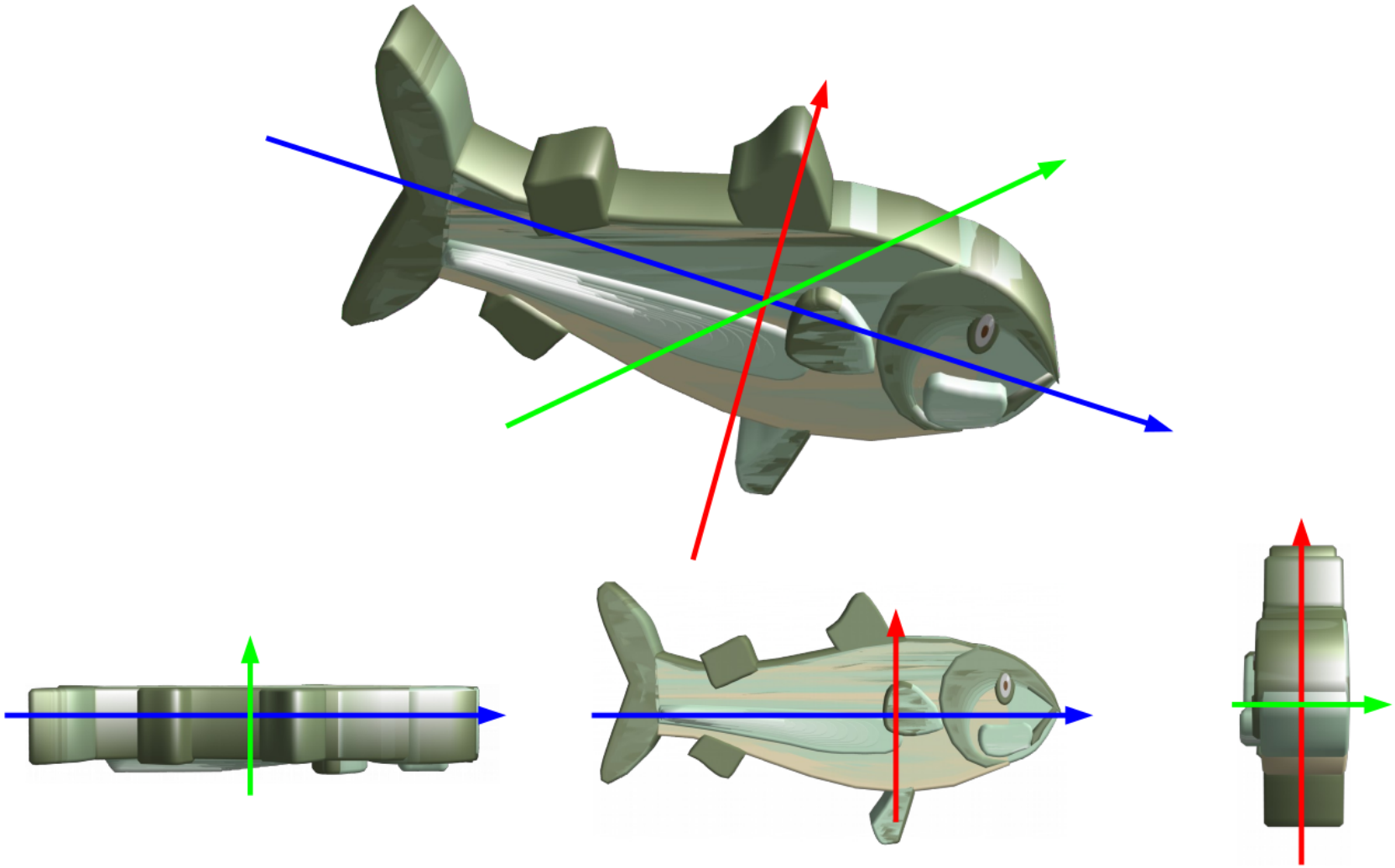
Statistique

Benjamin Bobbia

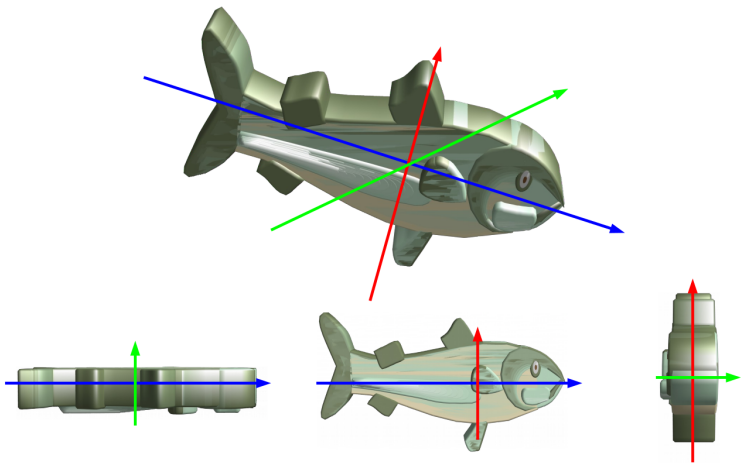
ISAE

1.0 Analyse en composantes principales (ACP)

Motivation : un problème de projection



Motivation : un problème de projection



Parmi les projections en 2D, toutes ne permettent pas de reconnaître aussi facilement l'objet initial, *i.e.* elles ne contiennent pas toutes la même **quantité d'information** et n'ont pas la même **capacité à résumer**.

