

Semaine 4 : Apprentissage non-supervisé et réduction de dimension

2 : Analyse en composantes principales

Laurent Risser
Ingénieur de Recherche CNRS

Partie théorique basée sur le cours de Jean-Michel Loubes (Pr Univ. Toulouse 3)

0 : Préambule – réduire la dimension pour explorer les données

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
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

Comment faire un classement général entre les pays ???

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

0 : Préambule – réduire la dimension pour explorer les données

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
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

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 x vecteur :

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

0 : Préambule – réduire la dimension pour explorer les données

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
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

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 = $M \cdot v_1$

0 : Préambule – réduire la dimension pour explorer les données

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
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

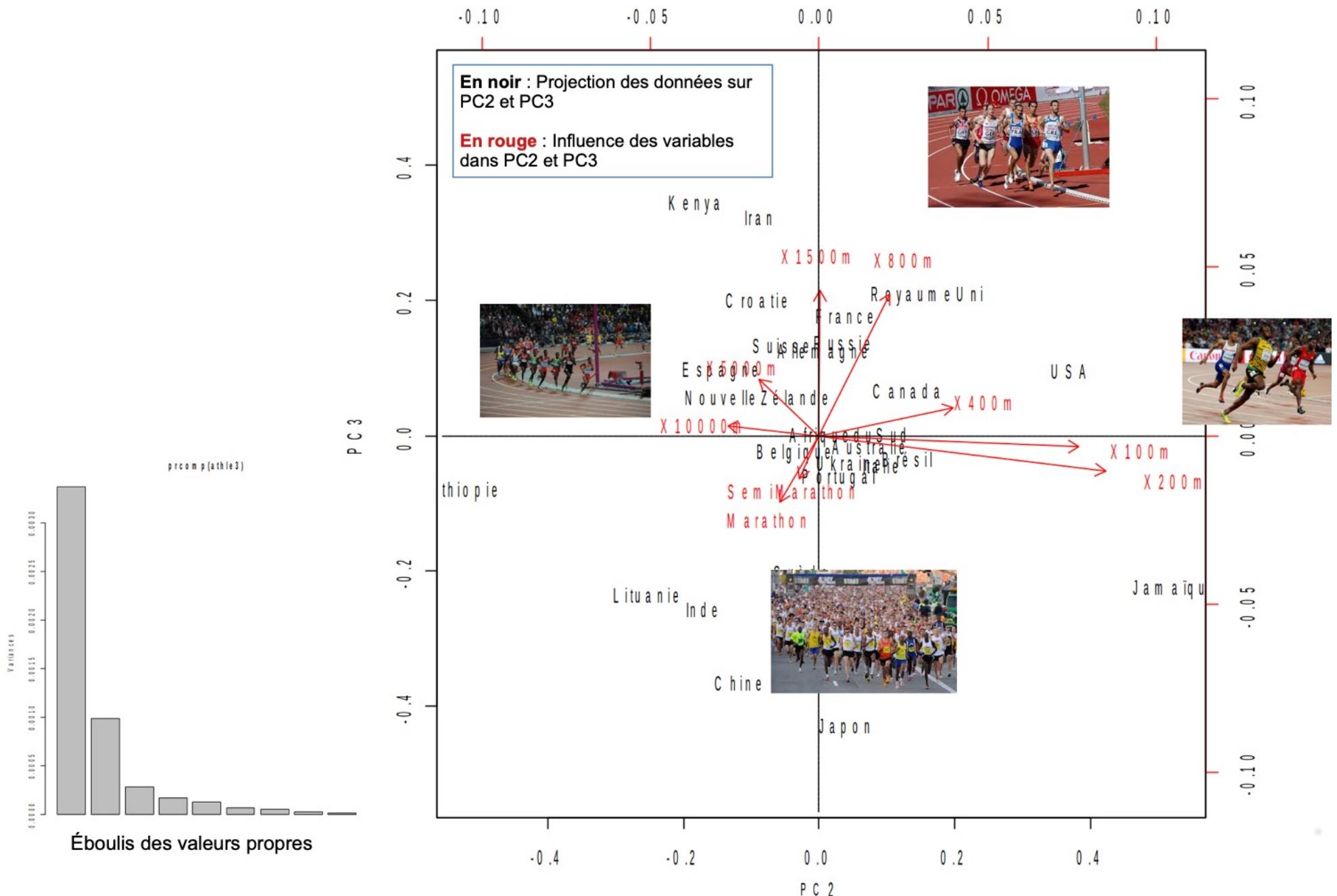
Une fois enlevée l'influence de v_1 , on cherche le vecteur (de norme 1) qui maximise la variabilité

- 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

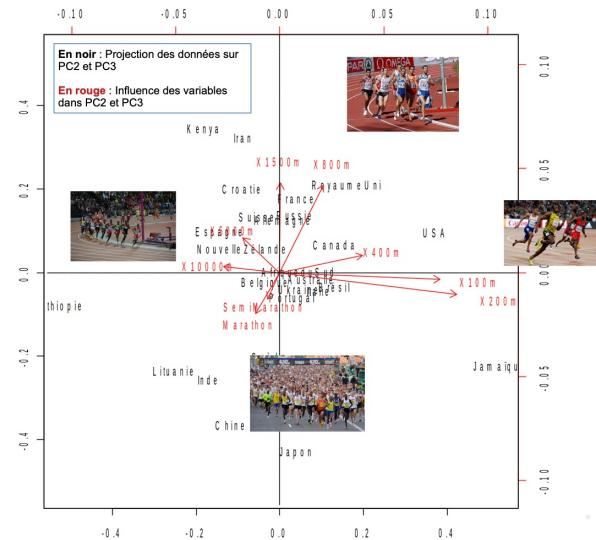
0 : Préambule – réduire la dimension pour explorer les données



0 : Préambule – réduire la dimension pour explorer les données

On passe d'une dimension 9 à une dimension 3 en gardant plus de 90% de la variabilité dans les données...

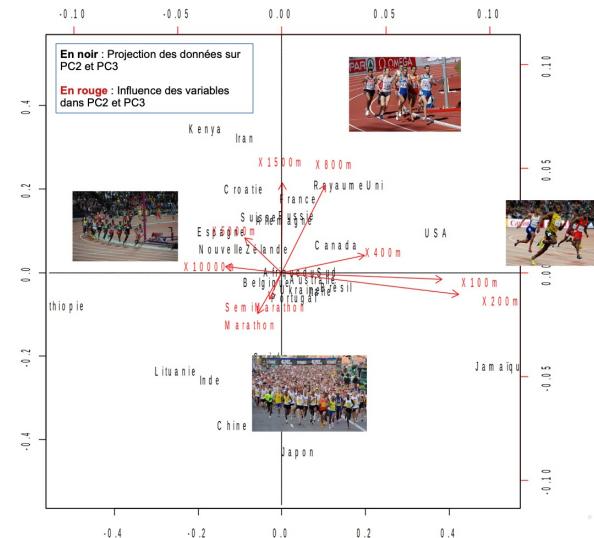
	100m	200m	400m	800m	1500m	5000m	10000m	SemiMarathon	Marathon
Australie	9.93	20.06	44.38	104.40	211.96	775.76	1649.73	3602	7671
Belgique	10.02	20.19	44.78	103.86	214.13	769.71	1612.30	3605	7640
Brésil	10.00	19.89	44.29	101.77	213.25	799.43	1648.12	3573	7565
RoyaumeUni	9.87	19.87	44.36	101.73	209.67	780.41	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



0 : Préambule – réduire la dimension pour explorer les données

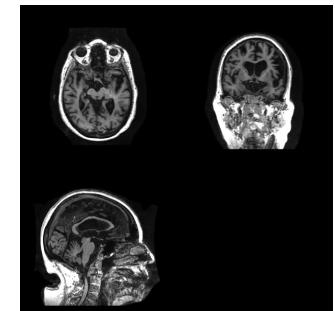
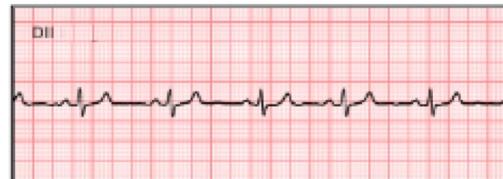
On passe d'une dimension 9 à une dimension 3 en gardant plus de 90% de la variabilité dans les données...

	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



On pourrait faire de même sur :

