient de axele principale
- dacă structura din date este NELINEARĂ (norul de puncte se curbează prin spațiul p-dimensional), axele principale nu sunt un rezumat eficient și informativ al datelor.