

Fondements statistiques de l'apprentissage automatique

Chapitre 4 : réduction de dimension par ACP et PLS

Records nationaux (en secondes) de $p = 9$ épreuves d'athlétisme pour $n = 26$ pays

	100m	200m	400m	800m	1500m	5000m	10000m	SemiMarathon	Marathon
Australie	9.93	20.06	44.38	104.40	211.96	775.76	1649.73	3602	7671
Belgique	10.02	20.19	44.78	103.86	214.13	769.71	1612.30	3605	7640
Brésil	10.00	19.89	44.29	101.77	213.25	799.43	1648.12	3573	7565
RoyaumeUni	9.87	19.87	44.36	101.73	209.67	780.41	1638.14	3609	7633
Canada	9.84	20.17	44.44	103.68	211.71	793.96	1656.01	3650	7809
Chine	10.17	20.54	45.25	106.44	216.49	805.14	1670.00	3635	7695
Croatie	10.25	20.76	45.64	104.07	213.30	817.76	1704.32	3827	8225
Ethiopie	10.50	21.08	45.89	106.08	211.13	757.35	1577.53	3535	7439
France	9.99	20.16	44.46	103.15	208.98	778.83	1642.78	3658	7596
Allemagne	10.06	20.20	44.33	103.65	211.58	774.70	1641.53	3634	7727
Inde	10.30	20.73	45.48	105.77	218.00	809.70	1682.89	3672	7920
Iran	10.29	21.11	46.37	104.74	218.80	833.40	1762.65	4103	8903
Italie	10.01	19.72	45.19	103.17	212.78	785.59	1636.50	3620	7642
Jamaïque	9.58	19.19	44.49	105.21	219.19	813.10	1712.44	3816	8199
Japon	10.00	20.03	44.78	106.18	217.42	793.20	1655.09	3625	7576
Kenya	10.26	20.43	44.18	102.01	206.34	759.74	1587.85	3513	7467
Lituanie	10.33	20.88	45.73	106.64	220.90	797.90	1651.50	3851	7955
NouvelleZélande	10.11	20.42	46.09	104.30	212.17	790.19	1661.95	3732	7815
Portugal	9.86	20.01	46.11	104.91	210.07	782.86	1632.47	3665	7596
Russie	10.10	20.23	44.60	102.47	212.28	791.99	1673.12	3675	7747
AfriqueduSud	10.06	20.11	44.59	102.69	213.56	794.16	1649.94	3678	7593
Espagne	10.14	20.59	44.96	103.83	208.95	782.54	1634.44	3592	7562
Suède	10.18	20.30	44.56	105.54	216.49	797.59	1675.74	3655	7838
Suisse	10.16	20.41	44.99	102.55	211.75	787.54	1673.16	3686	7643
Ukraine	10.07	20.00	45.11	105.08	210.33	790.78	1679.80	3711	7635
USA	9.69	19.32	43.18	102.60	209.30	776.27	1633.98	3583	7538

Comment faire un classement général entre les pays ???

→ Somme pondérée des scores, puis classement en fonction du rang.

Records nationaux (en secondes) de $p = 9$ épreuves d'athlétisme pour $n = 26$ pays

	100m	200m	400m	800m	1500m	5000m	10000m	SemiMarathon	Marathon
Australie	9.93	20.06	44.38	104.40	211.96	775.76	1649.73	3602	7671
Belgique	10.02	20.19	44.78	103.86	214.13	769.71	1612.30	3605	7640
Brésil	10.00	19.89	44.29	101.77	213.25	799.43	1648.12	3573	7565
RoyaumeUni	9.87	19.87	44.36	101.73	209.67	780.41	1638.14	3609	7633
Canada	9.84	20.17	44.44	103.68	211.71	793.96	1656.01	3650	7809
Chine	10.17	20.54	45.25	106.44	216.49	805.14	1670.00	3635	7695
Croatie	10.25	20.76	45.64	104.07	213.30	817.76	1704.32	3827	8225
Ethiopie	10.50	21.08	45.89	106.08	211.13	757.35	1577.53	3535	7439
France	9.99	20.16	44.46	103.15	208.98	778.83	1642.78	3658	7596
Allemagne	10.06	20.20	44.33	103.65	211.58	774.70	1641.53	3634	7727
Inde	10.30	20.73	45.48	105.77	218.00	809.70	1682.89	3672	7920
Iran	10.29	21.11	46.37	104.74	218.80	833.40	1762.65	4103	8903
Italie	10.01	19.72	45.19	103.17	212.78	785.59	1636.50	3620	7642
Jamaïque	9.58	19.19	44.49	105.21	219.19	813.10	1712.44	3816	8199
Japon	10.00	20.03	44.78	106.18	217.42	793.20	1655.09	3625	7576
Kenya	10.26	20.43	44.18	102.01	206.34	759.74	1587.85	3513	7467
Lituanie	10.33	20.88	45.73	106.64	220.90	797.90	1651.50	3851	7955
NouvelleZélande	10.11	20.42	46.09	104.30	212.17	790.19	1661.95	3732	7815
Portugal	9.86	20.01	46.11	104.91	210.07	782.86	1632.47	3665	7596
Russie	10.10	20.23	44.60	102.47	212.28	791.99	1673.12	3675	7747
AfriqueduSud	10.06	20.11	44.59	102.69	213.56	794.16	1649.94	3678	7593
Espagne	10.14	20.59	44.96	103.83	208.95	782.54	1634.44	3592	7562
Suède	10.18	20.30	44.56	105.54	216.49	797.59	1675.74	3655	7838
Suisse	10.16	20.41	44.99	102.55	211.75	787.54	1673.16	3686	7643
Ukraine	10.07	20.00	45.11	105.08	210.33	790.78	1679.80	3711	7635
USA	9.69	19.32	43.18	102.60	209.30	776.27	1633.98	3583	7538

→ Matrice M

Comment faire un classement général entre les pays ???

→ Somme pondérée des scores, puis classement en fonction du rang.

Somme pondérée des scores est équivalente à une multiplication matrice \times vecteur :

→ Vecteur contenant les scores = $M \cdot w$

Records nationaux (en secondes) de $p = 9$ épreuves d'athlétisme pour $n = 26$ pays

	100m	200m	400m	800m	1500m	5000m	10000m	SemiMarathon	Marathon
Australie	9.93	20.06	44.38	104.40	211.96	775.76	1649.73	3602	7671
Belgique	10.02	20.19	44.78	103.86	214.13	769.71	1612.30	3605	7640
Brésil	10.00	19.89	44.29	101.77	213.25	799.43	1648.12	3573	7565
RoyaumeUni	9.87	19.87	44.36	101.73	209.67	780.41	1638.14	3609	7633
Canada	9.84	20.17	44.44	103.68	211.71	793.96	1656.01	3650	7809
Chine	10.17	20.54	45.25	106.44	216.49	805.14	1670.00	3635	7695
Croatie	10.25	20.76	45.64	104.07	213.30	817.76	1704.32	3827	8225
Ethiopie	10.50	21.08	45.89	106.08	211.13	757.35	1577.53	3535	7439
France	9.99	20.16	44.46	103.15	208.98	778.83	1642.78	3658	7596
Allemagne	10.06	20.20	44.33	103.65	211.58	774.70	1641.53	3634	7727
Inde	10.30	20.73	45.48	105.77	218.00	809.70	1682.89	3672	7920
Iran	10.29	21.11	46.37	104.74	218.80	833.40	1762.65	4103	8903
Italie	10.01	19.72	45.19	103.17	212.78	785.59	1636.50	3620	7642
Jamaïque	9.58	19.19	44.49	105.21	219.19	813.10	1712.44	3816	8199
Japon	10.00	20.03	44.78	106.18	217.42	793.20	1655.09	3625	7576
Kenya	10.26	20.43	44.18	102.01	206.34	759.74	1587.85	3513	7467
Lituanie	10.33	20.88	45.73	106.64	220.90	797.90	1651.50	3851	7955
NouvelleZélande	10.11	20.42	46.09	104.30	212.17	790.19	1661.95	3732	7815
Portugal	9.86	20.01	46.11	104.91	210.07	782.86	1632.47	3665	7596
Russie	10.10	20.23	44.60	102.47	212.28	791.99	1673.12	3675	7747
AfriqueduSud	10.06	20.11	44.59	102.69	213.56	794.16	1649.94	3678	7593
Espagne	10.14	20.59	44.96	103.83	208.95	782.54	1634.44	3592	7562
Suède	10.18	20.30	44.56	105.54	216.49	797.59	1675.74	3655	7838
Suisse	10.16	20.41	44.99	102.55	211.75	787.54	1673.16	3686	7643
Ukraine	10.07	20.00	45.11	105.08	210.33	790.78	1679.80	3711	7635
USA	9.69	19.32	43.18	102.60	209.30	776.27	1633.98	3583	7538

→ Matrice M

On peut aussi chercher le vecteur (de norme 1) qui maximise la variabilité entre les scores

- Vecteur optimal = 1^{er} vecteur propre (v_1) de l'ACP
- Niveau de variabilité = 1^{ere} valeur propre (λ_1) de l'ACP
- Vecteur de scores avec la plus grande variabilité possible = $M \cdot v_1$

Records nationaux (en secondes) de $p = 9$ épreuves d'athlétisme pour $n = 26$ pays

