MAP 553
Apprentissage statistique

Christophe Giraud

Université Paris Sud et Ecole Polytechnique

http://www.cmap.polytechnique.fr/~giraud/MAP553/MAP553.html

PC1

Apprentissage?

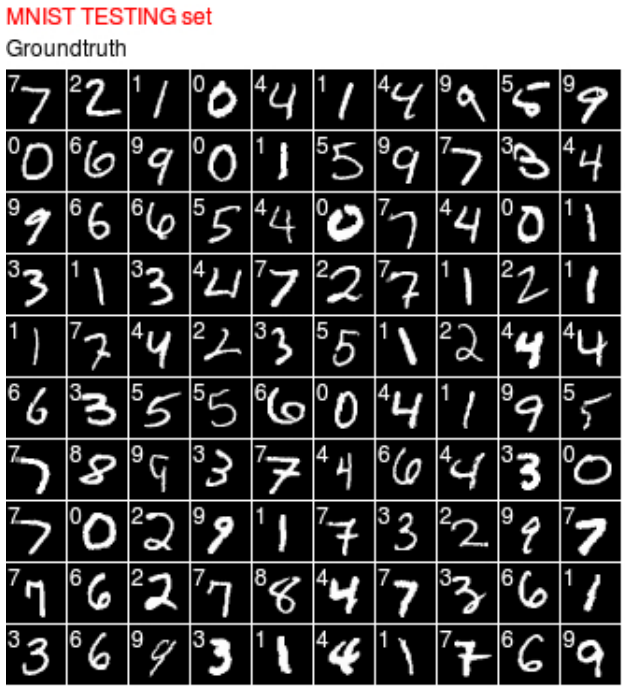
L'apprentissage au "quotidien"

1. **filtres SPAM**
2. **Reconnaissance de chiffre**: lecture automatique de codes postaux
3. **Diagnostique médical**: de cancers, alzheimer, diabète, etc
4. **In silico chemometrics**: recherche "virtuelle" de médicaments
5. **Business analytics, Google ranking, web-data, etc**

`http://c-command.com/spamsieve/`

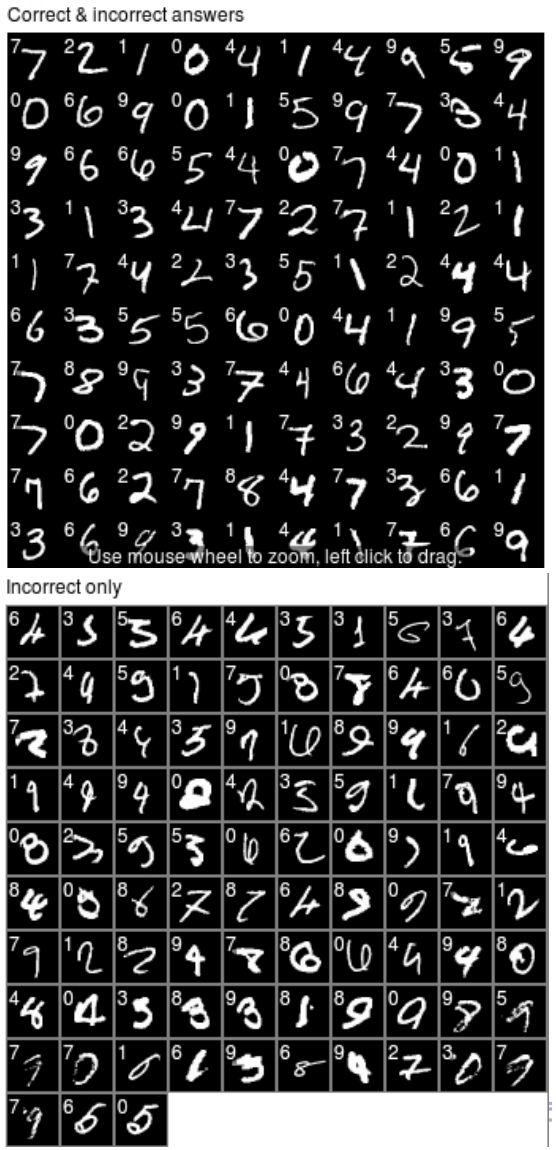
L'apprentissage au "quotidien"

1. filtres SPAM
2. **Reconnaissance de chiffre**: lecture automatique de codes postaux
3. **Diagnostique médical**: de cancers, alzheimer, diabète, etc
4. **In silico chemometrics**: recherche "virtuelle" de médicaments
5. **Business analytics, Google ranking, web-data, etc**



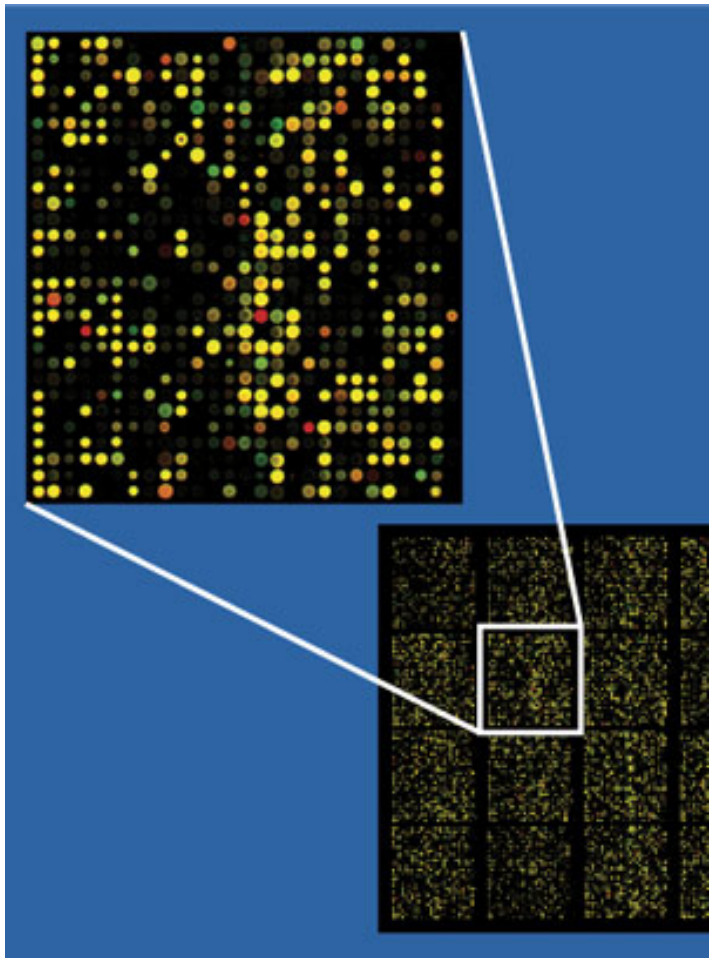
L'apprentissage au "quotidien"

1. filtres SPAM
2. **Reconnaissance de chiffre**: lecture automatique de codes postaux
3. **Diagnostique médical**: de cancers, alzheimer, diabète, etc
4. **In silico chemometrics**: recherche "virtuelle" de médicaments
5. **Business analytics, Google ranking, web-data, etc**



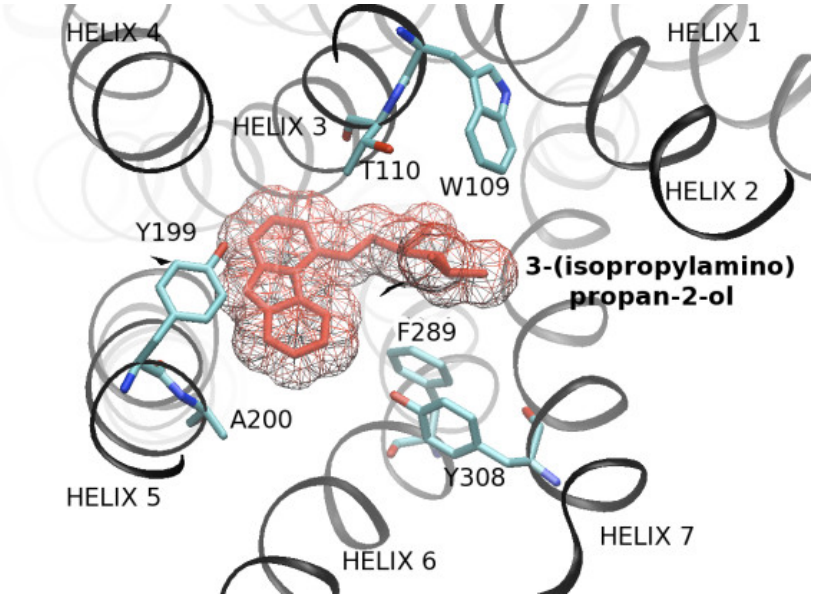
L'apprentissage au "quotidien"

1. **filtres SPAM**
2. **Reconnaissance de chiffre**: lecture automatique de codes postaux
3. **Diagnostique médical**: de cancers, alzheimer, diabète, etc
4. **In silico chemometrics**: recherche "virtuelle" de médicaments
5. **Business analytics, Google ranking, web-data, etc**



L'apprentissage au "quotidien"

1. filtres SPAM
2. **Reconnaissance de chiffre**: lecture automatique de codes postaux
3. **Diagnostique médical**: de cancers, alzheimer, diabète, etc
4. **In silico chemometrics**: recherche "virtuelle" de médicaments
5. **Business analytics, Google ranking, web-data, etc**



L'apprentissage au "quotidien"

1. **filtres SPAM**
2. **Reconnaissance de chiffre**: lecture automatique de codes postaux
3. **Diagnostique médical**: de cancers, alzheimer, diabète, etc
4. **In silico chemometrics**: recherche "virtuelle" de médicaments
5. **Business analytics, Google ranking, web-data, etc**

Apprentissage?