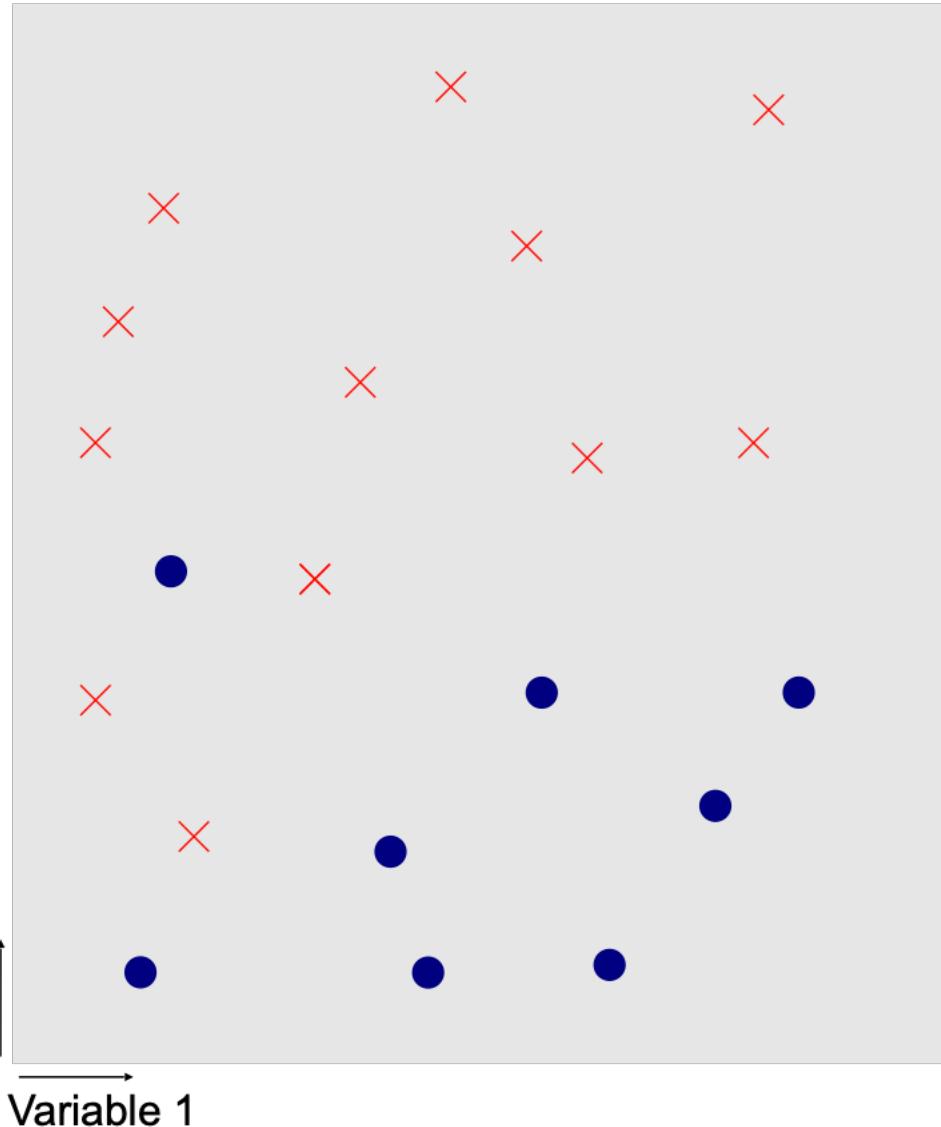
- Des tableaux de données avec $p >> 1$
- Des signaux
- Des images
- Du texte



Intérêt en apprentissage ?

0 : Préambule – intérêt en apprentissage

Exemple : apprentissage supervisé par arbre de décisions



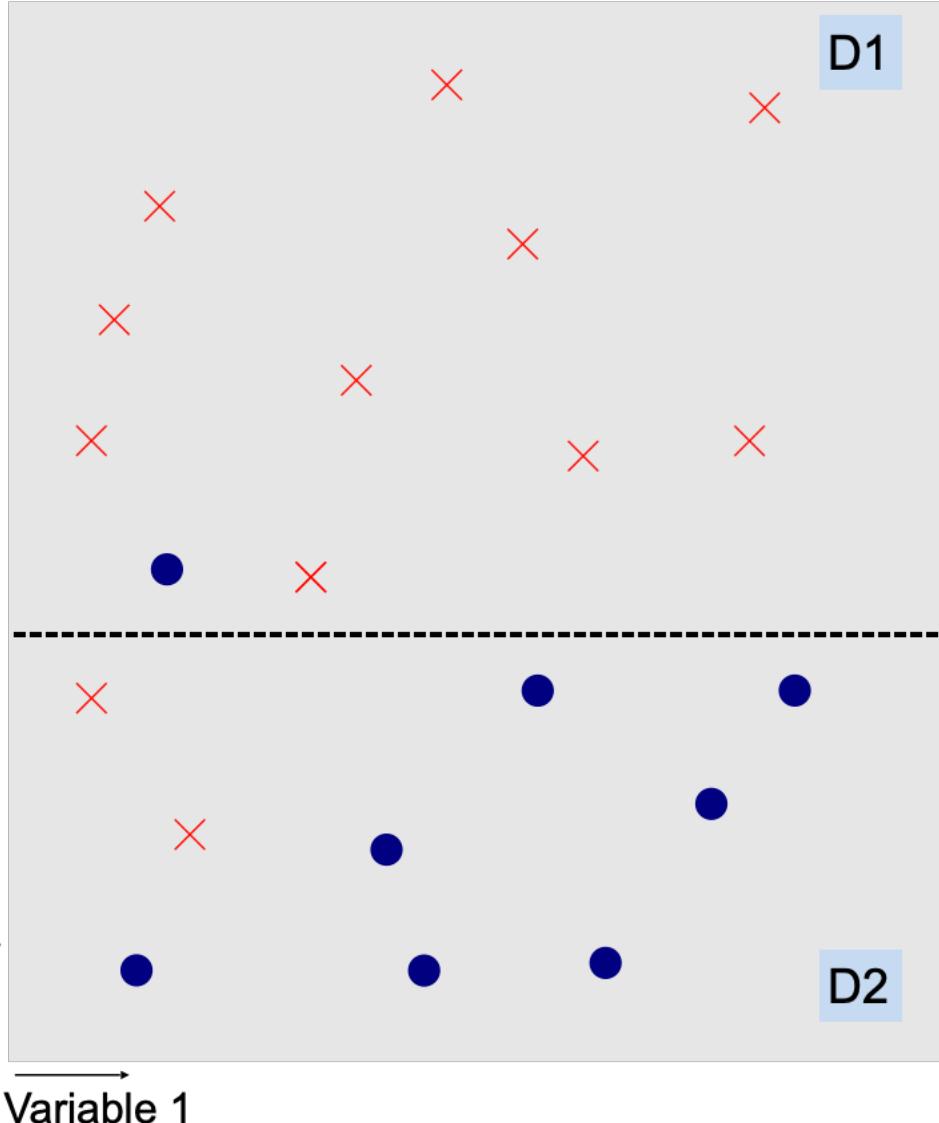
$\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$ les observations
(ici : x_i est la coordonnée du point en 2D)

y_1, y_2, \dots, y_N les labels
(ici : $\text{X} \rightarrow 1$ et $\text{O} \rightarrow -1$)

On coupe le domaine en *sous domaines* pour minimiser la variance dans chaque sous domaine (CART).

0 : Préambule – intérêt en apprentissage

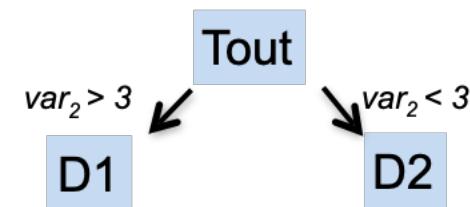
Exemple : apprentissage supervisé par arbre de décisions



$\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$ les observations
(ici : x_i est la coordonnée du point en 2D)

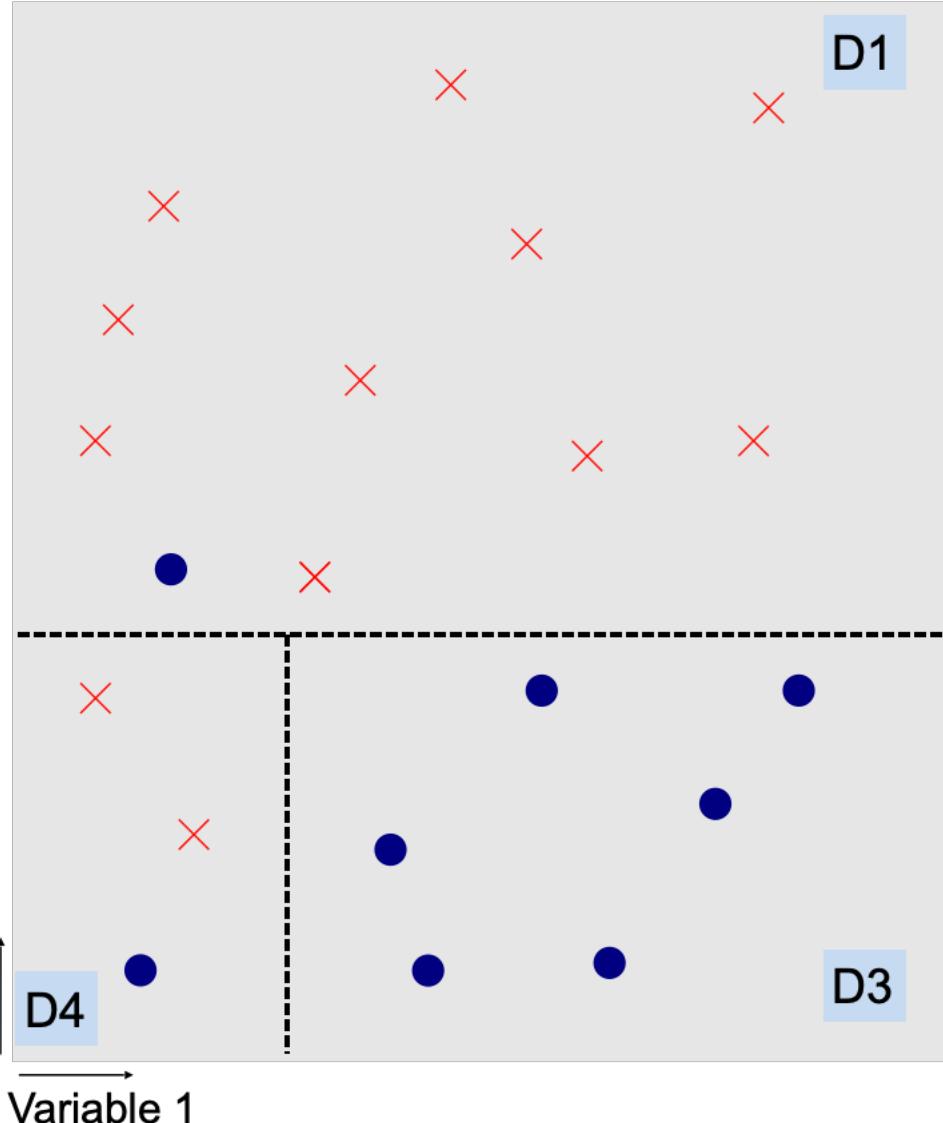
y_1, y_2, \dots, y_N les labels
(ici : $\text{X} \rightarrow 1$ et $\text{O} \rightarrow -1$)

On coupe le domaine en *sous domaines* pour minimiser la variance dans chaque sous domaine (CART).