	100m	200m	400m	800m	1500m	5000m	10000m	SemiMarathon	Marathon
Australie	9.93	20.06	44.38	104.40	211.96	775.76	1649.73	3602	7671
Belgique	10.02	20.19	44.78	103.86	214.13	769.71	1612.30	3605	7640
Brésil	10.00	19.89	44.29	101.77	213.25	799.43	1648.12	3573	7565
RoyaumeUni	9.87	19.87	44.36	101.73	209.67	780.41	1638.14	3609	7633
Canada	9.84	20.17	44.44	103.68	211.71	793.96	1656.01	3650	7809
Chine	10.17	20.54	45.25	106.44	216.49	805.14	1670.00	3635	7695
Croatie	10.25	20.76	45.64	104.07	213.30	817.76	1704.32	3827	8225
Ethiopie	10.50	21.08	45.89	106.08	211.13	757.35	1577.53	3535	7439
France	9.99	20.16	44.46	103.15	208.98	778.83	1642.78	3658	7596
Allemagne	10.06	20.20	44.33	103.65	211.58	774.70	1641.53	3634	7727
Inde	10.30	20.73	45.48	105.77	218.00	809.70	1682.89	3672	7920
Iran	10.29	21.11	46.37	104.74	218.80	833.40	1762.65	4103	8903
Italie	10.01	19.72	45.19	103.17	212.78	785.59	1636.50	3620	7642
Jamaïque	9.58	19.19	44.49	105.21	219.19	813.10	1712.44	3816	8199
Japon	10.00	20.03	44.78	106.18	217.42	793.20	1655.09	3625	7576
Kenya	10.26	20.43	44.18	102.01	206.34	759.74	1587.85	3513	7467
Lithuanie	10.33	20.88	45.73	106.64	220.90	797.90	1651.50	3851	7955
NouvelleZélande	10.11	20.42	46.09	104.30	212.17	790.19	1661.95	3732	7815
Portugal	9.86	20.01	46.11	104.91	210.07	782.86	1632.47	3665	7596
Russie	10.10	20.23	44.60	102.47	212.28	791.99	1673.12	3675	7747
Afrique du Sud	10.06	20.11	44.59	102.69	213.56	794.16	1649.94	3678	7593
Espagne	10.14	20.59	44.96	103.83	208.95	782.54	1634.44	3592	7562
Suède	10.18	20.30	44.56	105.54	216.49	797.59	1675.74	3655	7838
Suisse	10.16	20.41	44.99	102.55	211.75	787.54	1673.16	3686	7643
Ukraine	10.07	20.00	45.11	105.08	210.33	790.78	1679.80	3711	7635
USA	9.69	19.32	43.18	102.60	209.30	776.27	1633.98	3583	7538

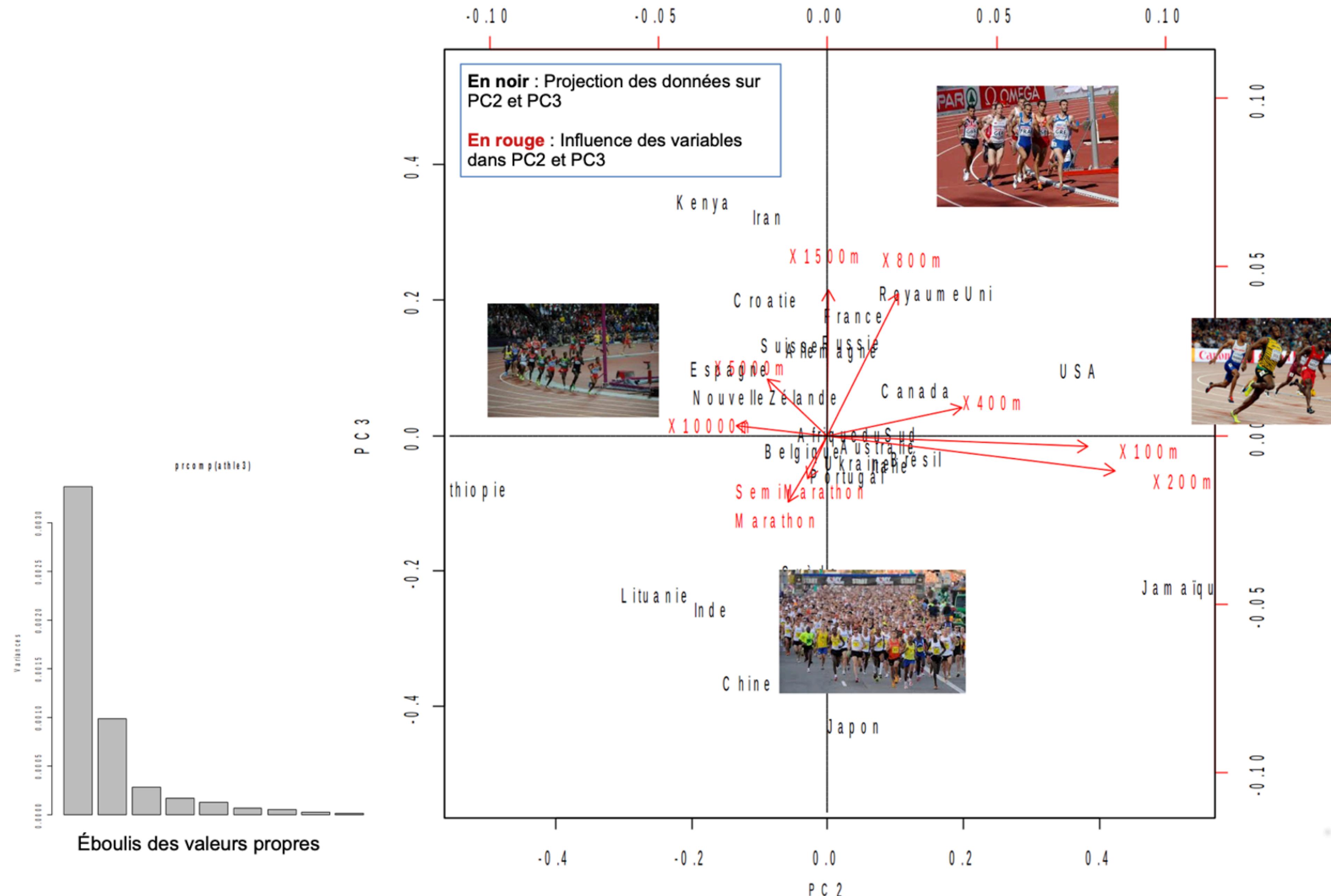
→ Matrice **M**

Une fois enlevée l'influence de v_1 , on cherche le vecteur (de norme 1) qui maximise la variabilité

- Vecteur optimal = 2^{er} vecteur propre (v_2) de l'ACP
- Niveau de variabilité = 2^{ere} valeur propre (λ_2) de l'ACP

...

Calculable de manière analytique

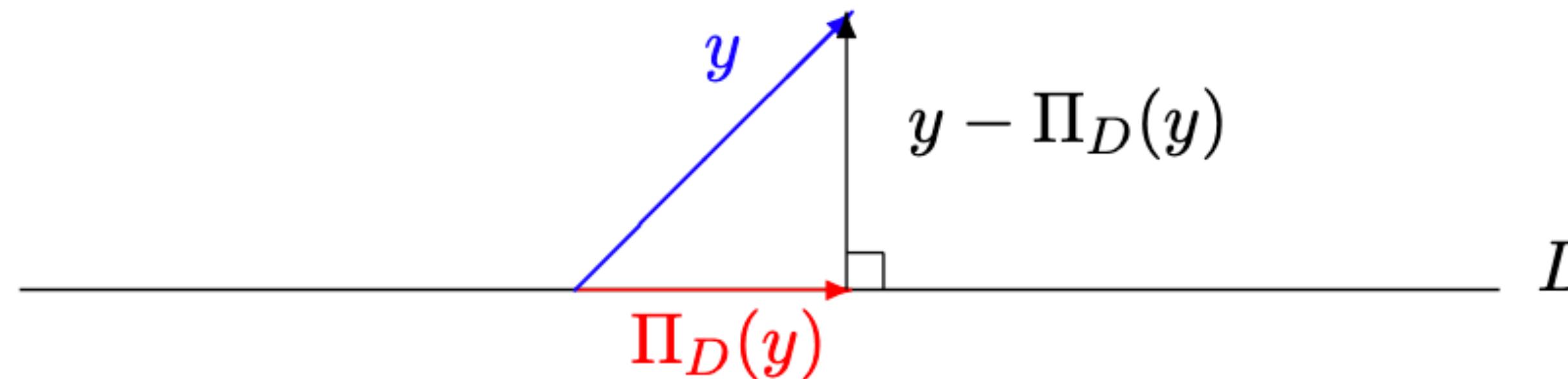


Soit x un vecteur non nul de E et soit D la droite vectorielle engendrée par x , i.e. $D = \{\lambda x \mid \lambda \in \mathbb{R}\}$. Soit $y \in E$. On veut projeter le vecteur y orthogonalement sur la droite D . On note $\Pi_D(y)$ la projection orthogonale de y sur D . Elle est caractérisée par les deux propriétés suivantes :

- (i) $\Pi_D(y) \in D$, i.e. $\Pi_D(y)$ est colinéaire à x
- (ii) $y - \Pi_D(y)$ est orthogonal à x .