Les deux aspects de l'apprentissage:

$\longrightarrow$ **aspect statistique**

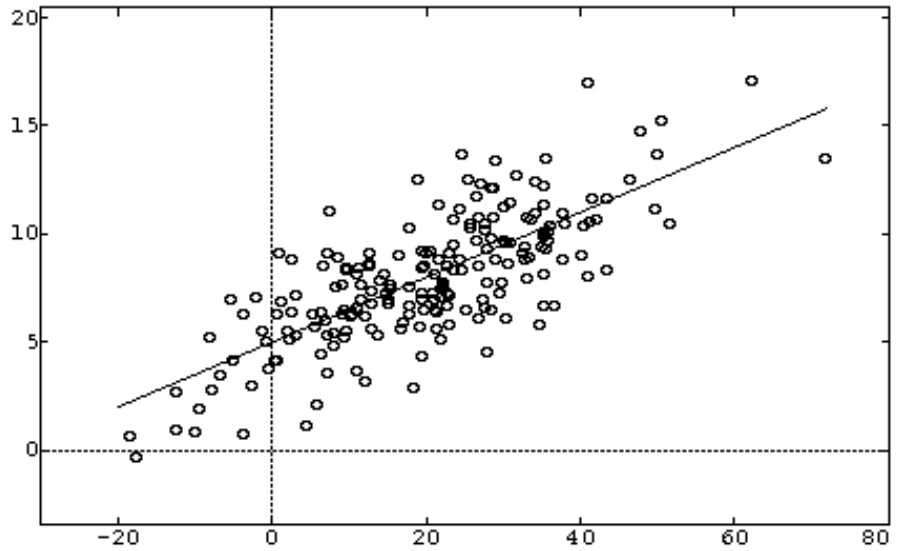
$\longrightarrow$ **aspect algorithmique**

Le fléau de la dimension

Renversement de point de vue

Cadre statistique classique:

- petit nombre p de paramètres
- grand nombre n d'expériences
- on étudie le comportement asymptotique des estimateurs lorsque $n \to \infty$ (résultats type théorème central limite)



Renversement de point de vue

Cadre statistique classique:

- petit nombre p de paramètres
- grand nombre n d'expériences
- on étudie le comportement asymptotique des estimateurs lorsque $n \to \infty$ (résultats type théorème central limite)

Données actuelles:

- inflation du nombre p de paramètres
- taille d'échantillon reste réduite: $n \approx p$ ou $n \ll p$

$\Longrightarrow$ penser différemment les statistiques!
(penser $n \to \infty$ ne convient plus)

Le fléau de la dimension: exemple 1

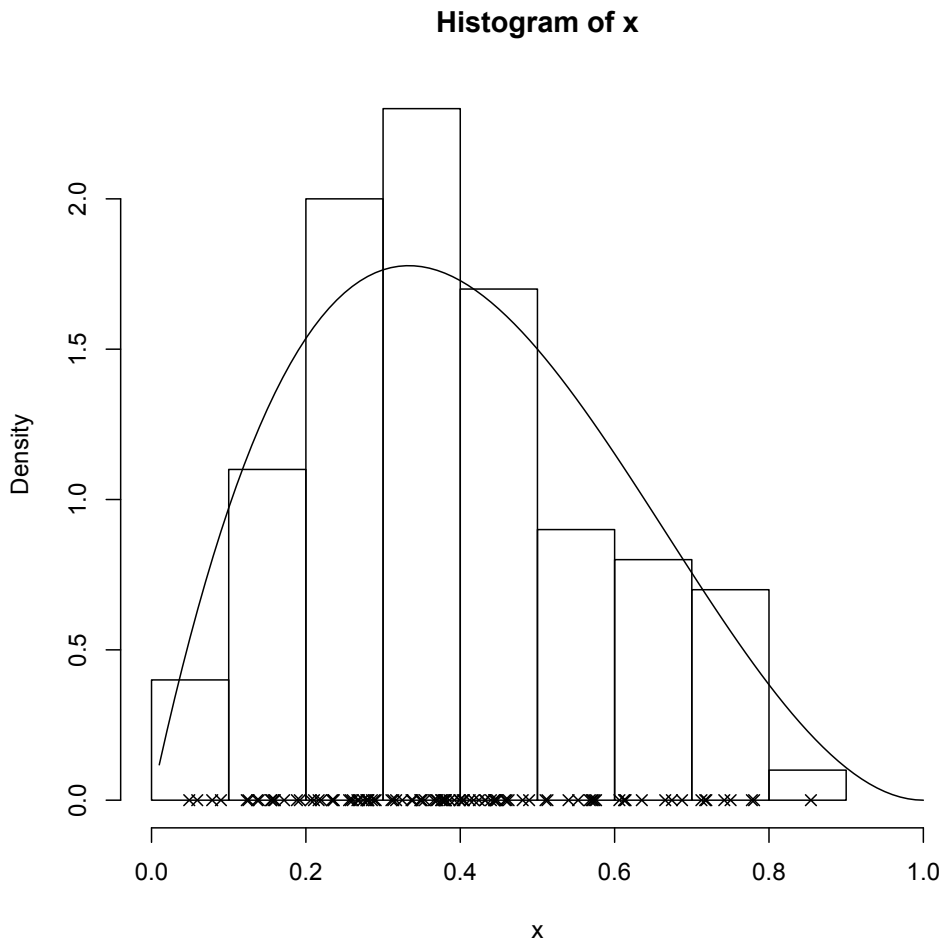
On observe $X_1, \ldots, X_n \in [0,1]^p$ i.i.d. selon une densité

$$f : [0,1]^p \to \mathbb{R} \quad \textbf{inconnue.}$$

On cherche à estimer f. Une idée naturelle est de faire un histogramme avec disons des "cases" de 0.1 de côté.

Le fléau de la dimension: exemple 1

En dimension $p = 1$:



Histogramme d'un échantillon de $n = 100$ tirages d'une loi beta.

Le fléau de la dimension: exemple 1

On observe $X_1, \ldots, X_n \in [0,1]^p$ i.i.d. selon une densité

$$f : [0,1]^p \to \mathbb{R} \quad \textbf{inconnue.}$$

On cherche à estimer f. Une idée naturelle est de faire un histogramme avec disons des "cases" de 0.1 de côté.

Questions :

1. Pour avoir en moyenne 10 observations par cases, quelle taille doit avoir n (en fonction de p)?
2. Conclusion? Comment faire avec des échantillons plus petits?

Le fléau de la dimension: exemple 2

Echec des méthodes locales en régression.
On observe $Y_1, \dots, Y_n \in \mathbb{R}$ et $X_1, \dots, X_n \in \mathbb{R}^p$ avec

$$Y_i = f(X_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

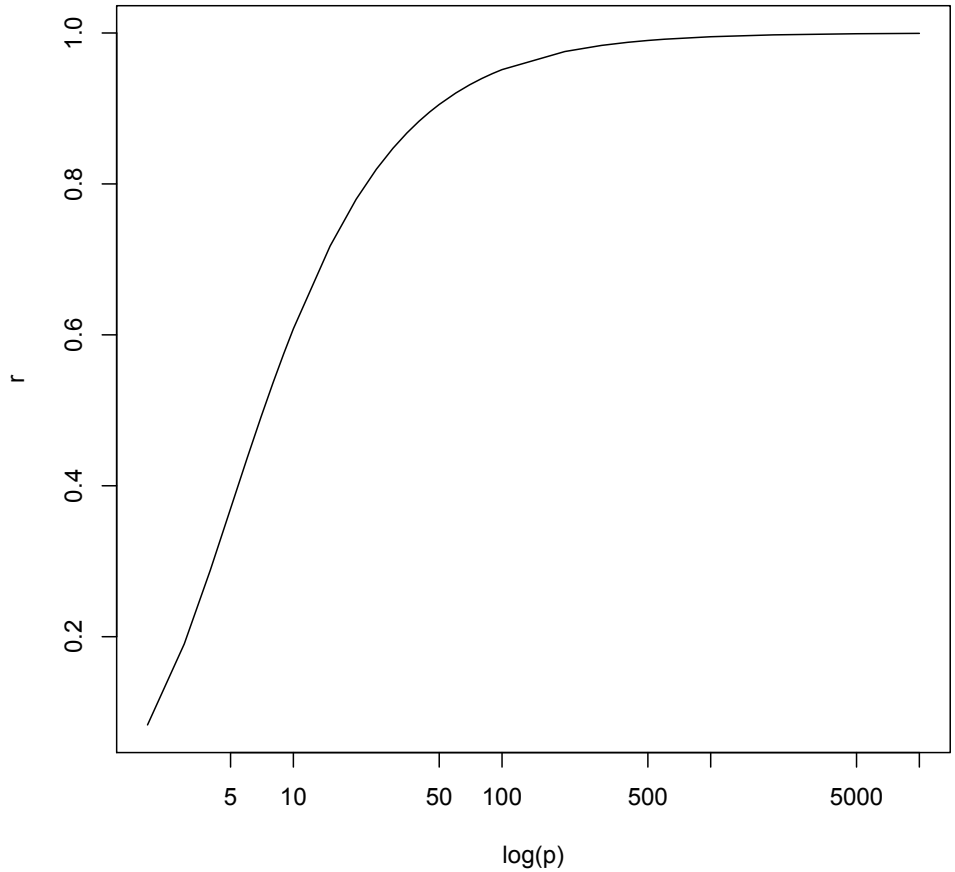
où f est **inconnue** et les $\mathbb{E}[\varepsilon_i] = 0$.

$\hat{f}(x) = \text{Moyenne}\{Y_i : X_i \in \mathcal{B}(x, r)\}$ avec un r petit.

on supposera les $X_i \overset{i.i.d.}{\sim} \mathcal{U}(B(0,1))$

1. Pour $r < 1$ montrer $\mathbb{P}(\exists\, X_i \in \mathcal{B}(0, r)) = 1 - (1 - r^p)^n$.
2. Pour quelle valeur de r est-ce supérieur à $1/2$?
3. Pour estimer $f(0)$ avec au moins un point, quel est l'ordre de grandeur du diamètre r minimal? Conclusion?