0 : Préambule – intérêt en apprentissage

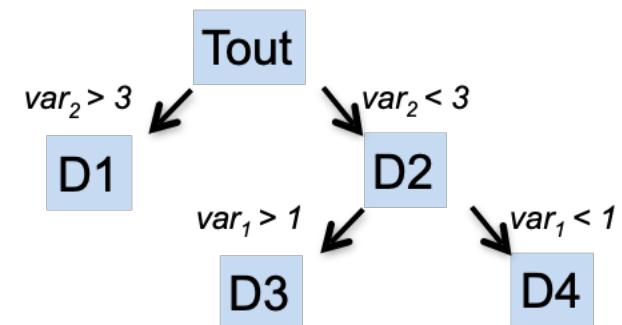
Exemple : apprentissage supervisé par arbre de décisions



$\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$ les observations
(ici : x_i est la coordonnée du point en 2D)

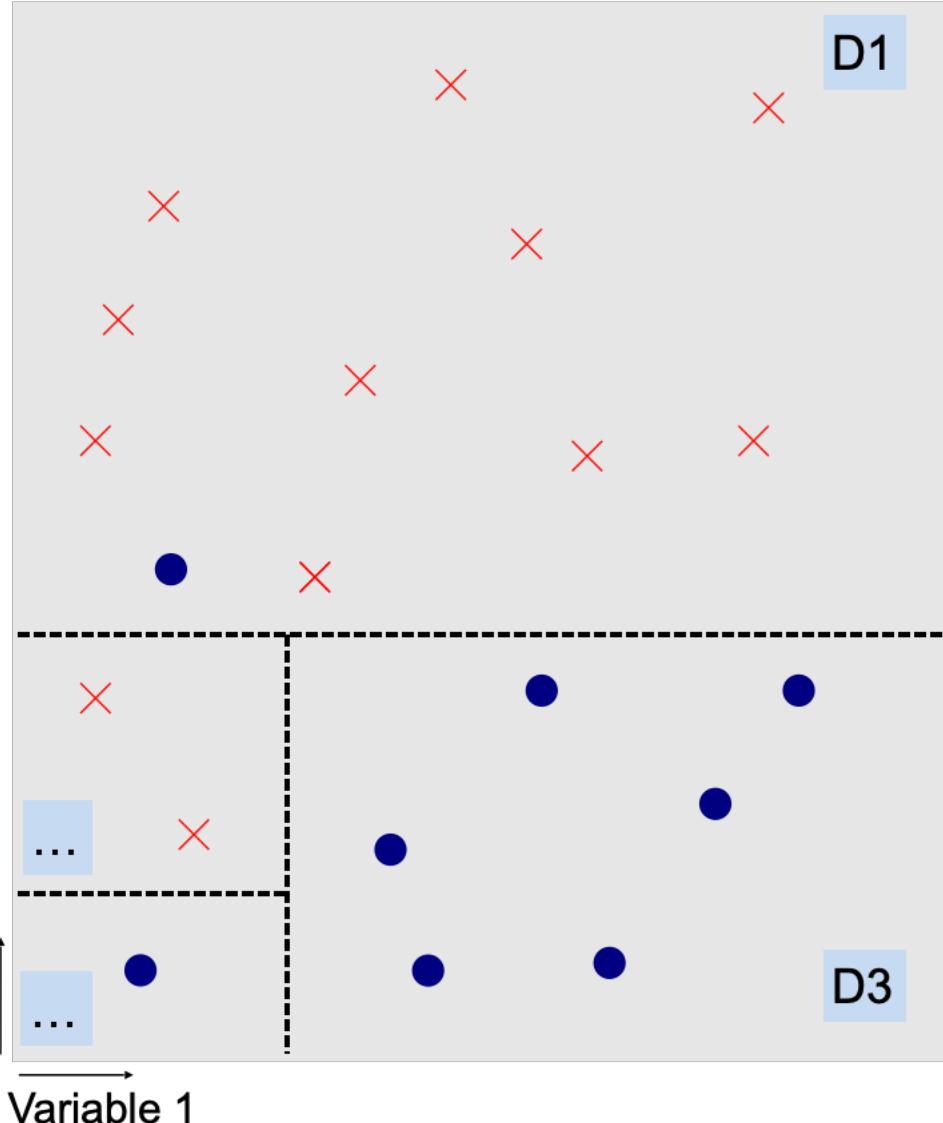
y_1, y_2, \dots, y_N les labels
(ici : $\text{X} \rightarrow 1$ et $\text{●} \rightarrow -1$)

On coupe le domaine en *sous domaines* pour minimiser la variance dans chaque sous domaine (CART).



0 : Préambule – intérêt en apprentissage

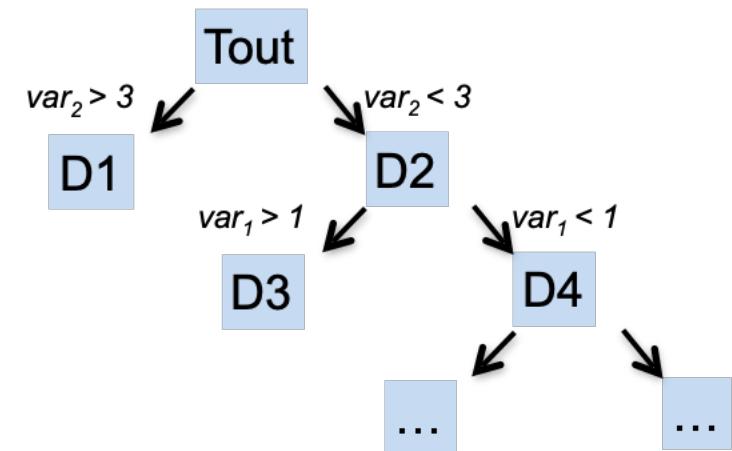
Exemple : apprentissage supervisé par arbre de décisions



$\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$ les observations
(ici : x_i est la coordonnée du point en 2D)

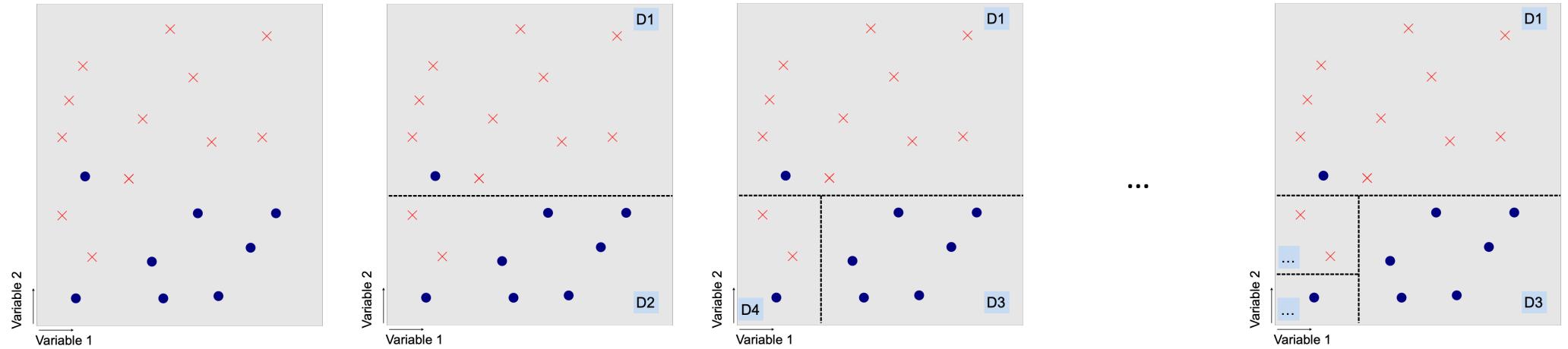
y_1, y_2, \dots, y_N les labels
(ici : $\text{X} \rightarrow 1$ et $\bullet \rightarrow -1$)

On coupe le domaine en *sous domaines* pour minimiser la variance dans chaque sous domaine (CART).



0 : Préambule – intérêt en apprentissage

Exemple : apprentissage supervisé par arbre de décisions



- A chaque itération recherche de la meilleure coupe sur p dimensions !!!
- Interprétabilité et stabilité du résultat de moins en moins bonne quand p augmente !

→ Criticité de réduire préalablement la dimension !