On en déduit la solution :

$$\Pi_D(y) = \frac{\langle x, y \rangle}{\|x\|^2} x .$$



Soit A une matrice de taille $m \times n$. Soit A_1, \dots, A_n ses colonnes. Ce sont des vecteurs de \mathbb{R}^m . Par exemple, si

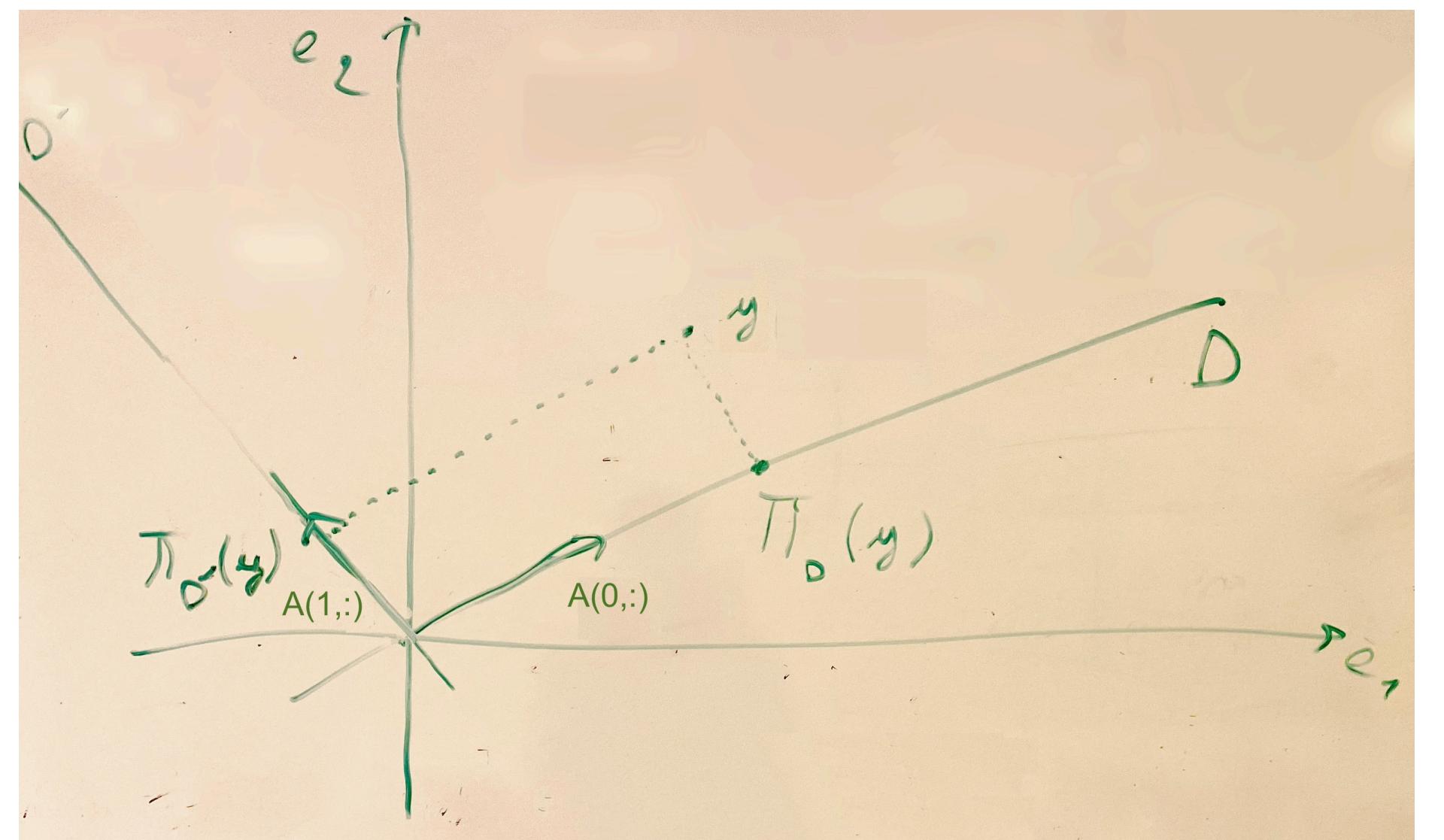
$$A = \begin{pmatrix} 1 & 2 \\ 3 & -4 \\ 2 & 0 \end{pmatrix},$$

ses colonnes sont

$$A_1 = \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 2 \\ -4 \\ 0 \end{pmatrix}$$

On cherche à déterminer la projection orthogonale d'un vecteur y sur le sous espace vectoriel de \mathbb{R}^m engendré par A_1, \dots, A_n . Cette projection, notée Π_A est caractérisée par les propriétés

- (i) $\Pi_A(y)$ est une combinaison linéaire des colonnes de A ;
- (ii) $y - \Pi_A(y)$ est orthogonal aux colonnes de A .

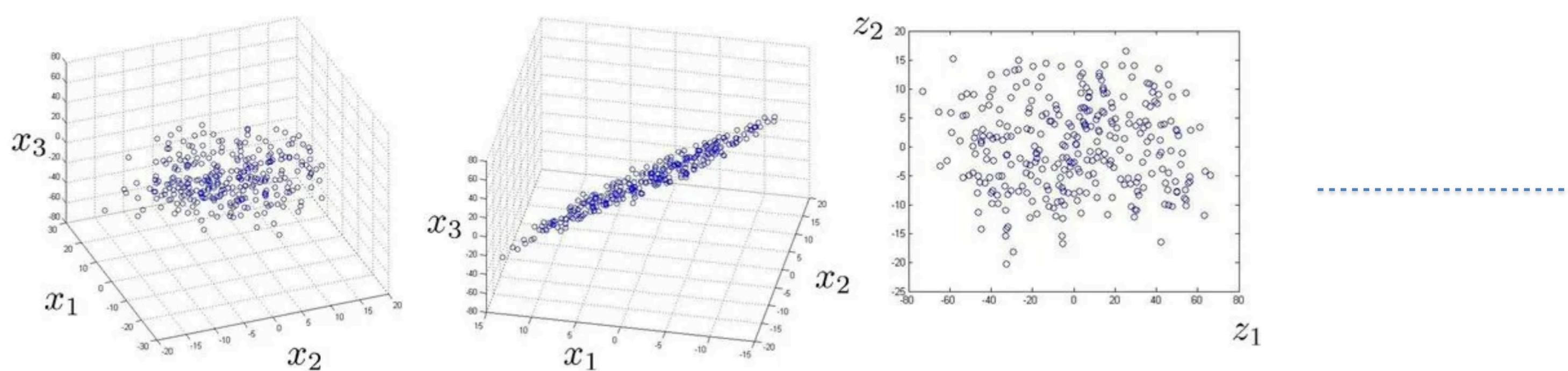


Dimension avant projection Dimension après projection

Théorème Soit A une matrice de taille $m \times n$ et soit $y \in \mathbb{R}^m$. Si la matrice ${}^t A A$ est inversible, alors la projection orthogonale de y sur le sous-espace de \mathbb{R}^m engendré par les colonnes de A est

$$\left(A \underbrace{\left(A^\top A \right)^{-1/2}}_{\text{↓}} \right) \left(\underbrace{\left(A^\top A \right)^{-1/2} A^\top}_{\text{↓}} \right) y = A \underbrace{\left(A^\top A \right)^{-1}}_{\text{↓}} A^\top y$$

- Matrice diagonale si colonnes de projection orthogonale
- Matrice unité si colonnes de projection orthonormales