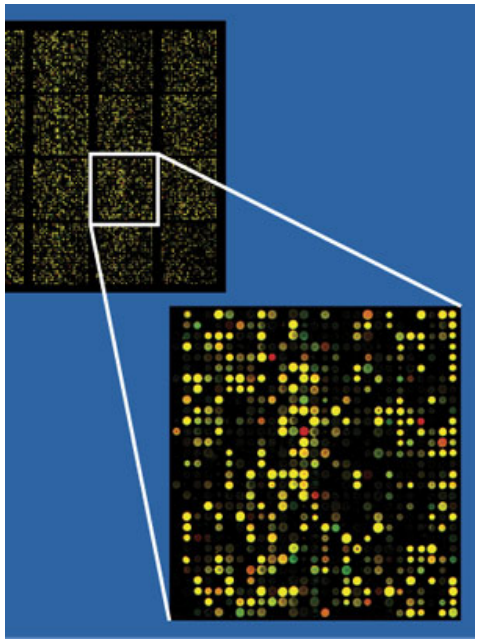
Le fléau de la dimension: exemple 2



valeurs de r pour lesquelles $(1 - r^p)^n = 1/2$, cas $n = 100$.

Le fléau de la dimension: exemple 3

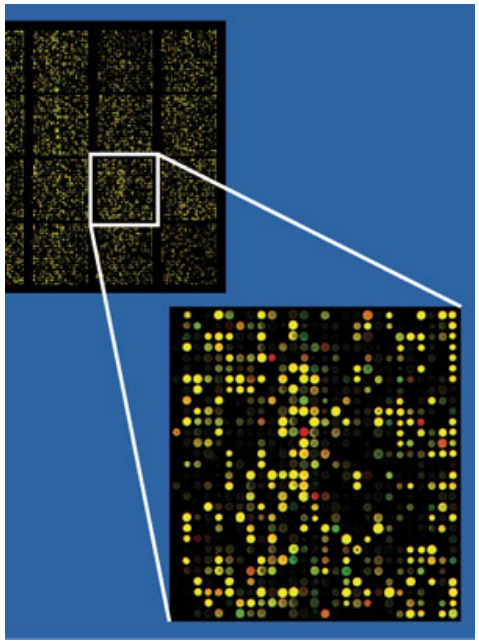
Puces ADN:



- **Modèle:** log-intensité du spot (après normalisation) $X_i = \theta_i + \epsilon_i$ avec $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$
- **Déviation gaussienne à 5%:** on a $\mathbb{P}\left[(\mathcal{N}(0,1))^2 > 3.84\right] \approx 5\%$

Le fléau de la dimension: exemple 3

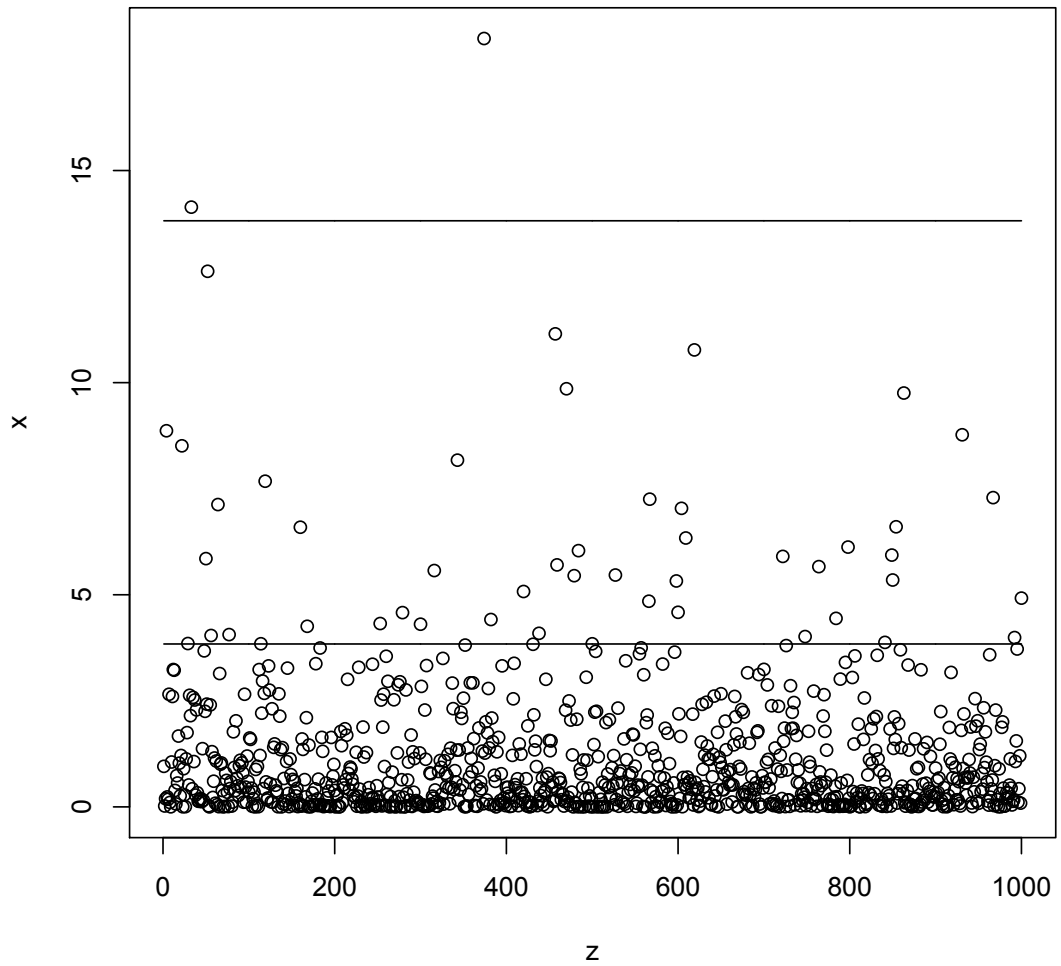
Puces ADN:



- **Modèle:** log-intensité du spot (après normalisation) $X_i = \theta_i + \epsilon_i$ avec $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$
- **Déviation gaussienne à 5%:** on a $\mathbb{P}\left[(\mathcal{N}(0,1))^2 > 3.84\right] \approx 5\%$

Les valeurs X_i^2 supérieures à 3.84 sont-elles significatives?

Le fléau de la dimension: exemple 3



Avec $p = 1000$ et $\theta_i = 0 \ \forall i$ (donc $X_i^2 = \varepsilon_i^2$).
Niveaux représentés: 3.84 et $2 \log p$.

Le fléau de la dimension: exemple 3

Combien de faux positifs ?

Supposons que $p = 5000$ et 4% des gènes sont positifs. Quel est le nombre moyen de faux positifs si on conserve tous les $X_i^2 > 3.84$?

Pourquoi un seuil à $2\log(p)$?

$$\mathbb{P}\left(\max_{i=1,\dots,p} \varepsilon_i^2 > t_p\right) \overset{t_p \sim \alpha \log p}{\longrightarrow} \begin{cases} 0 & \text{si } \alpha \geq 2 \\ 1 & \text{si } \alpha < 2 \end{cases}$$

Quel est le problème si p grand ?

Le fléau de la dimension: pour aller plus loin

- *Introduction to High-Dimensional Statistics.* To appear.

  `http://www.cmap.polytechnique.fr/~giraud/MSV/LectureNotes.pdf`

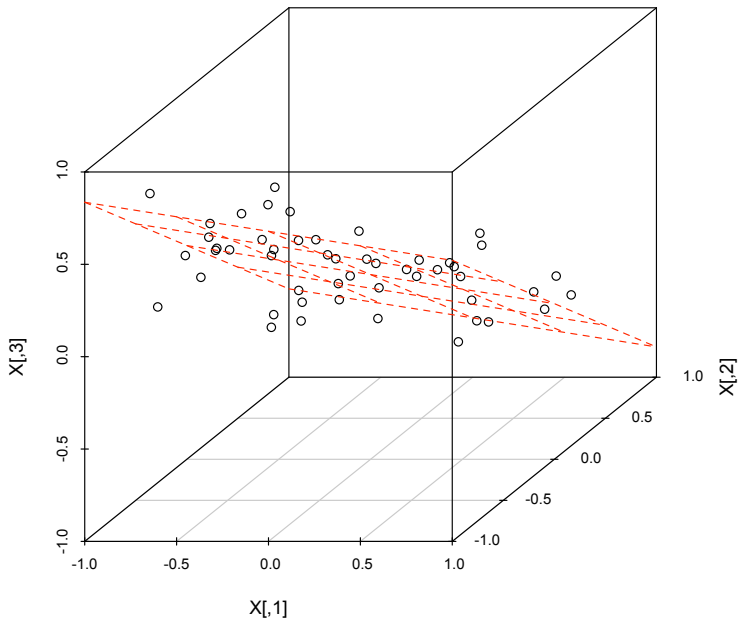
- Jiashun Jin. *Impossibility of successful classification when useful features are rare and weak*. Proceedings of the National Academy of Sciences of the USA. 106 (22); 2009. pp.8859-64.

  `http://www.pnas.org/content/106/22/8859.full`

Réduction de dimension : ACP

Réduire la dimension

Objectif: trouver un espace V de petite dimension tel que (simultanément) les observations $X_i \in \mathbb{R}^p$ soient proches de leur projection sur cet espace



Ex: dimension $p = 3$: meilleur plan approximant.