1 : Principes de l'Analyse en Composantes Principales

1.1 Rappels

- Une application P entre deux espaces E et F est une projection si et seulement si $P^2 = P$.
- Une application P entre deux espaces de Hilbert E et F est une projection orthogonale si et seulement si $\text{Im}(P) \perp \text{Ker}(P)$
- Projection sur un espace engendré par un vecteur $a \in \mathbb{R}^p$

$$P : \mathbb{R}^p \longrightarrow \mathbb{R}, \quad Px = a^T x$$

- Projection sur un espace engendré par une base orthonormale formée des vecteurs u_1, \dots, u_k . Soit $U = [u_1 \ u_2 \ \dots \ u_k]$,

$$P : \mathbb{R}^p \longrightarrow \mathbb{R}^k, \quad Px = UU^T x$$

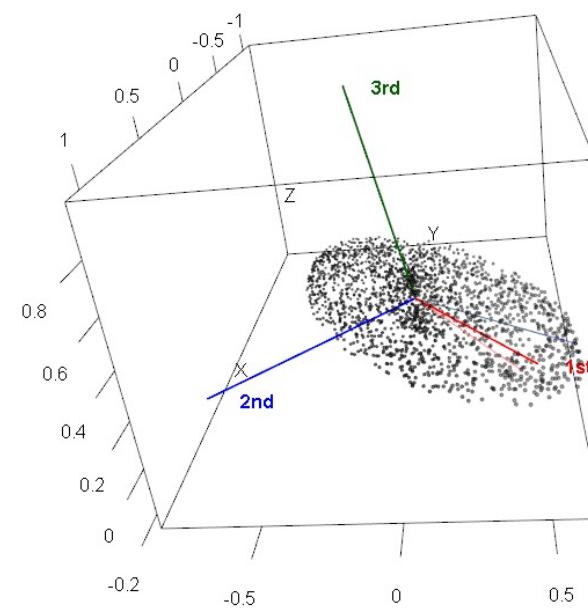
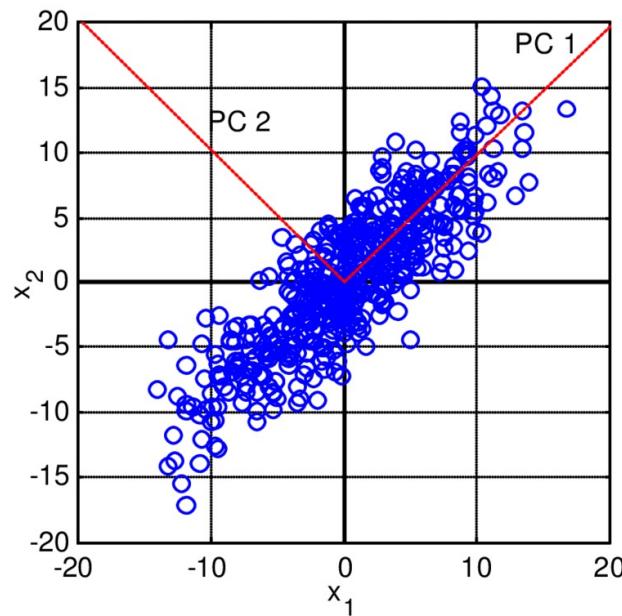
- La projection sur un espace affine donné par un point μ et un espace vectoriel S , $\tilde{S} = S + \mu$ s'écrit donc

$$Tx = \mu + P(x - \mu).$$

1 : Principes de l'Analyse en Composantes Principales

1.2 Principe général

L'objectif de l'ACP est de déterminer le *meilleur* 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.



1 : Principes de l'Analyse en Composantes Principales

1.3 Première interprétation

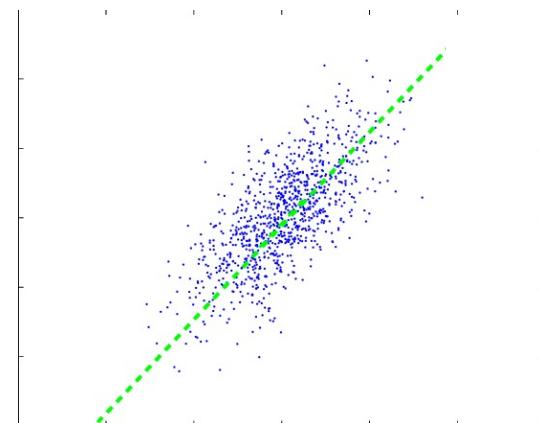
On dispose de n observations $X_i \in \mathbb{R}^p$.

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



1 : Principes de l'Analyse en Composantes Principales

1.3 Première interprétation

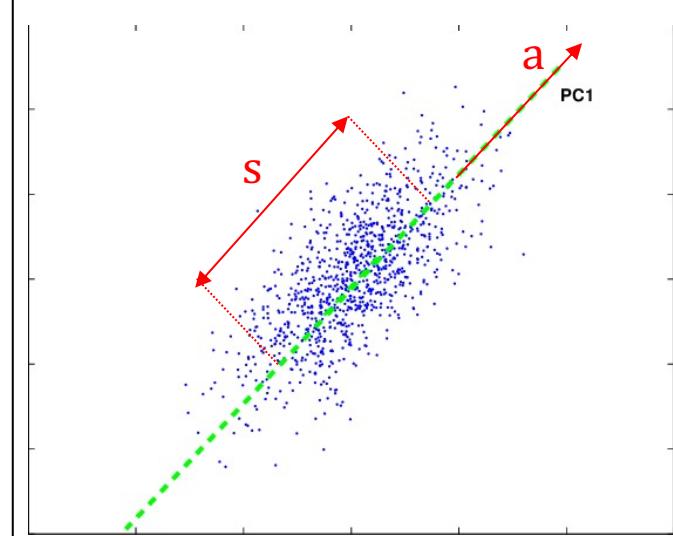
On dispose de n observations $X_i \in \mathbb{R}^p$.

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 .

1 : Principes de l'Analyse en Composantes Principales

1.4 Deuxième interprétation

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

Réduction de dimension. On cherche ici à trouver un espace de dimension $k << n$ sur lequel projeter les données et qui les interpole au mieux, au sens de la norme quadratique. On cherche donc à minimiser en S la perte

$$\min_S \sum_{i=1}^n \|X_i - P_S(X_i)\|^2.$$

avec $S = \{\mu + U_k z, z \in \mathbb{R}^k\}$, $\mu \in \mathbb{R}^p$ et $U \in \mathbb{R}^{p \times k}$

Représentation
latente de x_i

$$X_i = \begin{pmatrix} x_i^1 \\ x_i^2 \\ \vdots \\ x_i^p \end{pmatrix}$$

$$P_S(X_i) = \begin{pmatrix} \mu^1 \\ \mu^2 \\ \vdots \\ \mu^p \end{pmatrix} + \begin{pmatrix} U^{1,1} & \dots & U^{1,k} \\ U^{2,1} & \dots & U^{2,k} \\ \vdots & & \vdots \\ U^{p,1} & \dots & U^{p,k} \end{pmatrix} \begin{pmatrix} \delta_i^1 \\ \vdots \\ \delta_i^k \end{pmatrix}$$

1 : Principes de l'Analyse en Composantes Principales

1.4 Deuxième interprétation

On résout alors $\min_S \sum_{i=1}^n \|X_i - \mu - U_k z_i\|^2 = \min_{\mu, z, U_k} \sum_{i=1}^n \|X_i - \mu - U_k z_i\|^2.$

En minimisant pour U_l fixé on obtient que le minimum est atteint pour

$$\mu = \bar{x} = \frac{1}{n} X_i, \quad z_i = U_k^T (x_i - \bar{x}).$$

Ainsi la minimisation en U_k s'écrit

$$\begin{aligned} \min_{\mu, z, U_k} \sum_{i=1}^n \|X_i - \mu - U_k z_i\|^2 &= \min_{U_k} \sum_{i=1}^n \|(x_i - \bar{x}) - U_k U_k^T (x_i - \bar{x})\|^2 \\ &= \min_{U_k} \sum_{i=1}^n \|\tilde{X}_i - U_k U_k^T \tilde{X}_i\|^2 \end{aligned}$$

1 : Principes de l'Analyse en Composantes Principales

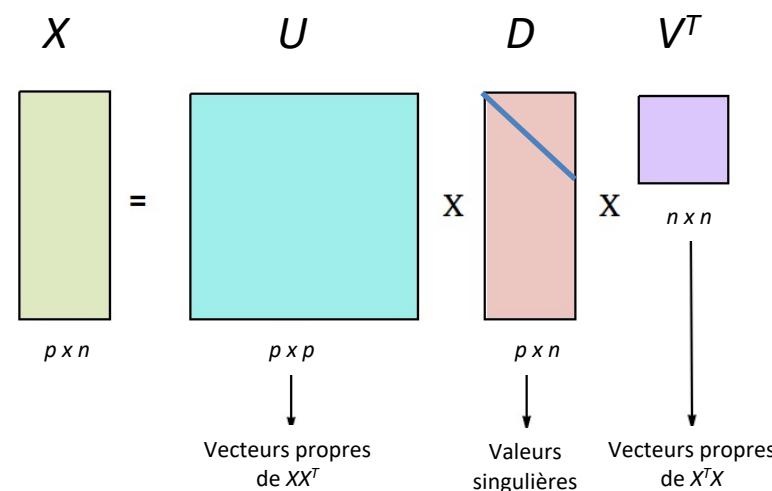
1.4 Deuxième interprétation

Résolution de $\min_{U_l} \sum_{i=1}^n \|\tilde{X}_i - U_k U_k^T \tilde{X}_i\|^2$

Pour cela, on fait la décomposition SVD de la matrice $X = \tilde{X}$ et on obtient

$$X = UDV^T, \quad U \in \mathbb{R}^{p \times p}, D \in \mathbb{R}^{p \times n}, V \in \mathbb{R}^{n \times n},$$

avec U et V des matrices orthogonales et D une matrice diagonale. On choisit alors pour U_k les k premières colonnes de U .



Nous allons voir ça plus en détails dans la prochaine partie !

1.4 Deuxième interprétation

On remarque que

$$S = XX^\top = UDV^\top VD^\top U^\top = UD^2U^\top$$

Ainsi trouver les vecteurs propres en utilisant la décomposition SVD de X ou trouver les vecteurs propres de S revient au même.

Ce point de vue est le plus général et permet de généraliser l'ACP dans des espaces plus complexes que \mathbb{R}^p . Trouver la première direction sur laquelle projeter revient à trouver la courbe géodésique qui interpole au mieux les données.

2 : Etude théorique de l'ACP

Considérons p variables statistiques réelles X^j ($j = 1, \dots, p$) observées sur n individus (numérotés $i = 1, \dots, n$) affectés des poids w_i :

$$\forall i = 1, \dots, n : w_i > 0 \text{ et } \sum_{i=1}^n w_i = 1 ;$$

$\forall i = 1, \dots, n : x_i^j = X^j(i)$, valeur prise par la variable X^j sur le $i^{ème}$ individu.