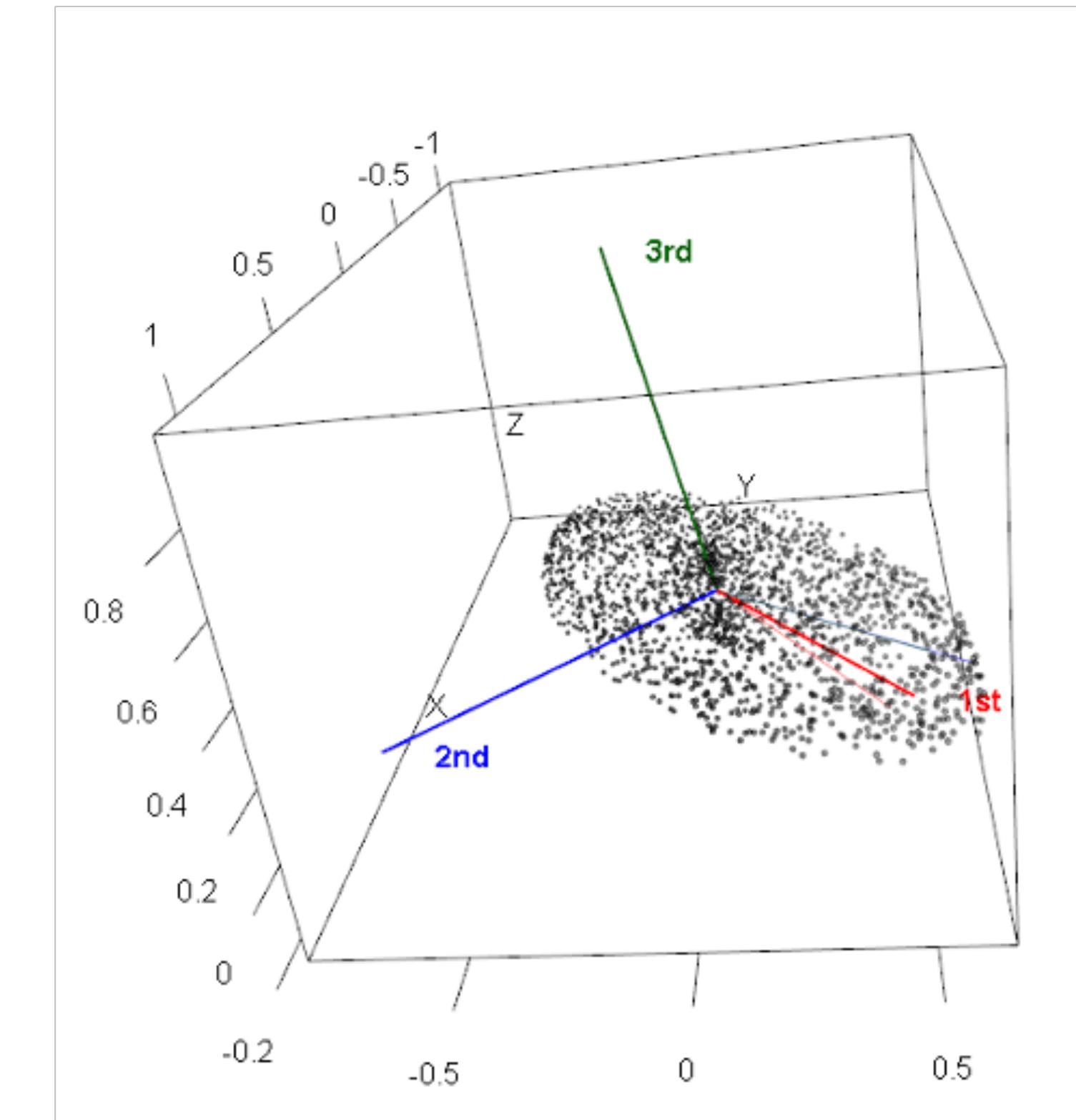
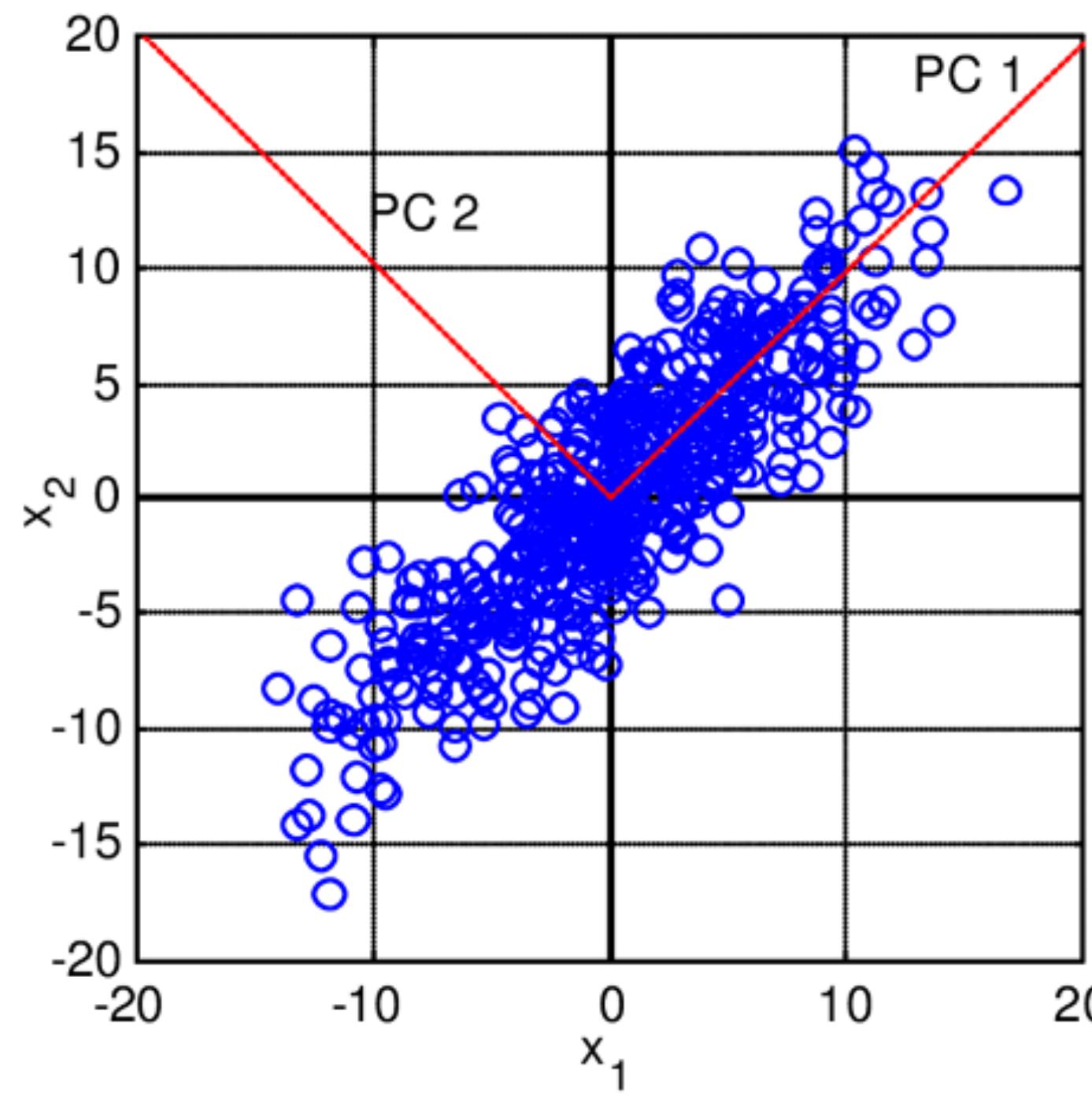


Données originales $y \in \mathbb{R}^3$

Projection $\left(\left(A^\top A \right)^{-1/2} A^\top \right) y \in \mathbb{R}^2$

Reconstruction de la projection
 $A \left(A^\top A \right)^{-1} A^\top y \in \mathbb{R}^3$

L'objectif de l'ACP est de déterminer le *méilleur* espace sur lequel projeter les données tout en conservant leur *structure*. L'idée principale consiste donc à trouver la projection dans laquelle les données projetées seront les plus *dispersées* possible.



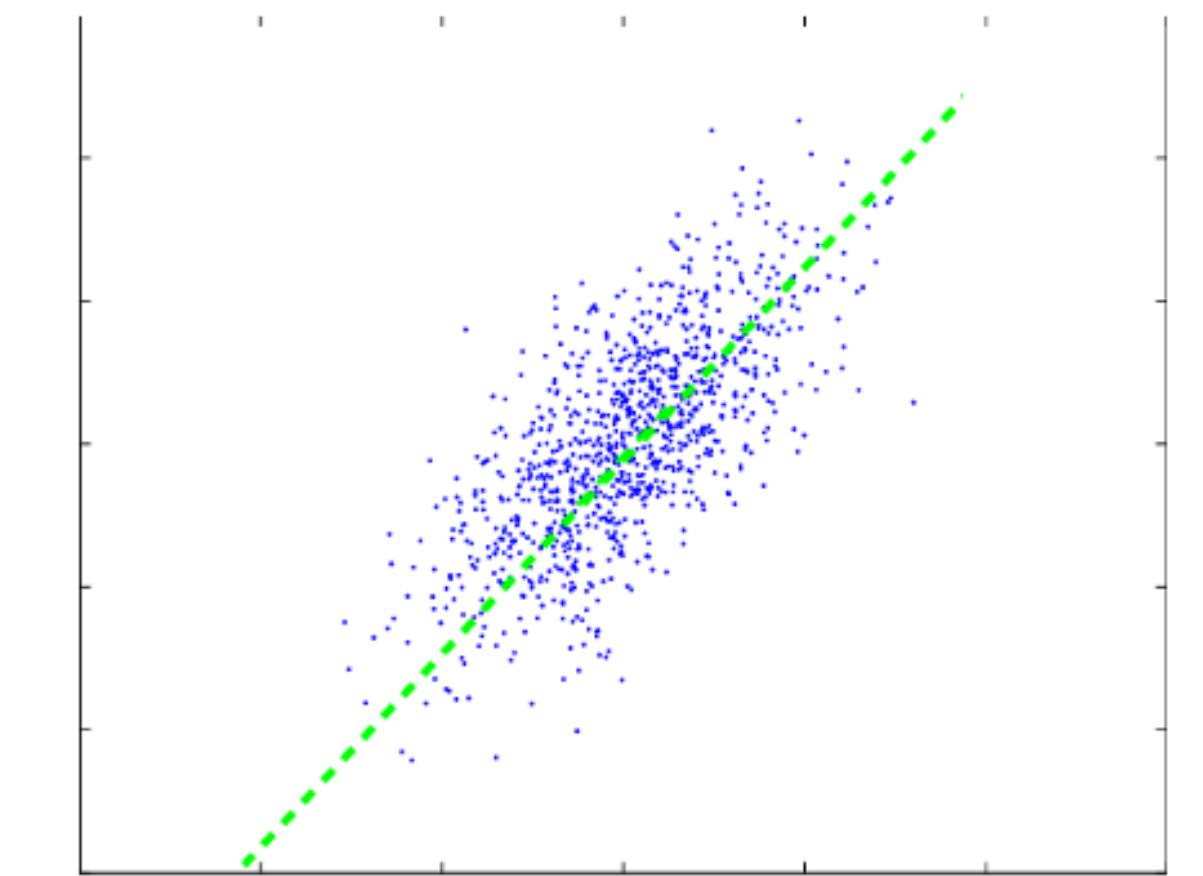
On dispose de n observations $X_i \in \mathbb{R}^p$.