Un exemple visuel : MNIST

Base MNIST : 1100 chiffres scannés

Figure : chaque image 16×16 correspond à un vecteur dans $\mathbb{R}^{256}$

Un exemple visuel : MNIST

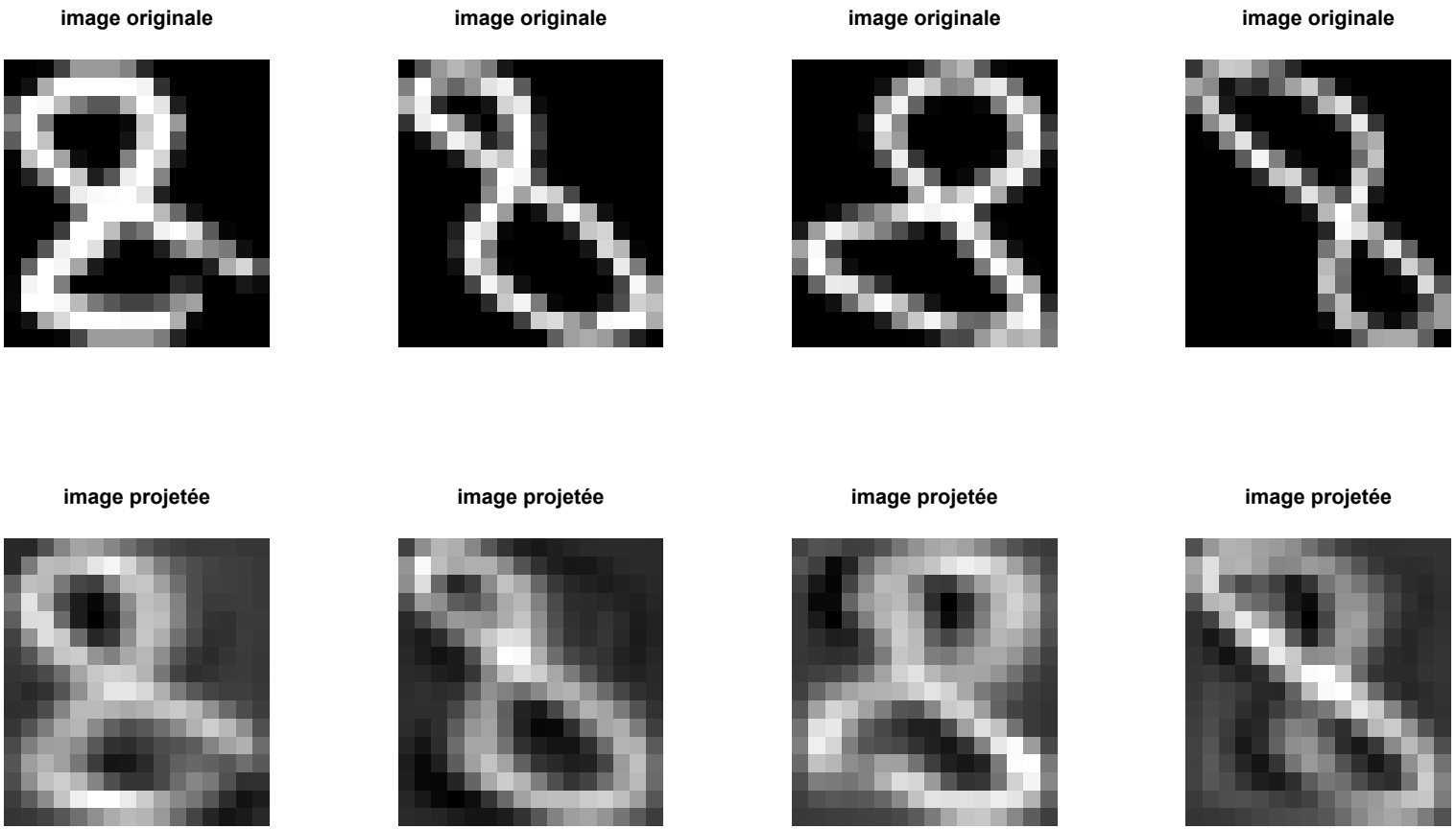


Figure : Projection des images sur un espace affine de dimension 10 donné par l'ACP

Un exemple visuel : MNIST

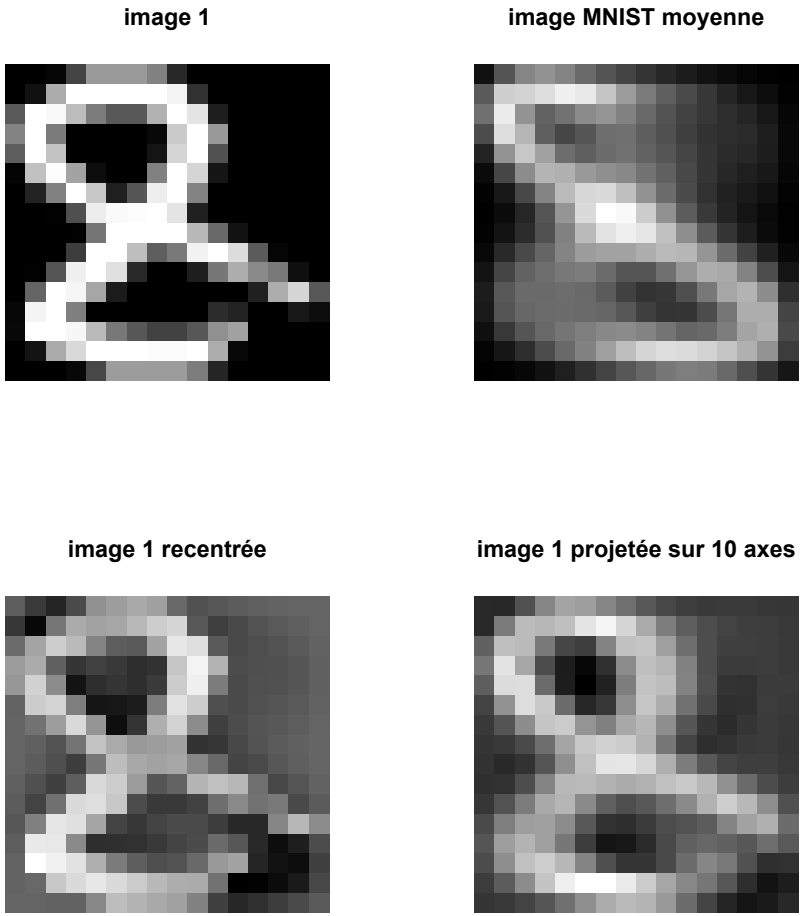


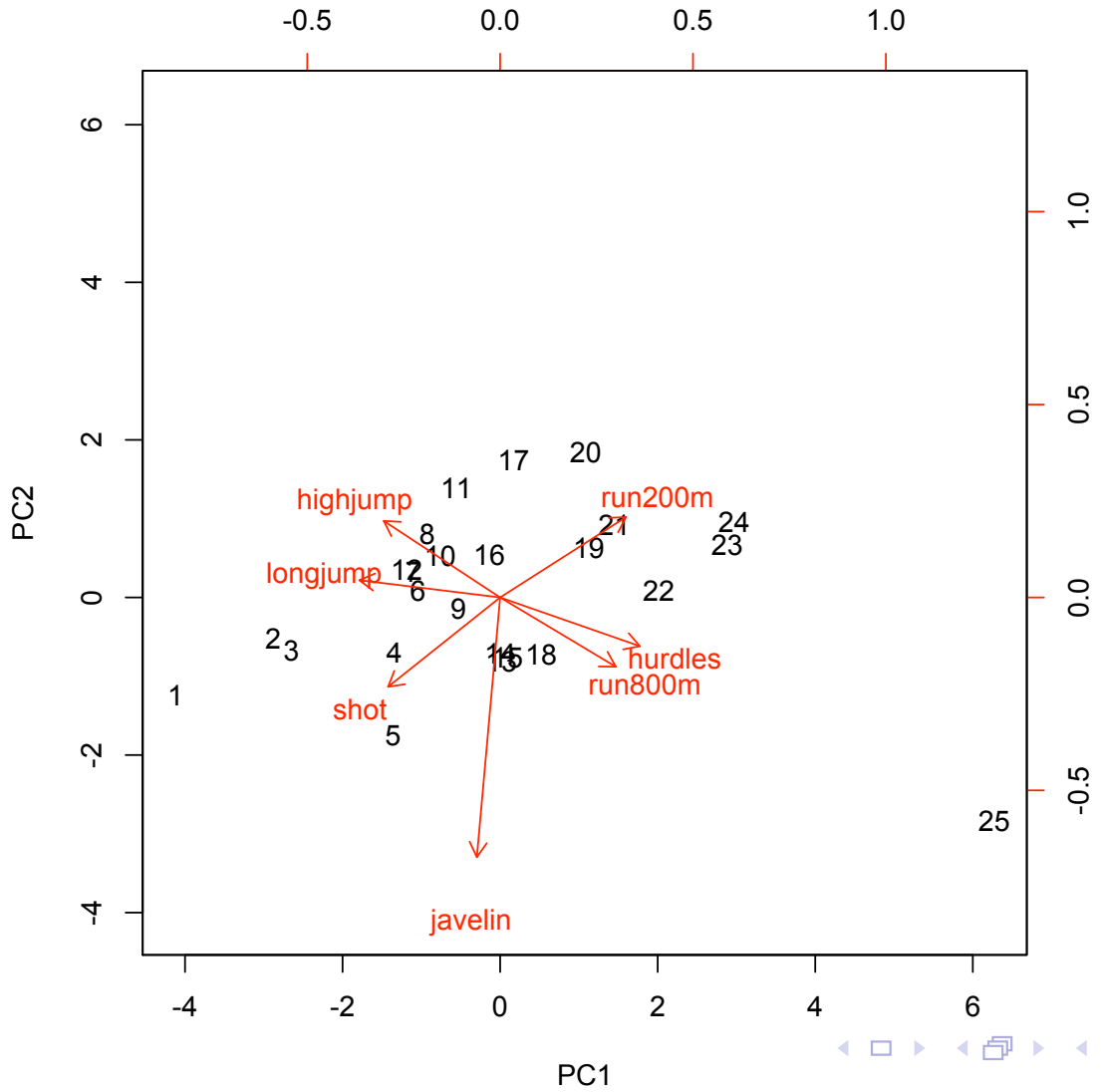
Figure : Réduction de dimension d'un facteur 25 par ACP