Si tous les individus ont le même poids, $w_i = \frac{1}{n}$ pour tout i .

Ces mesures sont regroupées dans une matrice \mathbf{X} d'ordre $(n \times p)$. On considère qu'on a préalablement centré les données.

	X^1	\dots	X^j	\dots	X^p
1	x_1^1	\dots	x_1^j	\dots	x_1^p
\vdots	\vdots		\vdots		\vdots
i	x_i^1	\dots	x_i^j	\dots	x_i^p
\vdots	\vdots		\vdots		\vdots
n	x_n^1	\dots	x_n^j	\dots	x_n^p

2 : Etude théorique de l'ACP

	X^1	\cdots	X^j	\cdots	X^p
1	x_1^1	\cdots	x_1^j	\cdots	x_1^p
\vdots	\vdots		\vdots		\vdots
i	x_i^1	\cdots	x_i^j	\cdots	x_i^p
\vdots	\vdots		\vdots		\vdots
n	x_n^1	\cdots	x_n^j	\cdots	x_n^p

- À chaque individu i est associé le vecteur \mathbf{x}_i contenant la i -ème ligne de \mathbf{X} mise en colonne. C'est un élément d'un espace vectoriel E de dimension p isomorphe à \mathbb{R}^p . E est alors appellé *espace des individus*.
- À chaque variable X^j est associé le vecteur \mathbf{x}^j contenant la j -ème colonne *centrée* (la moyenne de la colonne est retranchée à toute la colonne) de \mathbf{X} . C'est un élément d'un espace vectoriel F de dimension n avec pour métrique la matrice \mathbf{D} diagonale des *poids* lui conférant une structure d'espace euclidien et le rendant isomorphe à $(\mathbb{R}^n, \mathbf{D})$ avec $\mathbf{D} = \text{diag}(w_1, \dots, w_n)$; F est alors appellé *espace des variables*.

2 : Etude théorique de l'ACP

La matrice de covariance empirique est alors

$$S = \bar{\mathbf{X}}' \mathbf{D} \bar{\mathbf{X}} [= \frac{1}{n} \sum_{i=1}^n X_i X_i^T \quad \text{si } D = \frac{1}{n} I_n]$$

Définissons la matrice de covariance

$$\Sigma = \text{Var}(\mathbf{X}).$$

Cette matrice est une matrice symétrique et positive de dimension $p \times p$.
Ainsi il existe une matrice diagonale

$$\Lambda = \text{Diag}(\lambda_1, \dots, \lambda_p)$$

où λ_i sont les valeurs propres de Σ rangées en ordre décroissant $\lambda_1 \geq \lambda_2 \geq \lambda_p$,
les et des vecteurs orthonormés $\mathbf{u} = (\mathbf{u}^1, \dots, \mathbf{u}^p)$ vérifiant

$$\|\mathbf{u}^1\| = 1, \quad \mathbf{u}^j \mathbf{u}^{kT} = 0, \text{ pour } j \neq k$$

tels que

$$\Sigma = \mathbf{u} \Lambda \mathbf{u}^T$$

soit la décomposition spectrale de la matrice de covariance Σ . Soit $\mu = \mathbb{E}(X)$.
La variable aléatoire $\mathbf{v}^k = \mathbf{u}^{kT} (X - \mu)$ est appelée k -ième direction principale
de \mathbf{X} .

2 : Etude théorique de l'ACP

Theorem 5 Soit $\mathbf{X} \in \mathbb{R}^p$ de matrice de covariance Σ . Alors

$$\mathbf{u}^1 \in \arg \max_{a, : \|a\|=1} \text{Var}(a^T \mathbf{X}).$$

$$\mathbf{u}^k \in \arg \max_{\{a, : \|a\|=1, a \perp \mathbf{u}^j, j=1, \dots, k-1\}} \text{Var}(a^T \mathbf{X}).$$

2 : Etude théorique de l'ACP

Theorem 5 Soit $\mathbf{X} \in \mathbb{R}^p$ de matrice de covariance Σ . Alors

$$\mathbf{u}^1 \in \arg \max_{a, : \|a\|=1} \text{Var}(a^T \mathbf{X}).$$

$$\mathbf{u}^k \in \arg \max_{\{a, : \|a\|=1, a \perp \mathbf{u}^j, j=1, \dots, k-1\}} \text{Var}(a^T \mathbf{X}).$$

Démonstration. — En utilisant la décomposition spectrale de la matrice $\Sigma = \mathbf{u}\Lambda\mathbf{u}^T$, on a

$$\text{Var}(a^T \mathbf{X}) = \sum_{j=1}^p \lambda_j (a^T \mathbf{u}^j) (\mathbf{u}^{jT} a) = \sum_{j=1}^p \lambda_j v_j^2$$

avec $v_j = a^T \mathbf{u}^j$ la projection du vecteur . On a donc

$$\text{Var}(a^T \mathbf{X}) = \sum_{j=1}^p \lambda_j v_j^2 \leq \lambda_1 \sum_{j=1}^p v_j^2 \quad [\text{Definition des } \lambda'_i s] = \lambda_1 \|v\|^2 = \lambda_1 \|a\|^2$$

Ainsi la variance est maximisée pour $a = \mathbf{u}^1$ la première direction principale.

Définissons alors Soit $a \in \mathcal{A}_1 = \{a \in \mathbb{R}^n, \|a\| = 1, a \perp \mathbf{u}^1\}$, on a

$$\text{Var}(a^T \mathbf{X}) = \sum_{j=2}^p \lambda_j v_j^2 \quad [\text{car } c_1 = 0] \leq \sum_{j=2}^p \lambda_j v_j^2 \leq \lambda_2 \|a\|^2.$$

Ainsi la variance est maximisée sur l'ensemble \mathcal{A}_1 par le choix de $a = \mathbf{u}^2$, qui correspond à la seconde direction principale.

On procède de la même façon pour les autres directions principales.

3 : Choix de la dimension de l'ACP

La variance totale de X est définie par

$$\mathbb{E}\|X - \mu\|^2 = \mathbb{E}(X - \mu)^T(X - \mu) = \mathbb{E}(X - \mu)^T\mathbf{u}\mathbf{u}^T(X - \mu)$$

Donc, en utilisant la définition des directions principales

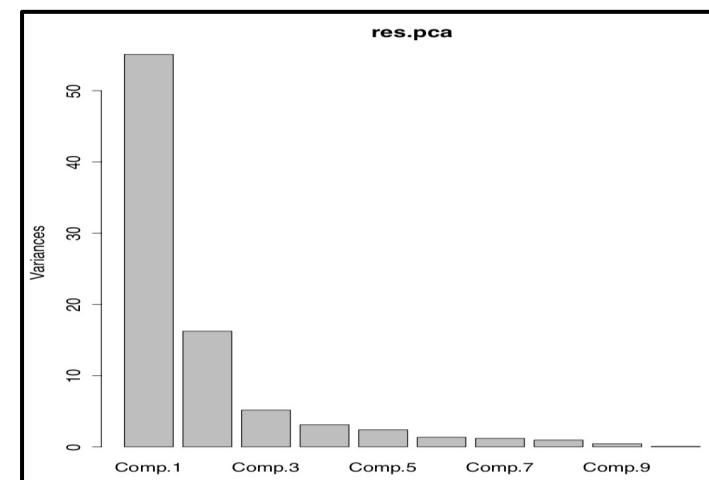
$$\mathbb{E}\|X - \mu\|^2 = \sum_{i=1}^p \mathbf{v}^{k^2} = \sum_{i=1}^p \lambda_i = \text{Tr}(\Sigma).$$

En projetant les données sur les k premières directions principales, on peut donc comparer la variance expliquée par cette projection par rapport à la variation totale des données

$$\frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^p \lambda_i}$$

La variance expliquée par une direction principale \mathbf{v}^k vaut

$$\frac{\lambda_k}{\sum_{i=1}^p \lambda_i}$$



4 : ACP fonctionnelle

Comment étendre maintenant le principe de l'ACP à des fonctions ?