On a alors $X = [x_1, x_2, \dots, x_n] = \begin{bmatrix} x_1^1 & x_2^1 & \dots & x_n^1 \\ \vdots & \vdots & & \vdots \\ x_1^p & x_2^p & \dots & x_n^p \end{bmatrix} \in \mathbb{R}^{n \times p}$

et la matrice de covariance empirique est défini par : $S = \frac{1}{n-1} X X^T \in \mathbb{R}^{p \times p}$

Par exemple, on a en 2D : $S = \begin{pmatrix} \sigma(x, x) & \sigma(x, y) \\ \sigma(y, x) & \sigma(y, y) \end{pmatrix}$

avec $\sigma(x, y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$



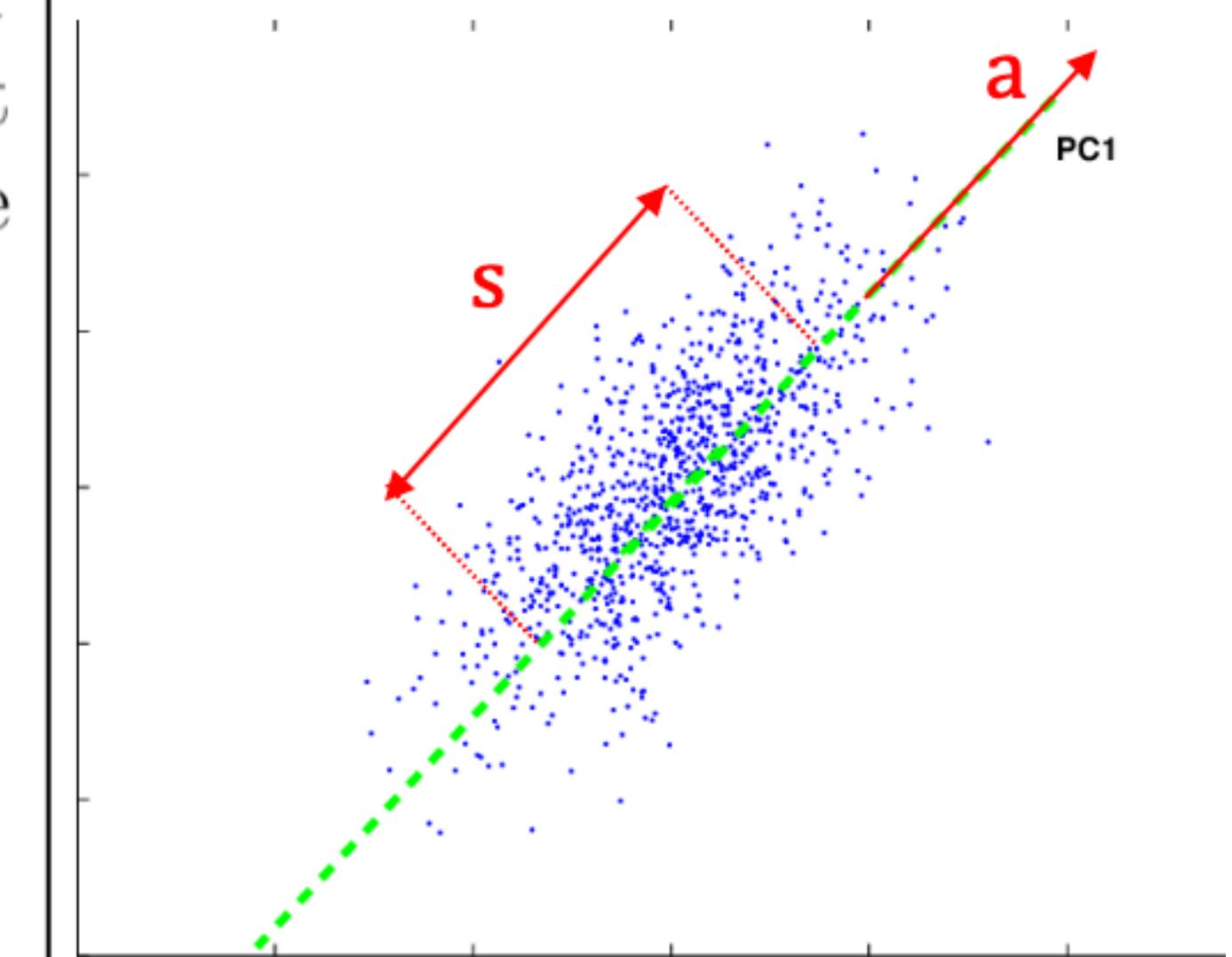
On dispose de n observations $X_i \in \mathbb{R}^p$.

Chercher la première direction sur laquelle projeter revient à trouver le vecteur $a \in \mathbb{R}^p$ tel que l'échantillon projeté $(a^T X_1, \dots, a^T X_n)$ ait une variance maximale. Ainsi l'information des données est assimilée à leur variabilité. La variance empirique s'écrit pour cet échantillon

$$\begin{aligned}s^2 &= \frac{1}{n} \sum_{i=1}^n (a^T X_i)^2 - \left(\frac{1}{n} \sum_{i=1}^n a^T X_i \right)^2 \\ &= a^T \frac{1}{n} \sum_{i=1}^n X_i X_i^T a - a^T \left(\frac{1}{n} \sum_{i=1}^n X_i \right) \left(\frac{1}{n} \sum_{i=1}^n X_i \right)^T a \\ &= a^T S a,\end{aligned}$$

où S est la variance empirique des X_1, \dots, X_n . Ainsi maximiser la variance des points projetés est équivalent à chercher à trouver la solution du problème de maximisation

$$\arg \max_{a, \|a\|=1} a^T S a.$$



Remarques :

- a est la direction principale empirique.
- L'opération peut être répétée itérativement en enlevant aux X_i leur projection sur a .

FS-AA - chapitre 4 : Réduction de dimension → 4.3 ACP (interprétation par maximisation des variances)

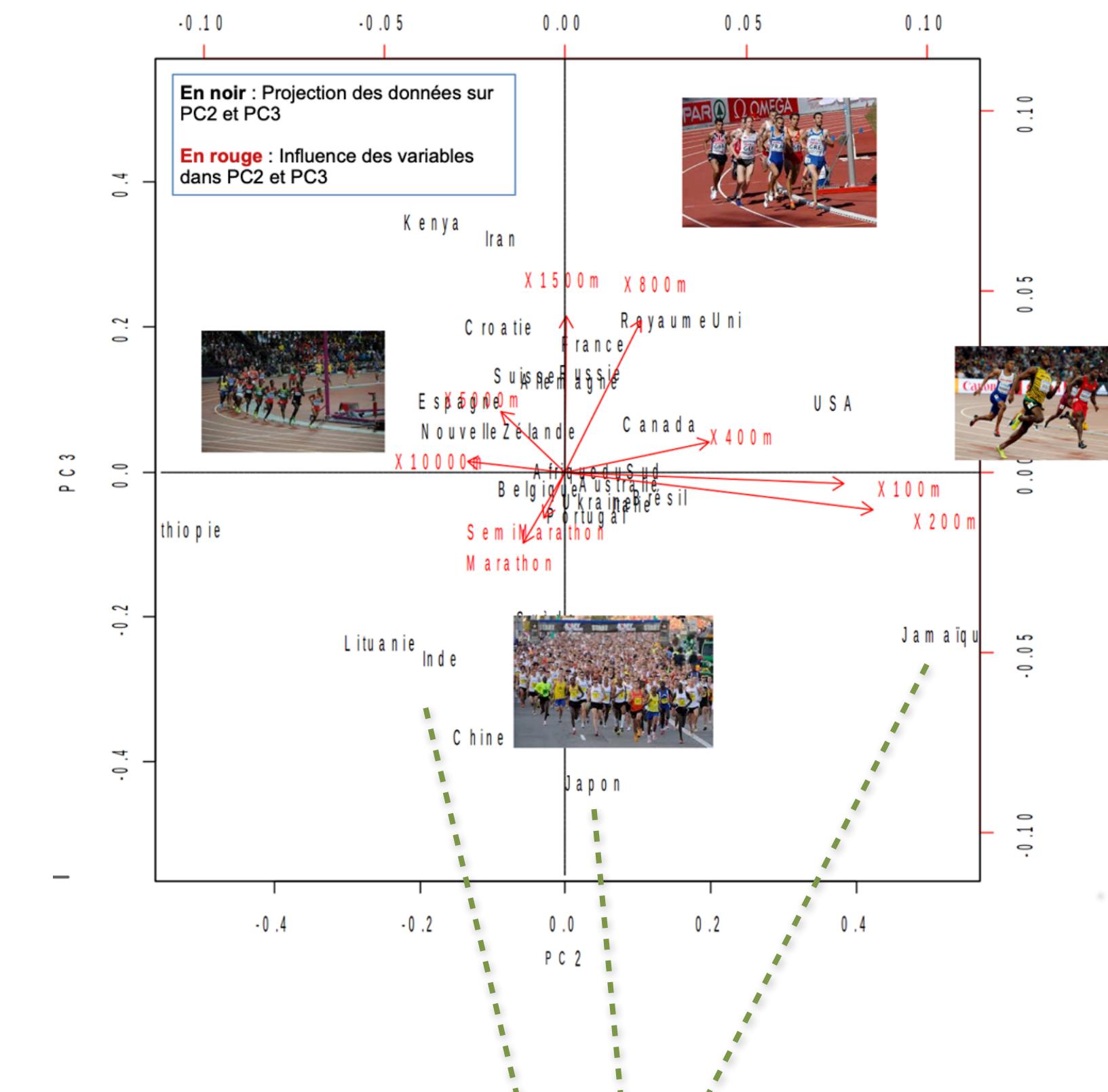
	100m	200m	400m	800m	1500m	5000m	10000m	SemiMarathon	Marathon
Australie	9.93	20.06	44.38	104.40	211.96	775.76	1649.73	3602	7671
Belgique	10.02	20.19	44.78	103.86	214.13	769.71	1612.30	3605	7640
Brésil	10.00	19.89	44.29	101.77	213.25	799.43	1648.12	3573	7565
RoyaumeUni	9.87	19.87	44.36	101.73	209.67	780.41	1638.14	3609	7633
Canada	9.84	20.17	44.44	103.68	211.71	793.96	1656.01	3650	7809
Chine	10.17	20.54	45.25	106.44	216.49	805.14	1670.00	3635	7695
Croatie	10.25	20.76	45.64	104.07	213.30	817.76	1704.32	3827	8225
Ethiopie	10.50	21.08	45.89	106.08	211.13	757.35	1577.53	3535	7439
France	9.99	20.16	44.46	103.15	208.98	778.83	1642.78	3658	7596
Allemagne	10.06	20.20	44.33	103.65	211.58	774.70	1641.53	3634	7727
Inde	10.30	20.73	45.48	105.77	218.00	809.70	1682.89	3672	7920
Iran	10.29	21.11	46.37	104.74	218.80	833.40	1762.65	4103	8903
Italie	10.01	19.72	45.19	103.17	212.78	785.59	1636.50	3620	7642
Jamaïque	9.58	19.19	44.49	105.21	219.19	813.10	1712.44	3816	8199
Japon	10.00	20.03	44.78	106.18	217.42	793.20	1655.09	3625	7576
Kenya	10.26	20.43	44.18	102.01	206.34	759.74	1587.85	3513	7467
Lituanie	10.33	20.88	45.73	106.64	220.90	797.90	1651.50	3851	7955
NouvelleZélande	10.11	20.42	46.09	104.30	212.17	790.19	1661.95	3732	7815
Portugal	9.86	20.01	46.11	104.91	210.07	782.86	1632.47	3665	7596
Russie	10.10	20.23	44.60	102.47	212.28	791.99	1673.12	3675	7747
Afrique du Sud	10.06	20.11	44.59	102.69	213.56	794.16	1649.94	3678	7593
Espagne	10.14	20.59	44.96	103.83	208.95	782.54	1634.44	3592	7562
Suède	10.18	20.30	44.56	105.54	216.49	797.59	1675.74	3655	7838
Suisse	10.16	20.41	44.99	102.55	211.75	787.54	1673.16	3686	7643
Ukraine	10.07	20.00	45.11	105.08	210.33	790.78	1679.80	3711	7635
USA	9.69	19.32	43.18	102.60	209.30	776.27	1633.98	3583	7538