Epreuve d'heptathlon, jeux olympiques de Seoul, 1988.

	hurdles	highjump	shot	run200m	longjump	javelin	run800m
Joyner-Kersee (USA)	12.69	1.86	15.80	22.56	7.27	45.66	128.51
John (GDR)	12.85	1.80	16.23	23.65	6.71	42.56	126.12
Behmer (GDR)	13.20	1.83	14.20	23.10	6.68	44.54	124.20
Sablovskaite (URS)	13.61	1.80	15.23	23.92	6.25	42.78	132.24
Choubenkova (URS)	13.51	1.74	14.76	23.93	6.32	47.46	127.90
Schulz (GDR)	13.75	1.83	13.50	24.65	6.33	42.82	125.79
Fleming (AUS)	13.38	1.80	12.88	23.59	6.37	40.28	132.54
Greiner (USA)	13.55	1.80	14.13	24.48	6.47	38.00	133.65
Lajbnerova (CZE)	13.63	1.83	14.28	24.86	6.11	42.20	136.05
Bouraga (URS)	13.25	1.77	12.62	23.59	6.28	39.06	134.74
Wijnsma (HOL)	13.75	1.86	13.01	25.03	6.34	37.86	131.49
Dimitrova (BUL)	13.24	1.80	12.88	23.59	6.37	40.28	132.54
Scheider (SWI)	13.85	1.86	11.58	24.87	6.05	47.50	134.93
Braun (FRG)	13.71	1.83	13.16	24.78	6.12	44.58	142.82
Ruotsalainen (FIN)	13.79	1.80	12.32	24.61	6.08	45.44	137.06
Yuping (CHN)	13.93	1.86	14.21	25.00	6.40	38.60	146.67
Hagger (GB)	13.47	1.80	12.75	25.47	6.34	35.76	138.48
Brown (USA)	14.07	1.83	12.69	24.83	6.13	44.34	146.43
Mulliner (GB)	14.39	1.71	12.68	24.92	6.10	37.76	138.02
Hautenauve (BEL)	14.04	1.77	11.81	25.61	5.99	35.68	133.90
Kytola (FIN)	14.31	1.77	11.66	25.69	5.75	39.48	133.35
Geremias (BRA)	14.23	1.71	12.95	25.50	5.50	39.64	144.02
Hui-Ing (TAI)	14.85	1.68	10.00	25.23	5.47	39.14	137.30
Jeong-Mi (KOR)	14.53	1.71	10.83	26.61	5.50	39.26	139.17
Launa (PNG)	16.42	1.50	11.78	26.16	4.88	46.38	163.43

ACP en action

Résultat d'une ACP sur les données d'heptathlon.



ACP : notations

Epreuve d'heptathlon, jeux olympiques de Seoul, 1988.

tableau $n \times p$: $\mathbf{X} = \left[X_i^{(v)}\right]_{\substack{i=1\dots n \\ v=1\dots p}} = \left[X^{(1)}, \dots, X^{(p)}\right] = \begin{bmatrix} X_1^T \\ \vdots \\ X_n^T \end{bmatrix}$

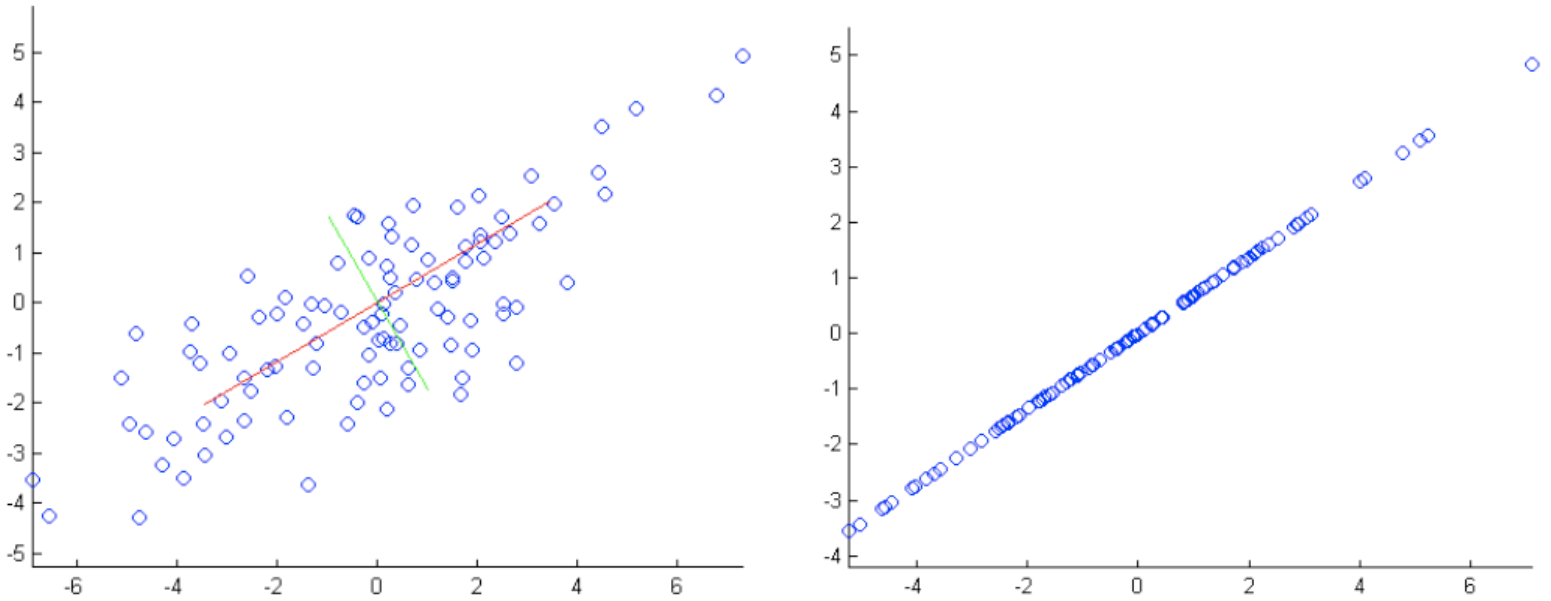
$p = 7$ **variables:**

1. hurdles: results 100m hurdles.
2. highjump: results high jump.
3. shot: results shot.
4. run200m: results 200m race.
5. longjump: results long jump.
6. javelin: results javelin.
7. run800m: results 800m race.

$n = 25$ **athlètes.**

Réduire la dimension

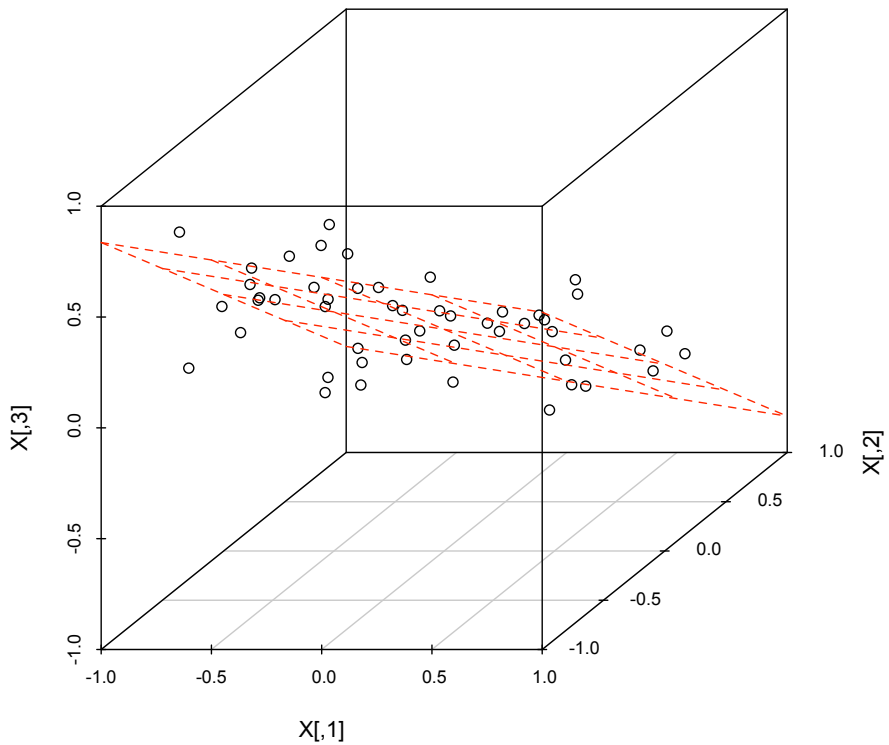
But: représenter les obervations $X_i \in \mathbb{R}^p$ dans un espace de plus petite dimension avec le moins de perte d'information possible.



Ex: avec $p = 2$ variables: axes de projections (1er en rouge, 2nd en vert).

Réduire la dimension

But: représenter les obervations $X_i \in \mathbb{R}^p$ dans un espace de plus petite dimension avec le moins de perte d'information possible.



Ex: avec $p = 3$ variables: meilleur plan approximant.

Etape préliminaire

Normalisation: étape préliminaire de normalisation des données:

- centrer: $X^{(v)} \leftarrow X^{(v)} - \bar{X}^{(v)}$
- réduire: $X^{(v)} \leftarrow X^{(v)} / \sqrt{\overline{var}\left(X^{(v)}\right)}$ (sauf si comparables)

Dorénavant on supposera les données **centrées**.

Objectif : Obtenir un espace vectoriel V de petite dimension tel que $\mathrm{Proj}_V(X_i) \approx X_i$ pour $i = 1, \ldots, n$.