Par exemple : Courbes de températures dans différentes villes Canadiennes

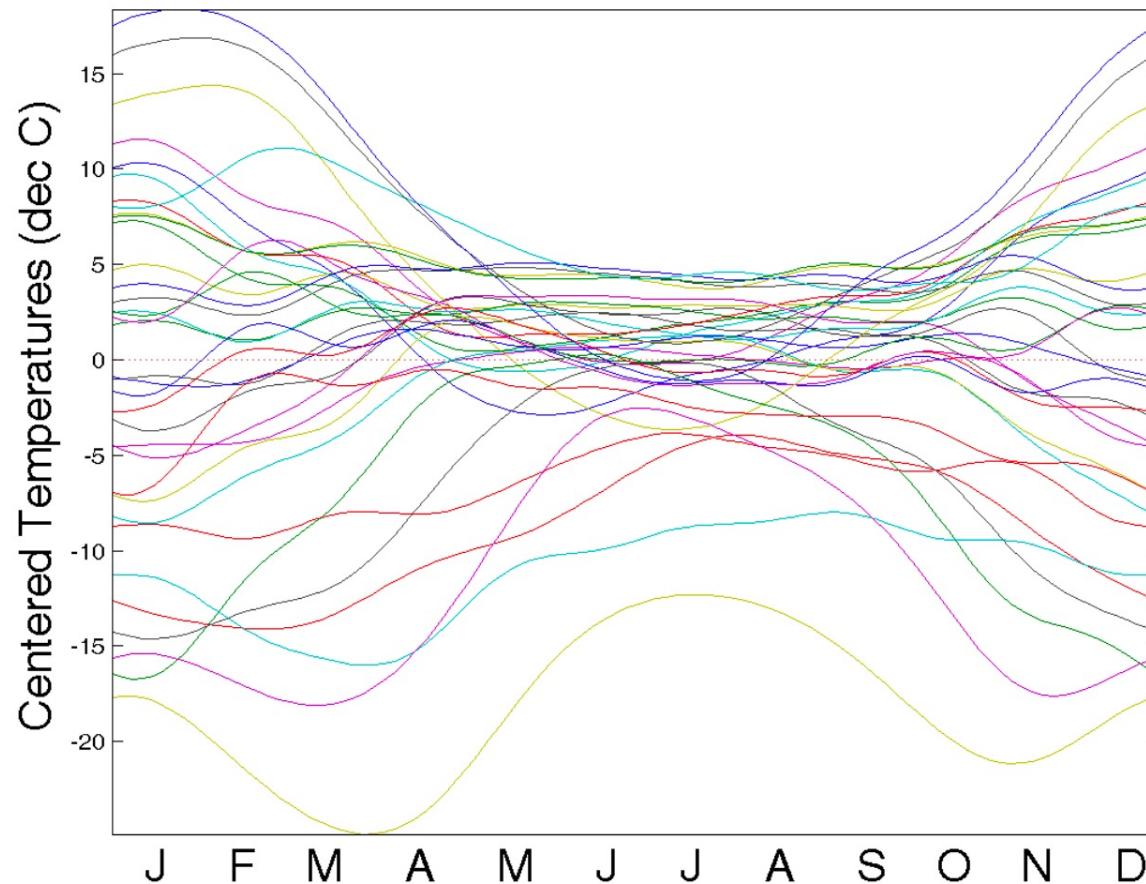


Illustration : cours McGill

4 : ACP fonctionnelle

On suppose qu'on observe n fonctions f_1, \dots, f_n centrées observées en p points t_1, \dots, t_p . L'ACP se généralise de la manière suivante : on recherche la première composante fonctionnelle $t \mapsto \Phi_1(t)$ qui maximise la variance

$$\Phi \mapsto \sum_{i=1}^n \left\| \int \Phi(t) f_i(t) dt \right\|^2$$

sous la contrainte

$$\int \Phi^2(t) dt = \|\Phi\|^2 = 1.$$

La seconde composante principale Φ_2 est définie comme réalisant le maximum de la variance

$$\sum_{i=1}^n \left\| \int \Phi_2(t) f_i(t) dt \right\|^2$$

sous les contraintes $\|\Phi_2\|^2 = 1$ et

$$\int \Phi_1(t) \Phi_2(t) dt = 0.$$

Les autres composantes principales sont calculées de la même façon en itérant le processus.

Pour trouver pratiquement les composantes principales en statistique multivariée, nous avons vu qu'il suffit de trouver les vecteurs propres de la matrice de covariance. Dans le cadre fonctionnel, la covariance est un opérateur défini par

$$\sigma(s, t) = \mathbb{E}(f(s)f(t)).$$

La covariance empirique s'écrit donc

$$v(s, t) = \frac{1}{n} \sum_{i=1}^n f_i(s)f_i(t).$$

Chercher les vecteurs propres revient à résoudre l'équation intégrale

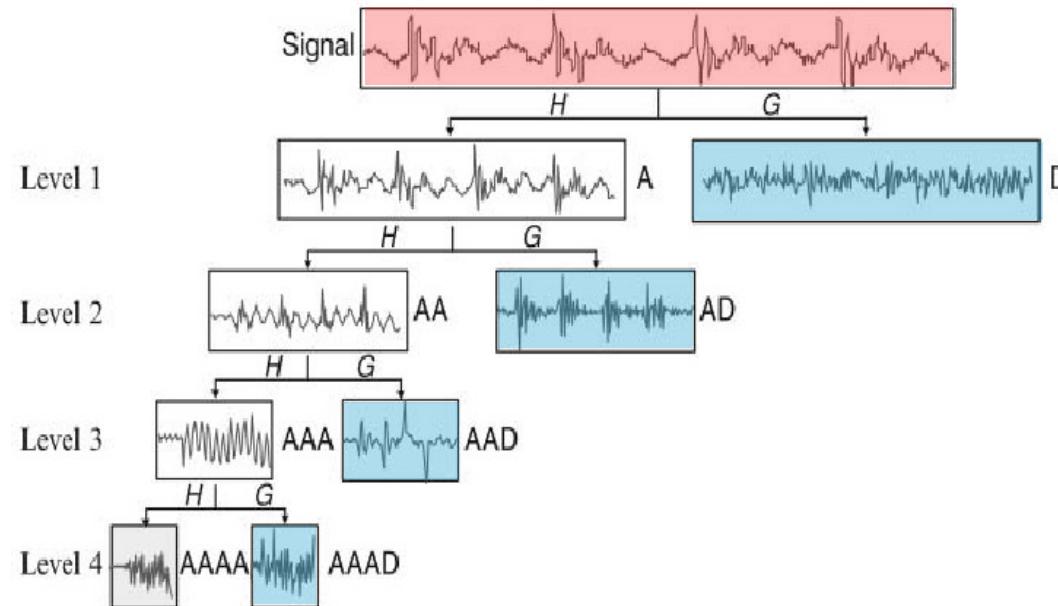
$$\int v(s, t)\Phi(t)dt = \lambda\Phi(s). \quad (11)$$

λ sera la valeur propre et $\Phi(t)$ sera la fonction vecteur propre. Pour résoudre cette équation intégrale fonctionnelle, on utilisera un décomposition des fonctions dans une base.

4 : ACP fonctionnelle

Exemple de décomposition d'un signal sur une base d'ondelettes.

Pour la fonction $f_i(t)$ (en rouge), la transformation en ondelette calcule les coefficients a_j^i (en bleu), chacun étant associé à un **élément de base** ρ_j qui représente l'impact semi-local de certaines fréquences



4 : ACP fonctionnelle

On suppose que chaque fonction f et Φ_j se décomposent (ou sont approchées) sur une base φ_j , $j = 1, \dots, m$ de la manière suivante

$$\begin{aligned}f_i(t) &= \sum_{l=1}^m a_l^i \varphi_l(t) = \mathbf{a}^{i'} \varphi(t) \\f(t) &= (f_1(t) \dots f_n(t)) = A\varphi(t) \\\Phi_j(t) &= \sum_{l=1}^m b_l^j \varphi_l(t) = \mathbf{b}^{j'} \varphi(t),\end{aligned}$$

en posant $\mathbf{a}' = (a_1 \dots a_m)'$, A la matrice formée par les \mathbf{a}^j et $\mathbf{b}' = (b_1 \dots b_m)'$. La covariance empirique est donc approchée par

$$v(s, t) = \frac{1}{n} \varphi'(t) A' A \varphi(t).$$

Ainsi (11) pourra s'écrire comme

$$\frac{1}{n} \varphi'(s) A' A \int \varphi(t) \varphi'(t) dt \mathbf{b}^j = \lambda \varphi'(s) \mathbf{b}^j,$$

soit en posant $G = \int \varphi(t) \varphi'(t) dt$

$$\frac{1}{n} \varphi'(s) A' A G \mathbf{b}^j = \lambda \varphi'(s) \mathbf{b}^j, \quad \forall s \in ?.$$

Ainsi nous obtenons les vecteurs propres en résolvant en \mathbf{b}^j , sous la contrainte $\|\Phi_j\|^2 = 1$, soit

$$\mathbf{b}^{j'} G \mathbf{b}^j = 1$$

l'équation

$$\left(\frac{1}{n} A' A G - \lambda I_p \right) \mathbf{b}^j = 0.$$

G est une matrice symétrique positive donc on peut définir $G^{\frac{1}{2}}$. Dès lors posons

$$u_j = G^{\frac{1}{2}} \mathbf{b}^j.$$

La contrainte s'écrit alors

$$u_j' u_j = 1$$

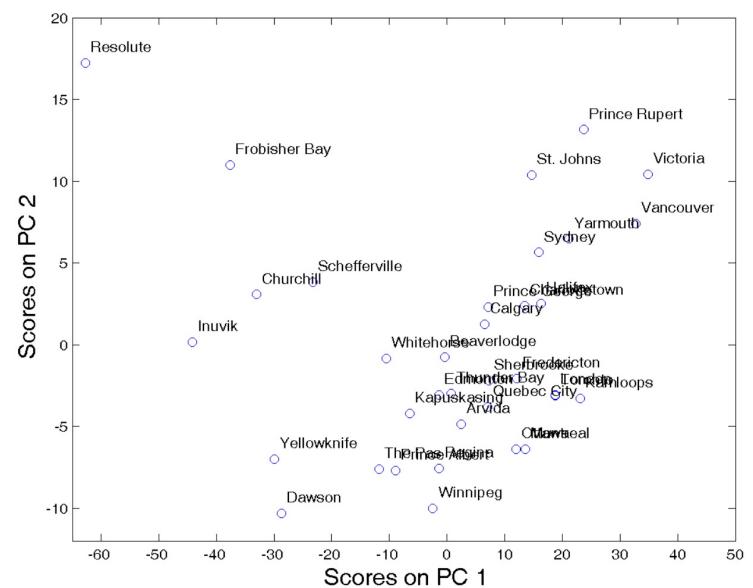
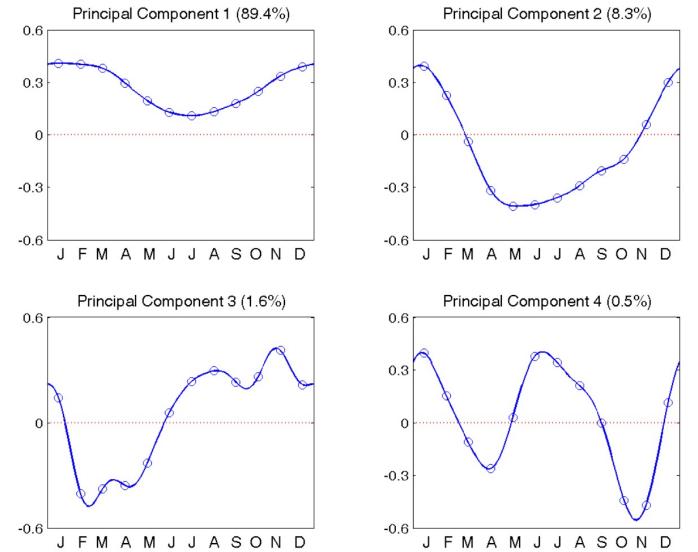
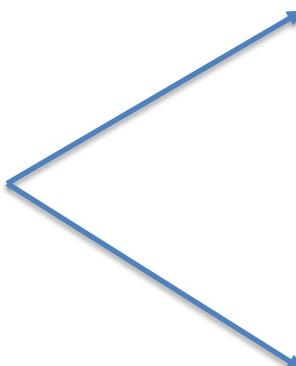
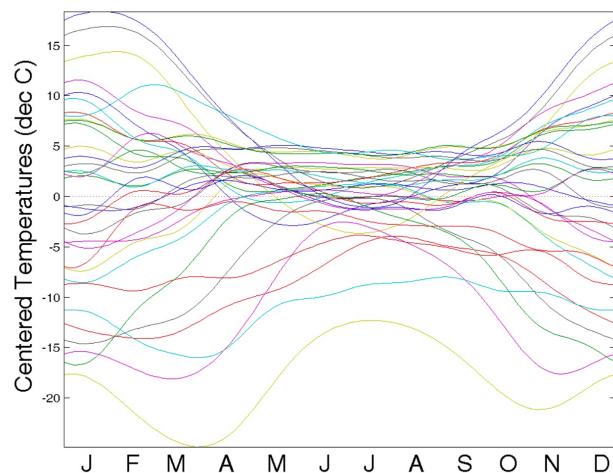
et l'équation s'écrit