Données X avant centrage-réduction

$$v_1 = \arg \max_{v \text{ t.q. } \|v\|_2=1} \sum_{i=1}^n (X_i v)^2$$

Premier vecteur propre

Variance de la projection des observations de X avec v

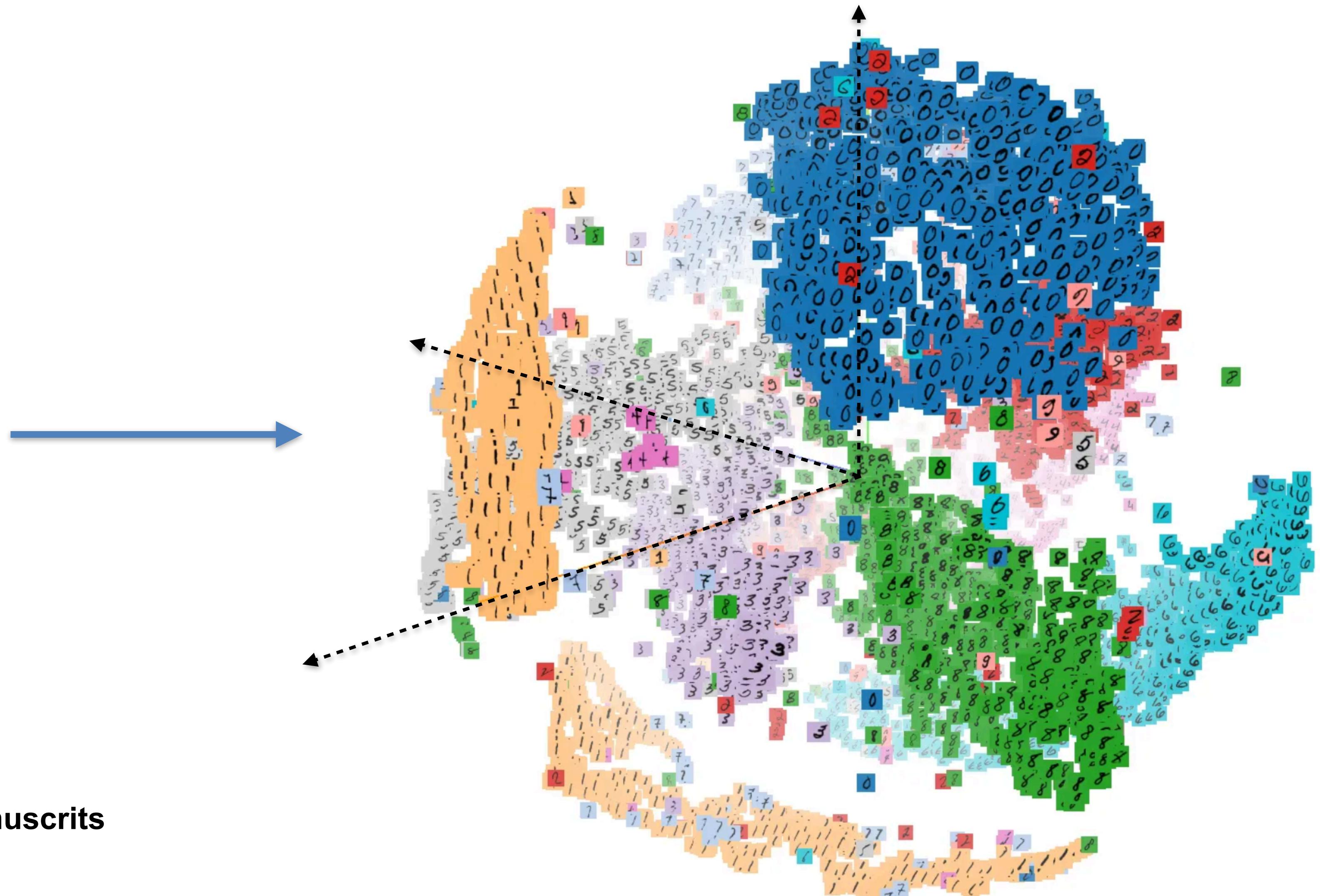


v_2 calculé en rajoutant la contrainte $\langle v_1, v_2 \rangle = 0$; v_3 calculé en rajoutant les contraintes $\langle v_1, v_3 \rangle = 0$ et $\langle v_2, v_3 \rangle = 0$; ...



$X_i \in \mathbb{R}^{28 \times 28}$: images représentant des chiffres manuscrits

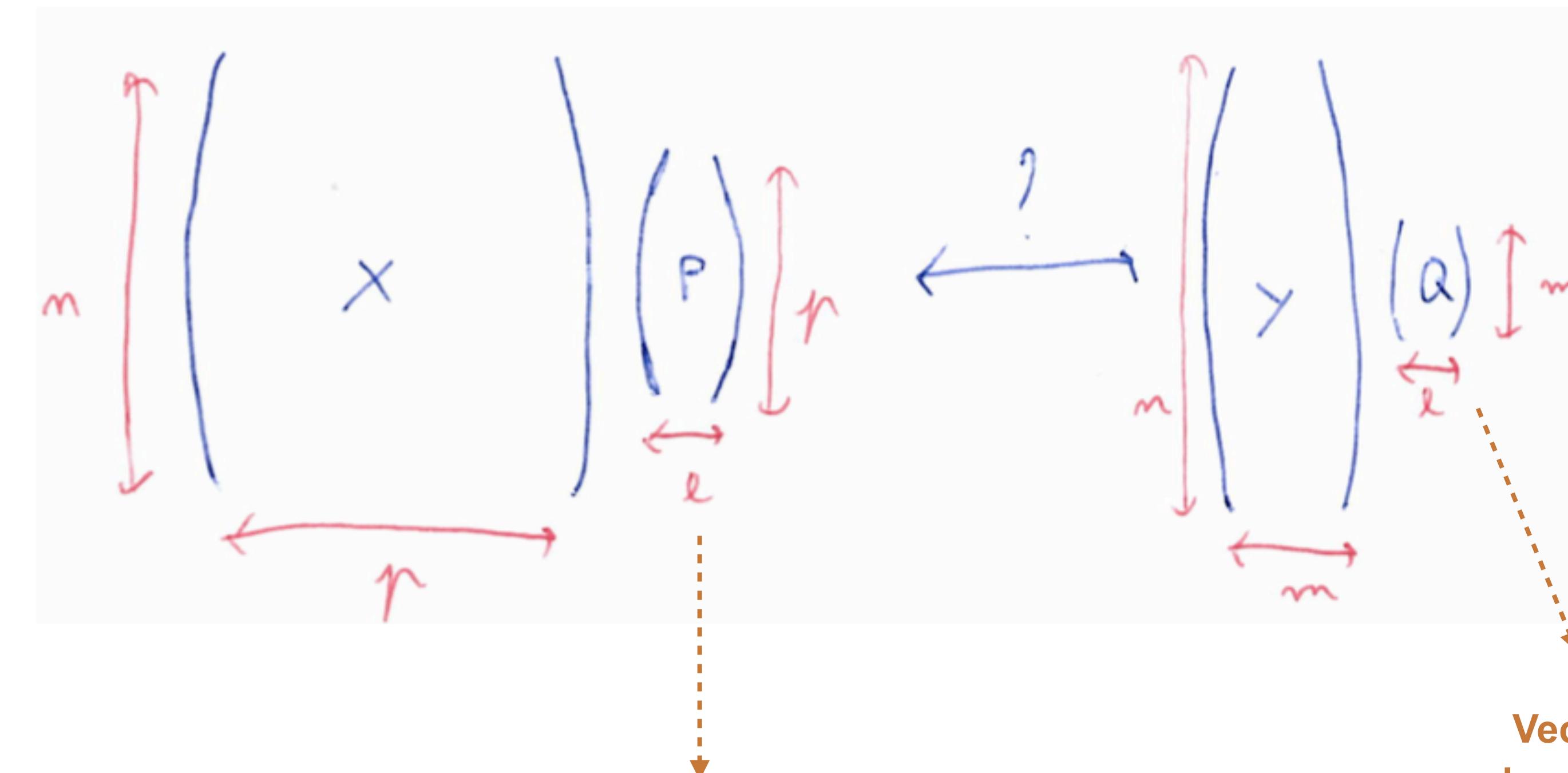
$y_i \in \{0, 1, \dots, 9\}$: chiffre correspondant