Questions : on notera $\widehat{\Sigma} = \frac{1}{n}\mathbf{X}^T\mathbf{X}$

1. Montrer que

$$\begin{aligned} V_d &:= \underset{\dim(V)=d}{\operatorname{argmin}} \sum_{i=1}^{n} \|X_i - \mathrm{Proj}_V(X_i)\|^2 \\ &= \underset{\dim(V)=d}{\operatorname{argmax}} \sum_{i=1}^{n} \|\mathrm{Proj}_V(X_i)\|^2 \end{aligned}$$

2. Par quels vecteurs V_d est-il engendré ? (commencer par $d = 1$)
3. Que vaut $\sum_{i=1}^{n} \|\mathrm{Proj}_{V_d}(X_i)\|^2$?

ACP : définitions

Axes principaux: $a^{(1)} \perp \ldots \perp a^{(d)} \in \mathbb{R}^p$ vecteurs propres orthonormés de $\widehat{\Sigma} = \frac{1}{n}\mathbf{X}^T\mathbf{X}$, ordonnés selon les valeurs propres décroissantes

Composantes principales: $c_k = \mathbf{X}a^{(k)} \in \mathbb{R}^n$ pour $k = 1, \ldots, d$

Remarques:

- $c_1 \perp \ldots \perp c_d \in \mathbb{R}^n$: car $\langle c_j, c_k \rangle = n(a^{(j)})^T \widehat{\Sigma}\, a^{(k)} = n\lambda_k \delta_{jk}$
- $\|a^{(k)}\| = 1$ et $\|c_k\|^2 = n\lambda_k$
- $c_1, \ldots, c_d \in \mathbb{R}^n$ vecteurs propres de $\mathbf{X}\mathbf{X}^T$

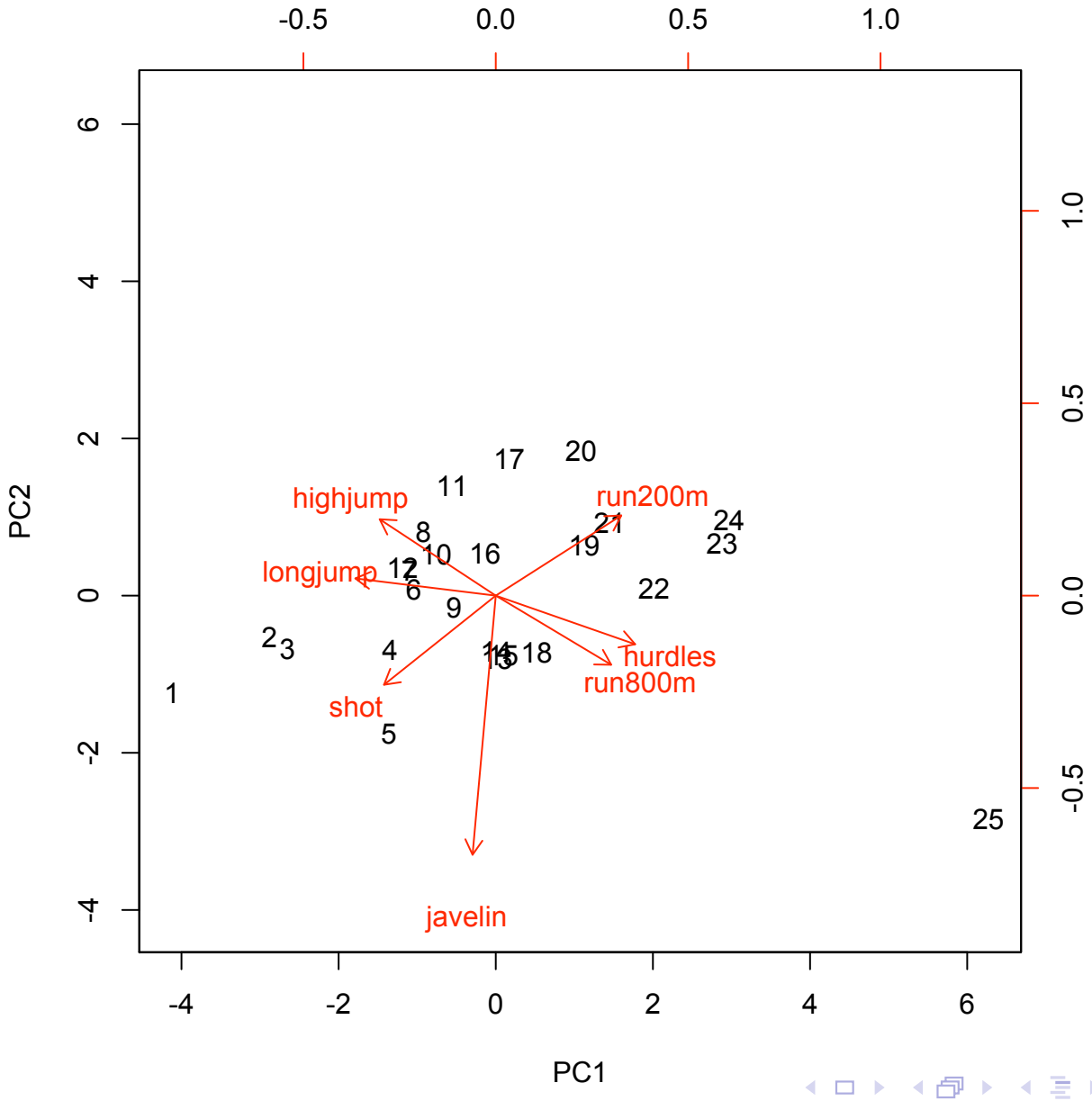
Projection des individus: $\langle X_i, a^{(k)} \rangle = (c_k)_i$ donc

$$\mathrm{Proj}_{V_d}(X_i) = \sum_{k=1}^{d} (c_k)_i \, a^{(k)}$$

Projection des variables: $\langle X^{(v)}, c_k \rangle / \|c_k\|^2 = (a^{(k)})_v$ donc

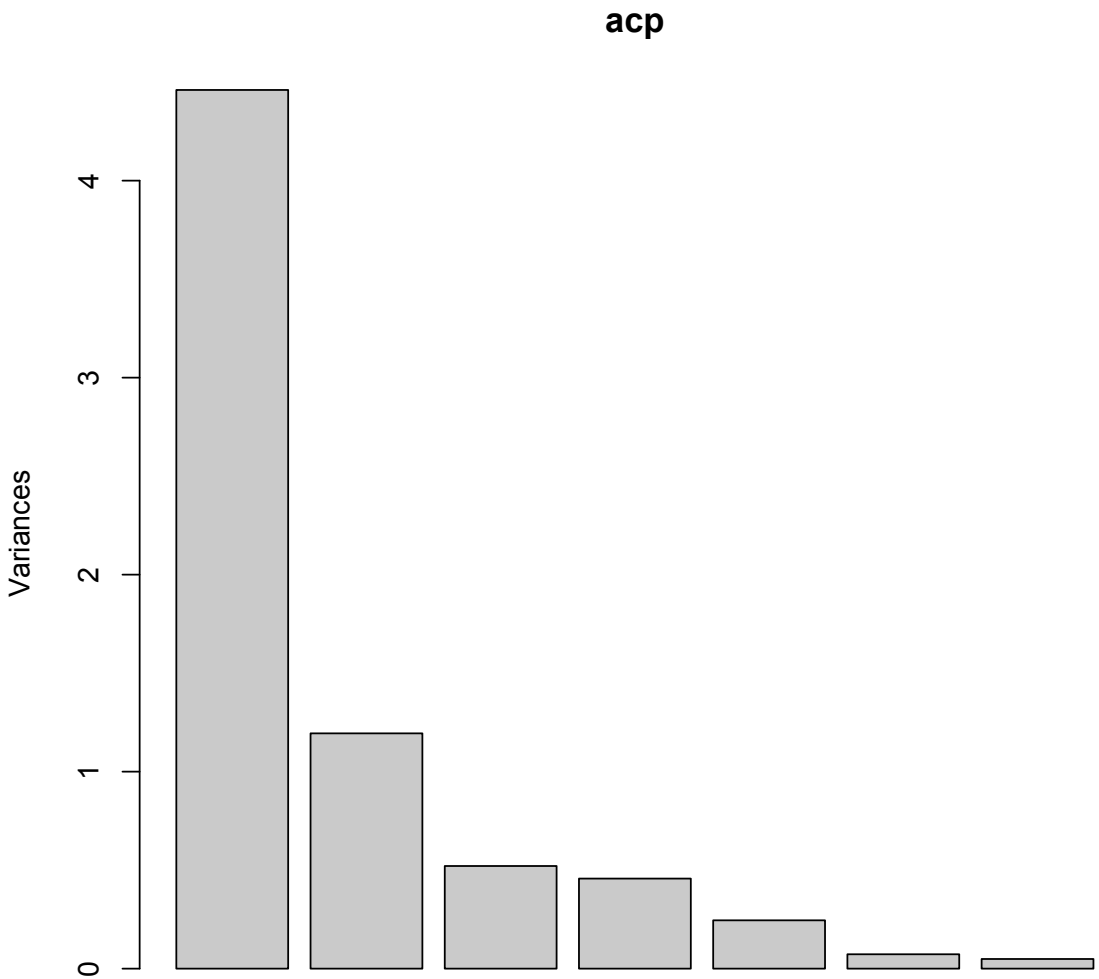
$$\mathrm{Proj}_{<c_1,\ldots,c_d>}(X^{(v)}) = \sum_{k=1}^{d} (a^{(k)})_v \, c_k$$

ACP: biplot


-0.5
0.0
0.5
1.0
6
4
2
0
-2
-4
PC2
1.0
0.5
0.0
-0.5
highjump
longjump
shot
javelin
run200m
hurdles
run800m
1
2
3
4
5
6
8
9
10
11
12
16
17
18
19
20
21
22
23
24
25
-4
-2
0
2
4
6
PC1

ACP : variance expliquée

valeurs propres pour les données d'heptathlon.