$$\left(\frac{1}{n} G^{\frac{1}{2}} A' A G^{\frac{1}{2}} - \lambda I_p \right) u_j = 0.$$

Cette équation matricielle est une équation usuelle de recherche de vecteur propre qui peut être résolue par les méthodes numériques traditionnelles.

4 : ACP fonctionnelle

Exemple des courbes de températures dans différentes villes



cours McGill

5 : ACP à noyaux

L'ACP à noyaux repose sur l'astuce du noyau usuelle : envoyer les données dans un espace de plus grande dimension dans lequel la variabilité des données peut être capturée par des méthodes linéaires comme l'ACP. Pour cela on considère qu'on observe x_1, \dots, x_n chacun dans \mathbb{R}^p . On note $X = [x_1 \dots x_n]$ la matrice des observations en colonne. Soit l'application Φ qui va augmenter la dimension des données dans l'objectif d'obtenir une représentation plus simple dans ce nouvel espace

$$\Phi : \mathbb{R}^p \longrightarrow F$$

5 : ACP à noyaux

$$\Phi : \mathbb{R}^p \longrightarrow F$$

On suppose que les observations dans ce nouvel espace sont centrées

$$\sum_{i=1}^n \Phi(x_i) = 0.$$

Soit C la matrice de covariance empirique des nouvelles données

$$C = \frac{1}{n} \sum_{i=1}^n \Phi(x_i) \Phi(x_i)^T.$$

La matrice des produits scalaires définit la matrice du noyau

$$K := (K_{ij})_{ij} = (\langle \Phi(x_i), \Phi(x_j) \rangle)_{ij}.$$

On cherche à appliquer l'ACP sur ces données transformées, donc ceci revient à rechercher les valeurs λ et vecteurs propres orthonormés $v \in F - \{0\}$ tels que

$$Cv = \lambda v.$$

5 : ACP à noyaux

On cherche à appliquer l'ACP sur ces données transformées, donc ceci revient à rechercher les valeurs λ et vecteurs propres orthonormés $v \in F - \{0\}$ tels que

$$Cv = \lambda v.$$

Or

$$Cv = \frac{1}{n} \sum_{i=1}^n \langle \Phi(x_i), v \rangle \Phi(x_i)$$

donc les solutions appartiennent à l'espace engendré par les Vect $\{\Phi(x_1), \dots, \Phi(x_n)\}$, c'est-à-dire qu'il existe des coefficients $\alpha_1, \dots, \alpha_n$ tels que

$$v = \sum_{i=1}^n \alpha_i \Phi(x_i).$$

Ces solutions reviennent à chercher

$$\lambda \langle \Phi(x_i), v \rangle = \langle \Phi(x_i), Cv \rangle, \quad \forall i = 1, \dots, n.$$

5 : ACP à noyaux

Ces deux dernières égalités donnent

$$\lambda \sum_{i=1}^n \alpha_i \langle \Phi(x_j), \Phi(x_i) \rangle = \frac{1}{n} \sum_{i=1}^n \alpha_i \langle \Phi(x_i), \sum_{j=1}^n \Phi(x_j) \rangle \langle \Phi(x_j), \Phi(x_i) \rangle,$$

soit en notant $\alpha = (\alpha_1, \dots, \alpha_n)^T$

$$n\lambda K\alpha = K^2\alpha.$$

Or K est une matrice symétrique définie positive donc $K\alpha$ engendre tout l'espace donc cela revient à chercher α et λ vérifiant

$$K\alpha = n\lambda\alpha.$$

La condition d'orthonormalité des composantes principales entraîne la contrainte suivante sur les α^k 's composantes du vecteur v_k

$$\begin{aligned} 1 &= \langle v_k, v_k \rangle \\ &= \sum_{i,j=1}^n \alpha_i^k \alpha_j^k \langle \Phi(x_i), \Phi(x_j) \rangle \\ &= \sum_{i,j=1}^n \alpha_i^k \alpha_j^k K_{ij} = \langle \alpha^k, K\alpha^k \rangle = \lambda_k \langle \alpha_k, \alpha_k \rangle \end{aligned}$$

5 : ACP à noyaux

PCA dans un espace à noyau

- *Calcul des points centrés*

$$\Phi(x_i) := \Phi(x_i) - \frac{1}{n} \sum_{i=1}^n \Phi(x_i)$$

- *Calcul de la matrice de noyau*

$$K_{ij} = \langle \Phi(x_i), \Phi(x_j) \rangle .$$

- *Calcul de la décomposition en vecteurs propres*

$$K = UDU^T, \quad U = [U_1, U_2 \dots U_p]$$

- Pour un choix de dimension k , on construit la matrice V_k ,

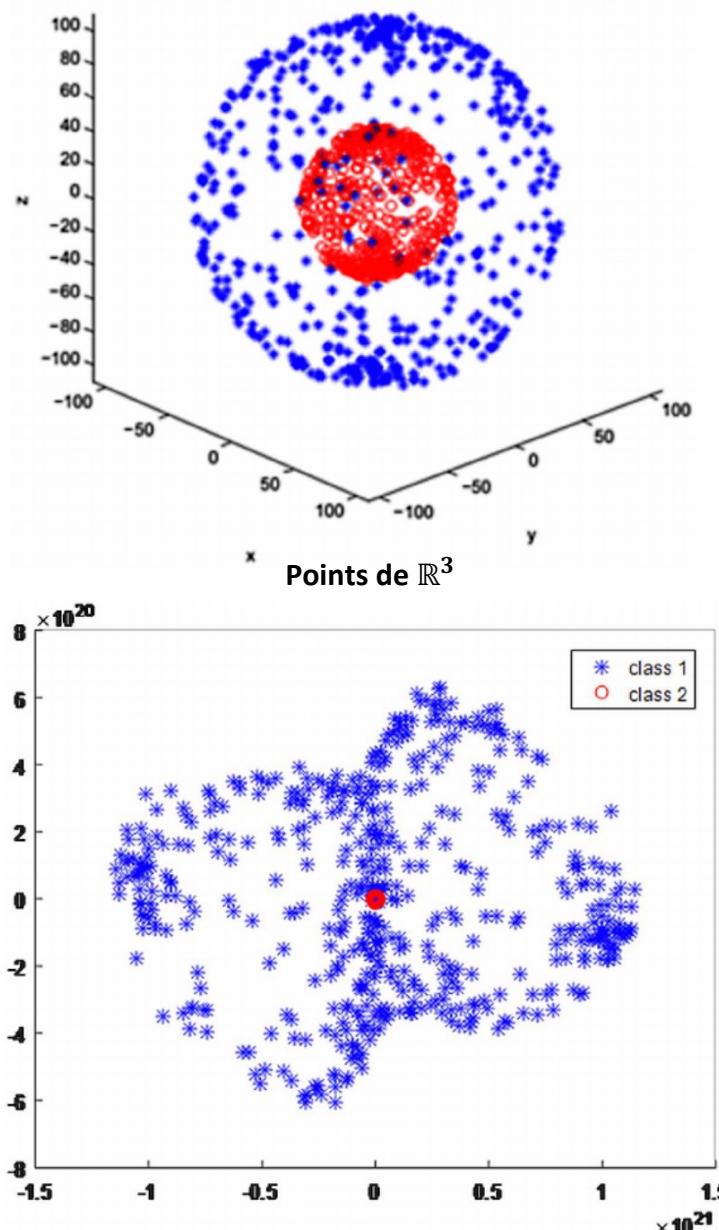
$$V_k = [U_1/\sqrt{\lambda_1} \dots U_k/\sqrt{\lambda_k}],$$

avec λ_j , $j = 1, \dots, k$ les valeurs propres de la matrice K .