Projection des images dans \mathbb{R}^3

Nous avons vu dans la première partie du cours comment mettre en lien des données d'entrée X et des données de sortie y à l'aide de la régression linéaire.

→ Voyons maintenant comment réduire la dimension de X , tout comme avec l'ACP, mais de manière à ce que ses projections permettent d'expliquer au mieux les y !



Vecteurs de projection
des observations de X

Vecteurs de projection
des observations de y (si
multi-dimensionnel)

on suppose :

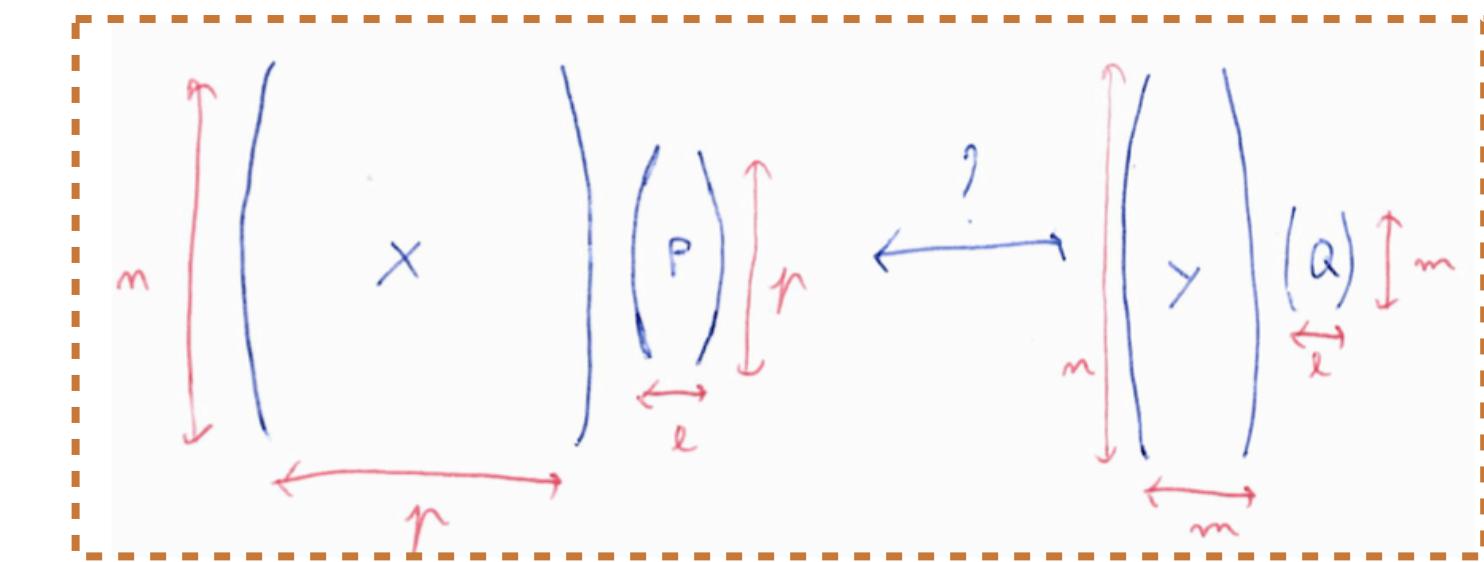
$$X = TP^\top + E$$

$$Y = UQ^\top + F$$

où

- **X** est la matrice $n \times p$ de prédicteurs. Elle est supposée centrée/réduite,
- **Y** est la matrice $n \times m$ de réponses. Elle est supposée centrée/réduite,
- **P** et **Q** sont respectivement des matrices $p \times l$ et $m \times l$ de projection.
Leurs colonnes sont orthonormés.
- **T** et **U** sont les projections de **X** et de **Y** respectivement par **P** et **Q**.
Elles sont de taille $n \times l$.
- **E** et **F** sont des termes d'erreur de même taille que **X** et **Y**. Ils sont supposés *i.i.d.* et distribués suivant une loi normale.

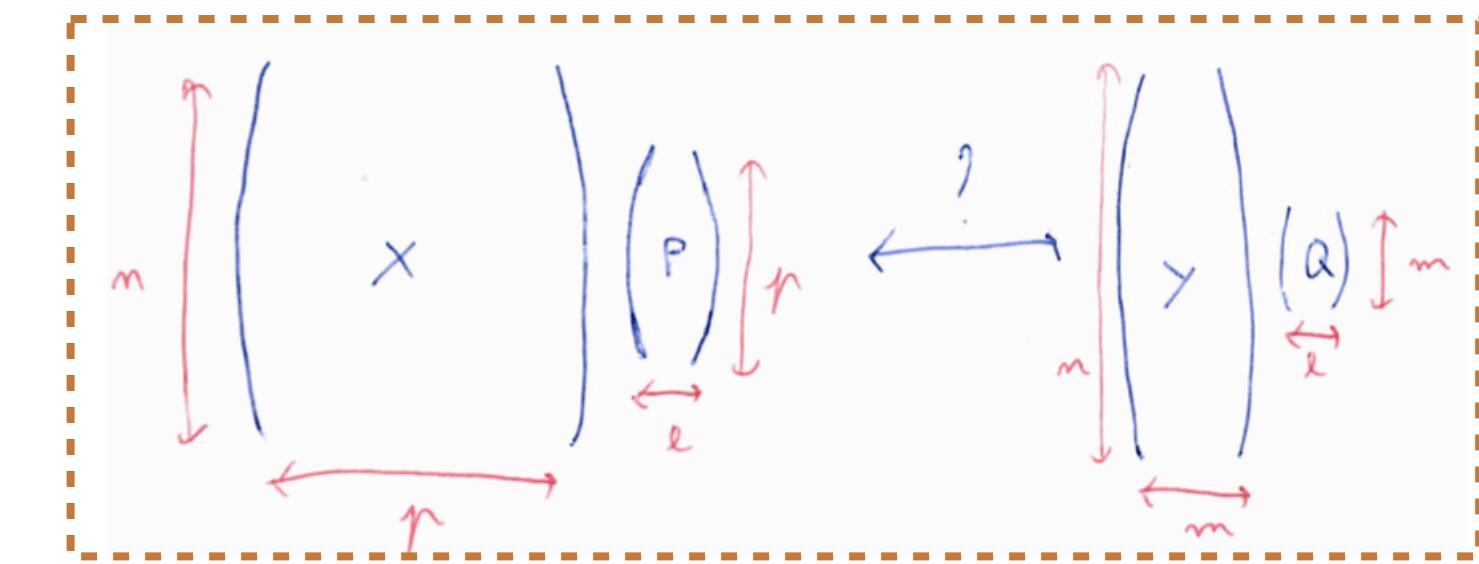
Les projections de **X** et de **Y** dans **T** et **U** sont aussi toutes deux de même taille $n \times l$ avec $l \leq p$.



on suppose :

$$X = TP^\top + E$$

$$Y = UQ^\top + F$$



La PLS consiste alors à calculer les projecteurs P et Q qui maximisent la covariance entre T et U , i.e. :

$$\sum_{j=1}^l \sum_{i=1}^n (T_{ij} - \bar{T}_i)(U_{ij} - \bar{U}_i)$$

où

- $T = XP$ et $U = YQ$ sont respectivement les projections de X et Y
- \bar{T}_j et \bar{U}_j sont les moyennes des valeurs des colonnes j de T et U .