Cercle des corrélations: pour chaque variable v on définit le vecteur $\rho^{(v)} \in \mathbb{R}^d$ par

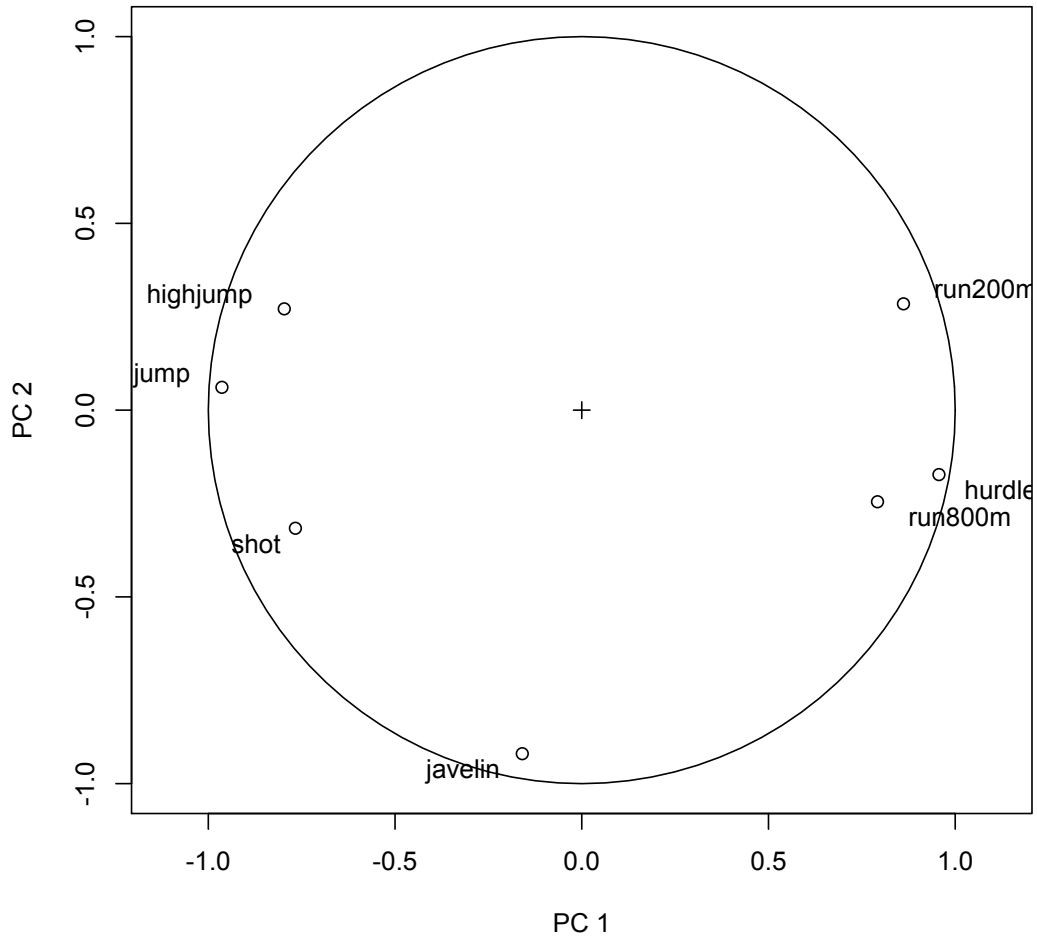
$$\rho_k^{(v)} = \overline{cor}(X^{(v)}, c_k) = \frac{\langle X^{(v)}, c_k \rangle}{\|X^{(v)}\| \, \|c_k\|}, \quad k = 1, \ldots, d.$$

On a

$$\|\rho^{(v)}\|^2 \;=\; \frac{\|\mathrm{Proj}_{<c_1,\ldots,c_d>}(X^{(v)})\|^2}{\|X^{(v)}\|^2} \;\leq\; 1.$$

La norme de $\|\rho^{(v)}\|$ représente la **qualité** de la représentation de la variable v par les d premiers axes.

Cercle des corrélations : $d = 2$



Les variables sont bien expliquées par les deux premières composantes (proche du cercle)

ACP: exemple 3

Exemple: budget de l'état français sur 24 années.

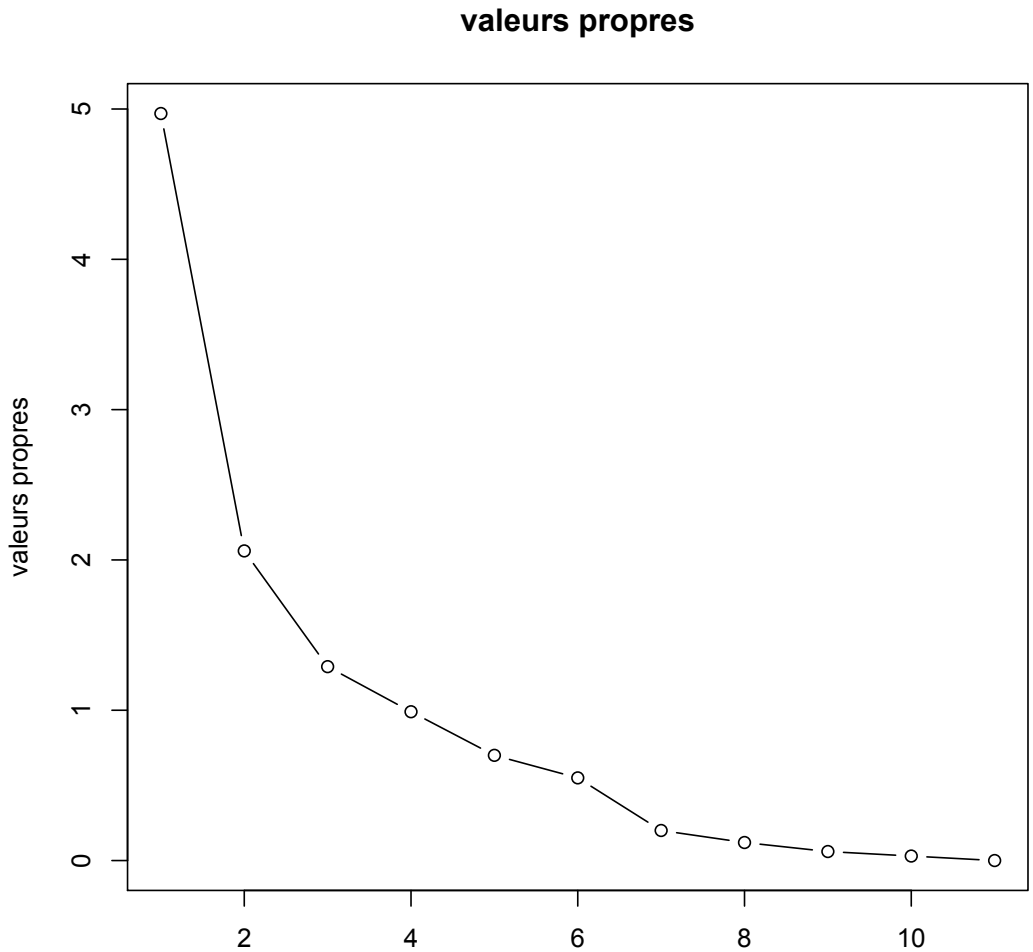
Les variables: part du budget alloué à différents postes (en pourcentage du budget)

PVP:	Pouvoirs publics	AGR:	Agriculture
CMI:	Commerce et industrie	TRA:	Travail
LOG:	Logement	EDU:	Éducation
ACS:	Action sociale	ANC:	Ancien combattants
DEF:	Défense	DET:	Remboursement dette
DIV:	Divers		

donc $p = 11$

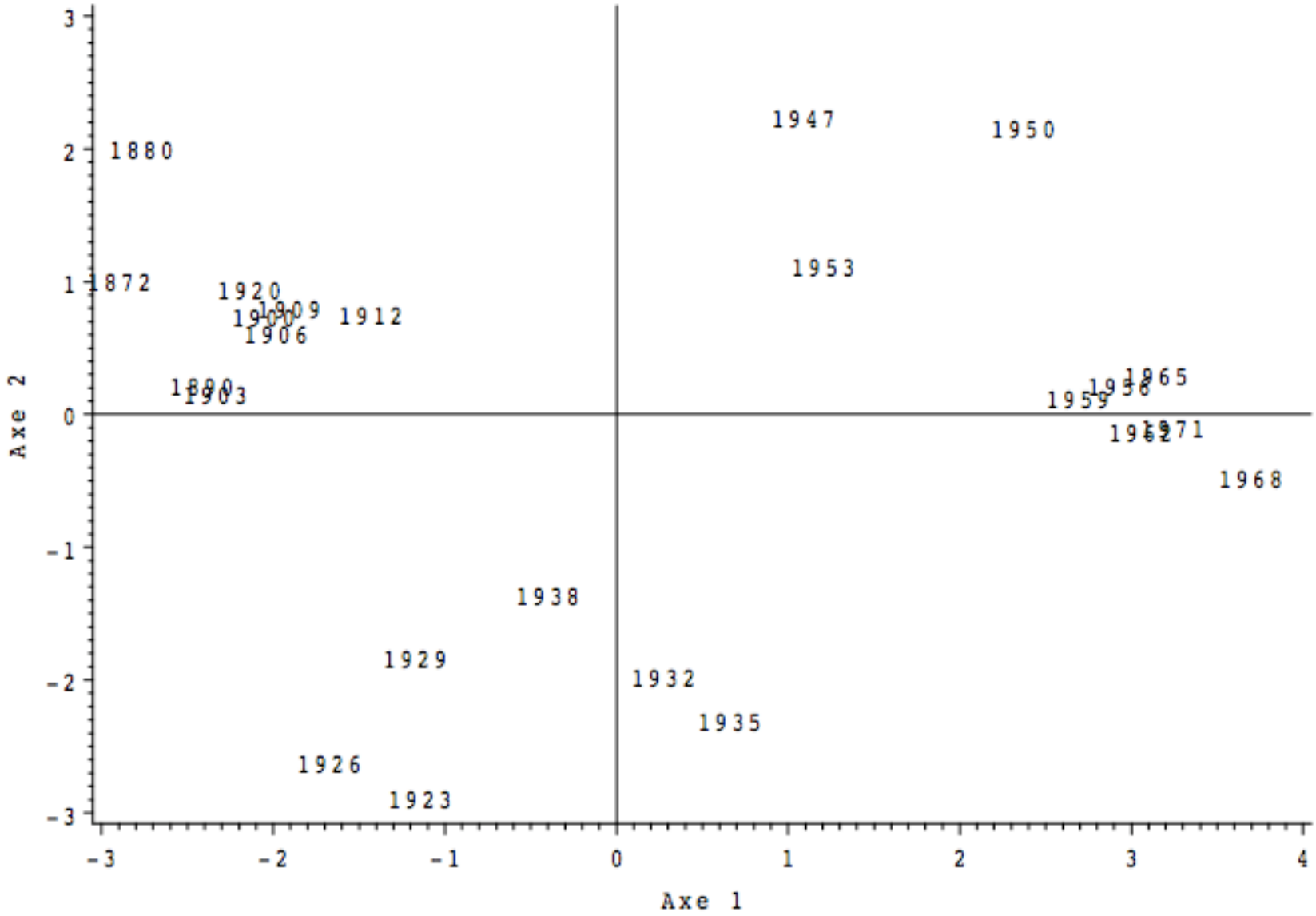
Observations: on a 24 observations pour chaque variable ($n = 24$)