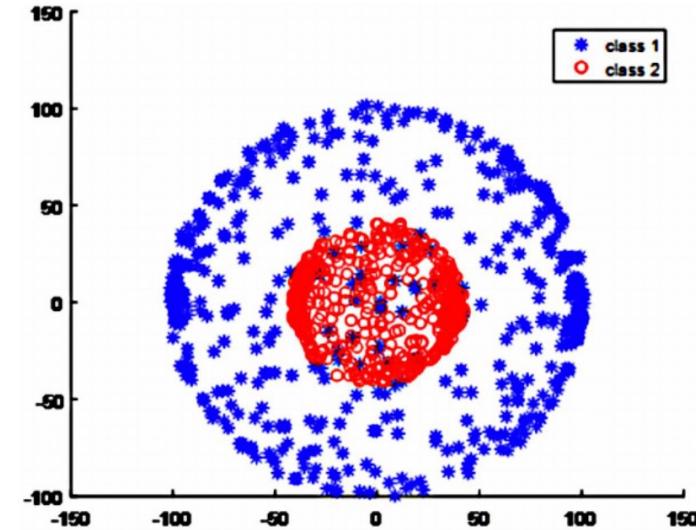
- On obtient dès lors une nouvelle base orthonormale donnée par les V_k (de coordonnées les $\alpha_1^k, \dots, \alpha_n^k$) qui sont les directions principales de l'ACP pour le noyau K . Pour projeter un nouveau point $x \mapsto \Phi(x)$ il suffit de calculer ses coordonnées dans l'espace formé par les vecteurs propres

$$\langle \Phi(x), v^k \rangle = \sum_{i=1}^n \alpha_i^k \langle \Phi(x_i), \Phi(x) \rangle .$$

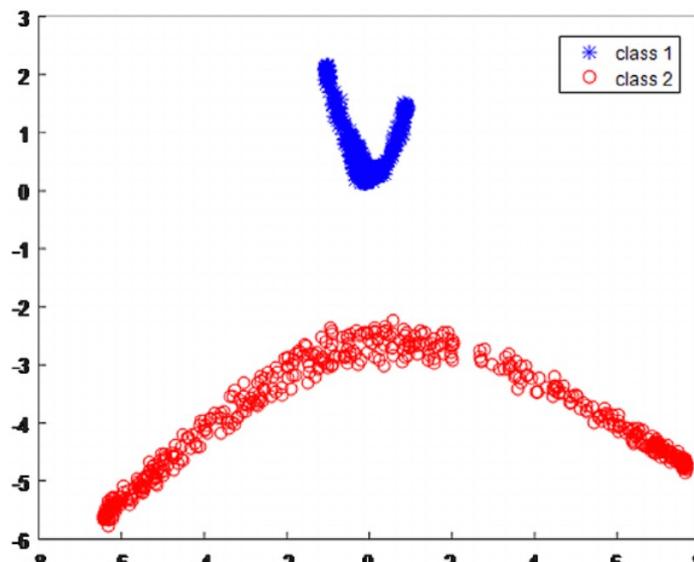
5 : ACP à noyaux



Projection sur les deux composantes principales d'une ACP à noyau $k(x, y) = (x'y + 1)^d$ avec $d = 5$.



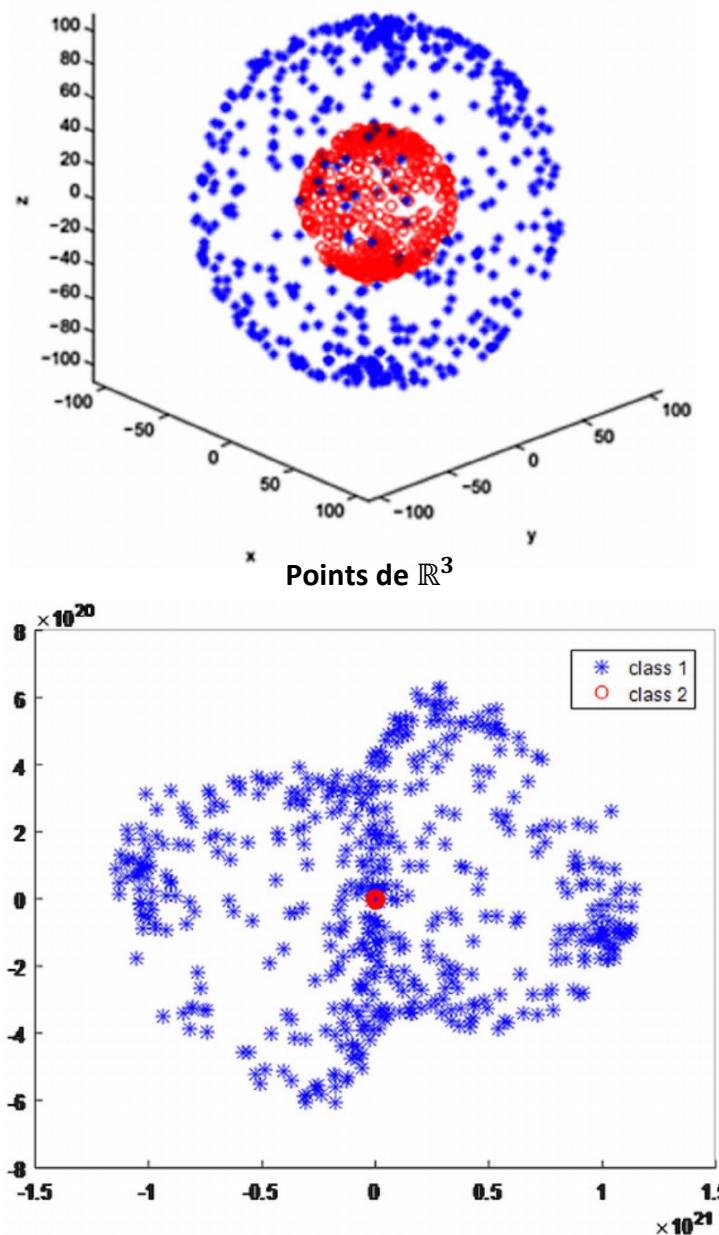
Projection sur les deux composantes principales d'une ACP classique



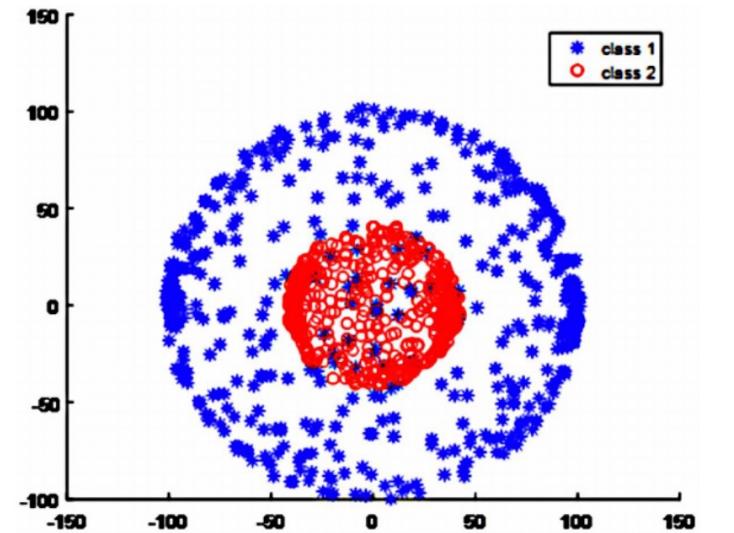
Projection sur les deux composantes principales d'une ACP à noyau $k(x, y) = \exp(-\frac{\|x-y\|^2}{2\sigma^2})$ avec $\sigma = 28$.

Illustration : I. Seikh Mazharul et al: Intelligent multidimensional data clustering and analysis. 2017

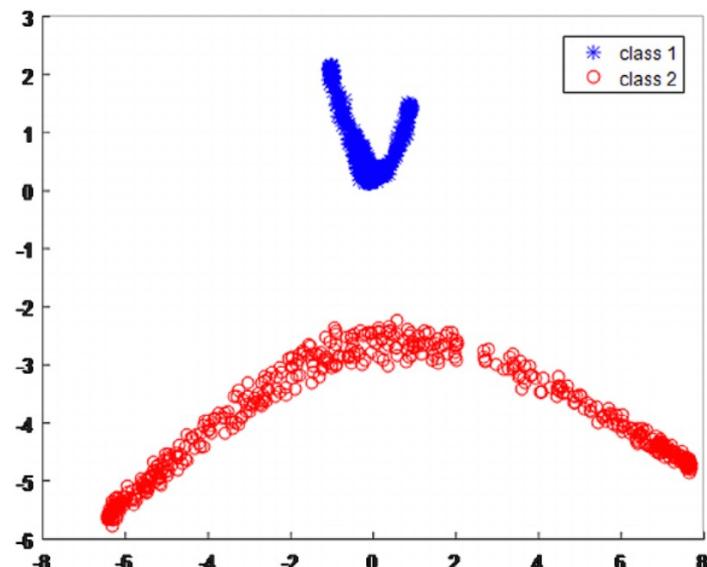
5 : ACP à noyaux



Projection sur les deux composantes principales d'une ACP à noyau $k(x, y) = (x'y + 1)^d$ avec $d = 5$.



Projection sur les deux composantes principales d'une ACP classique



Projection sur les deux composantes principales d'une ACP à noyau $k(x, y) = \exp(-\frac{\|x-y\|^2}{2\sigma^2})$ avec $\sigma = 28$.

Conclusions sur l'ACP

- Outils simple et efficace pour l'exploration et la compréhension de données.
- Outil puissant pour réduire la dimension d'un problème d'apprentissage.
- Possibilité de traiter des nuages de points mais aussi des fonctions.
- Intégration possible de noyaux pour rendre plus linéaires des problèmes d'apprentissage, en plus de réduire leur dimension.

Illustration : I. Seikh Mazharul et al: Intelligent multidimensional data clustering and analysis. 2017

Merci

Merci !