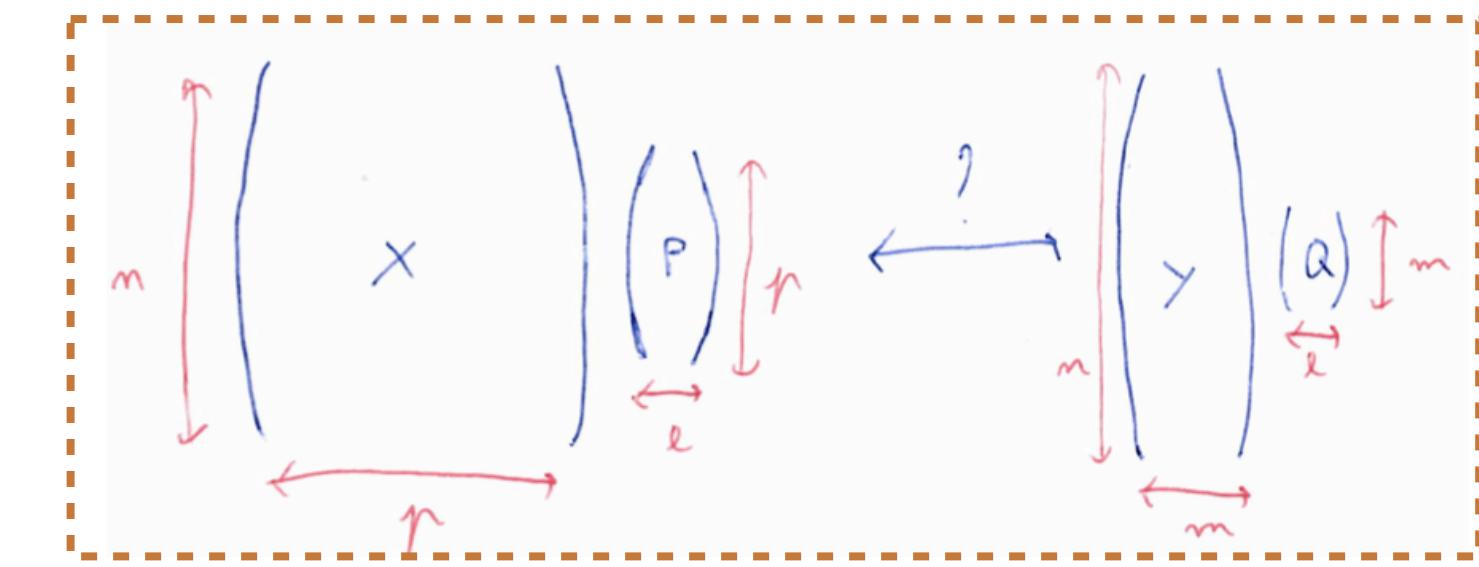
Alg. 1 Fonction $PLS1(\mathbf{X}, \mathbf{y}, l)$

```

1:  $\mathbf{X}^{(0)} \leftarrow \mathbf{X}$ 
2:  $\mathbf{w}^{(0)} \leftarrow \mathbf{X}' \mathbf{y} / |\mathbf{X}' \mathbf{y}|_2$ 
3: for  $k = 0, \dots, l - 1$  do
4:    $\mathbf{t}^{(k)} \leftarrow \mathbf{X}' \mathbf{w}^{(k)}$ 
5:    $t_k \leftarrow \mathbf{t}^{(k)'} \mathbf{t}^{(k)}$ 
6:    $\mathbf{t}^{(k)} \leftarrow \mathbf{t}^{(k)} / t_k$ 
7:    $\mathbf{p}^{(k)} \leftarrow \mathbf{X}^{(k)'} \mathbf{t}^{(k)}$ 
8:    $q_k \leftarrow \mathbf{y}' \mathbf{t}^{(k)}$ 
9:   if  $q_k = 0$  then
10:     $l \leftarrow k$  et sort de la boucle for (toute la variabilité est capturée).
11:   end if
12:   if  $k < (l - 1)$  then
13:      $\mathbf{X}^{(k+1)} \leftarrow \mathbf{X}^{(k)} - t_k \mathbf{t}^{(k)} \mathbf{p}^{(k)'}'$ 
14:      $\mathbf{w}^{(k+1)} \leftarrow \mathbf{X}^{(k+1)'} \mathbf{y} / |\mathbf{X}^{(k+1)'} \mathbf{y}|_2$ 
15:   end if
16: end for

18:  $\mathbf{P}$  est la matrice composée des colonnes  $\mathbf{p}^{(0)}, \mathbf{p}^{(1)}, \dots, \mathbf{p}^{(l-1)}$ .

```



On suppose ici que $m = 1$, c'est à dire que les observations de y sont en dimension 1.

→ Pas d'estimation de Q .

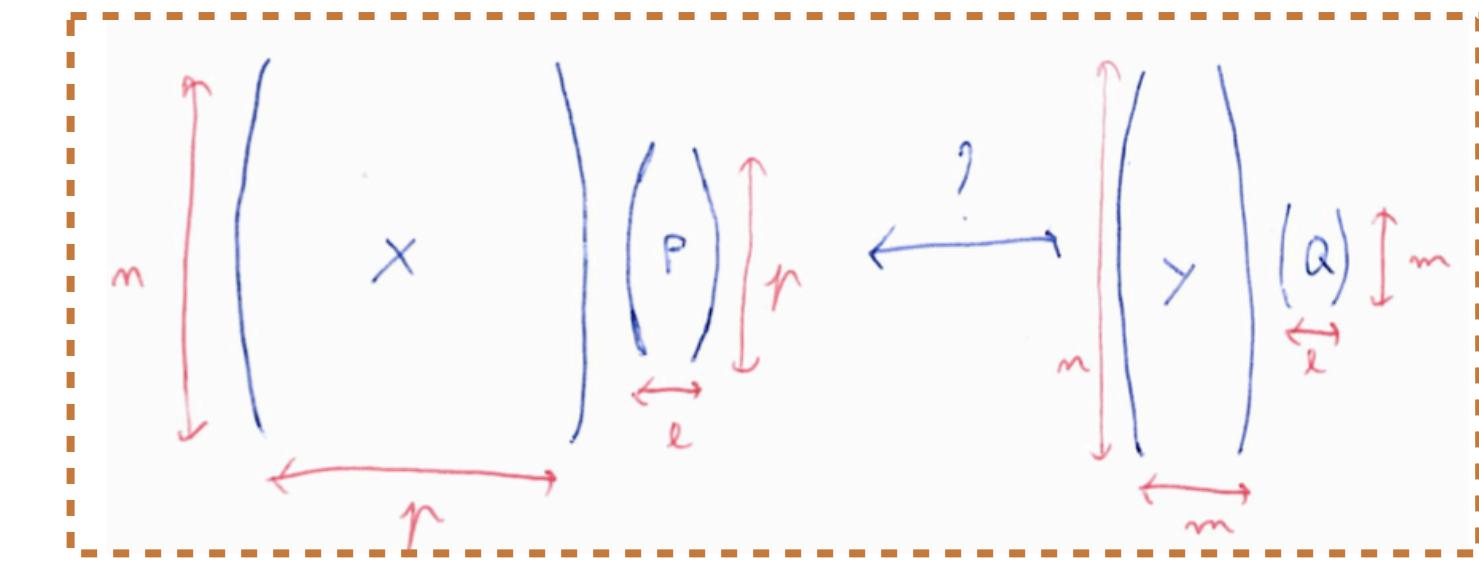
Alg. 1 Fonction $PLS1(\mathbf{X}, \mathbf{y}, l)$

```

1:  $\mathbf{X}^{(0)} \leftarrow \mathbf{X}$ 
2:  $\mathbf{w}^{(0)} \leftarrow \mathbf{X}' \mathbf{y} / |\mathbf{X}' \mathbf{y}|_2.$ 
3: for  $k = 0, \dots, l - 1$  do
4:    $\mathbf{t}^{(k)} \leftarrow \mathbf{X}_{\cdot}^{(k)} \mathbf{w}^{(k)}$ 
5:    $t_k \leftarrow \mathbf{t}^{(k)}' \mathbf{t}^{(k)}$ 
6:    $\mathbf{t}^{(k)} \leftarrow \mathbf{t}^{(k)} / t_k$ 
7:    $\mathbf{p}^{(k)} \leftarrow \mathbf{X}_{\cdot}^{(k)} \mathbf{t}^{(k)}$ 
8:    $q_k \leftarrow \mathbf{y}' \mathbf{t}^{(k)}$ 
9:   if  $q_k = 0$  then
10:     $l \leftarrow k$  et sort de la boucle for (toute la variabilité est capturée).
11:   end if
12:   if  $k < (l - 1)$  then
13:      $\mathbf{X}^{(k+1)} \leftarrow \mathbf{X}^{(k)} - t_k \mathbf{t}^{(k)} \mathbf{p}^{(k)'}'$ 
14:      $\mathbf{w}^{(k+1)} \leftarrow \mathbf{X}_{\cdot}^{(k+1)}' \mathbf{y} / |\mathbf{X}_{\cdot}^{(k+1)}' \mathbf{y}|_2$ 
15:   end if
16: end for

18:  $\mathbf{P}$  est la matrice composée des colonnes  $\mathbf{p}^{(0)}, \mathbf{p}^{(1)}, \dots, \mathbf{p}^{(l-1)}$ .

```



On suppose ici que $m = 1$, c'est à dire que les observations de y sont en dimension 1.

→ Pas d'estimation de Q .

Extension sparse-PLS

→ Sélection des variables les plus impactantes dans chaque vecteur de projection.