ACP: exemple 3



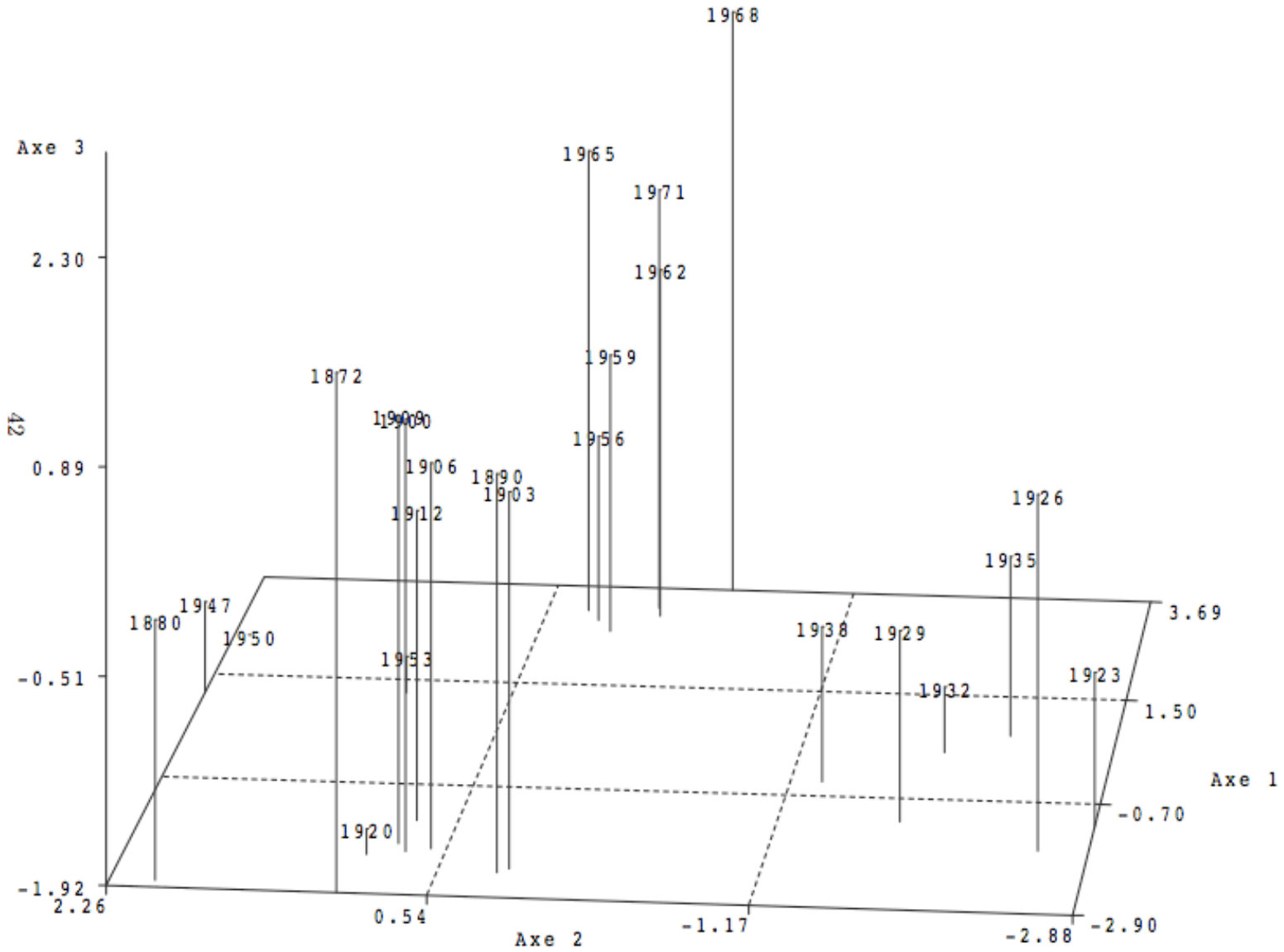
ACP: exemple 3

Projection sur les 2 premiers axes

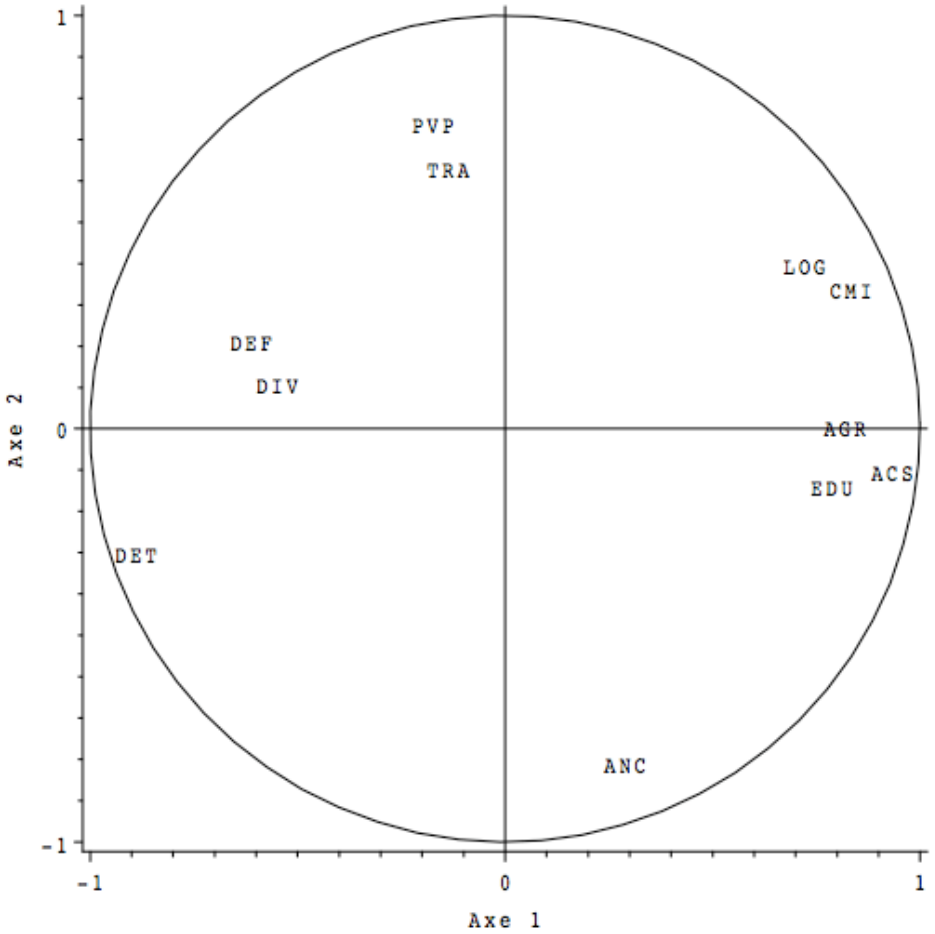


ACP: exemple 3

Projection sur les 3 premiers axes



Cercle des corrélations



Variables proches du cercle: bien expliquées par les deux premiers axes.

ACP : récapitulatif

Axes principaux: $a^{(1)} \perp \ldots \perp a^{(d)} \in \mathbb{R}^p$ vecteurs propres orthonormés de $\widehat{\Sigma} = \frac{1}{n}\mathbf{X}^T\mathbf{X}$,

Composantes principales: $c_1 \perp \ldots \perp c_d \in \mathbb{R}^n$, avec $c_k = \mathbf{X}a^{(k)}$

Projection des individus: $\mathrm{Proj}_{V_d}(X_i) = \sum_{k=1}^{d} (c_k)_i \, a^{(k)}$

Projection des variables: $\mathrm{Proj}_{<c_1,\ldots,c_d>}(X^{(v)}) = \sum_{k=1}^{d} (a^{(k)})_v \, c_k$

Ratio de variance expliquée: par les d premières composantes

$$\frac{\lambda_1 + \ldots + \lambda_d}{\lambda_1 + \ldots + \lambda_p}.$$