La projection du milieu apparaît comme la plus fidèle à l'original. Elle correspond au plan où l'objet initial s'étale le plus, *i.e.* admet la **plus grande variabilité**.

L'information apportée par la 3ème dimension est **minimale** et sa perte n'est pas préjudiciable.

Quelques notations

Dans cette section, nous considérons un jeu de données quantitatives issu de n observations de p variables $x^1, \dots, x^p$ à valeurs réelles,

$$x_1, \dots, x_n \in \mathbb{R}^p$$

avec, pour tout $k \in \{1, \dots, n\}$,

$$x_k = \begin{pmatrix} x_k^1 \\ \vdots \\ x_k^p \end{pmatrix} \in \mathbb{R}^p.$$

Autrement dit, pour tout $k \in \{1, \dots, n\}$ et $\ell \in \{1, \dots, p\}$, $x_k^\ell \in \mathbb{R}$ est la valeur prise par la variable x^ℓ sur le k-ème individu.

Quelques notations

Échantillon : $x_1, \dots, x_n \in \mathbb{R}^p$
La **matrice des données (centrées)** de taille $n \times p$ est

$$X = \begin{pmatrix} x_1^1 - \overline{x}^1 & \dots & x_1^p - \overline{x}^p \\ \vdots & \vdots & \vdots \\ x_n^1 - \overline{x}^1 & \dots & x_n^p - \overline{x}^p \end{pmatrix}$$

où $\overline{x}^1, \dots, \overline{x}^p \in \mathbb{R}$ sont les moyennes des observations des variables $x^1, \dots, x^p$ respectivement,

$$\forall \ell \in \{1, \dots, p\}, \ \overline{x}^\ell = \frac{1}{n} \sum_{k=1}^{n} x_k^\ell.$$

Pour tout $k \in \{1, \dots, n\}$ et $\ell \in \{1, \dots, p\}$, la k-ème ligne de la matrice X contient les observations faites sur le k-ème individu et la ℓ-ème colonne de la matrice X contient les observations centrées de la variable x^ℓ.

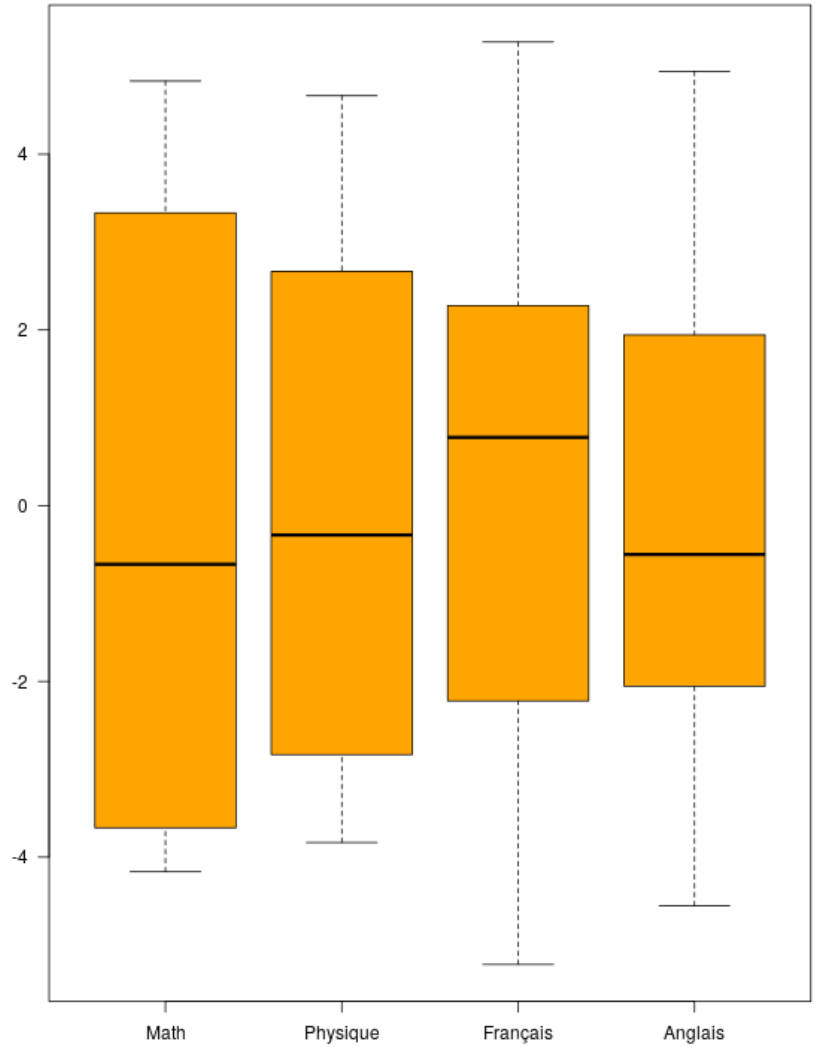
Exemple-jouet : observations

Considérons les notes obtenues par $n = 9$ étudiants dans $p = 4$ matières :

	Math	Physique	Français	Anglais
Benny	6.0	6.0	5.0	5.5
Bobby	8.0	8.0	8.0	8.0
Brandy	6.0	7.0	11.0	9.5
Coby	14.5	14.5	15.5	15.0
Daisy	14.0	14.0	12.0	12.5
Emily	11.0	10.0	5.5	7.0
Judy	5.5	7.0	14.0	11.5
Marty	13.0	12.5	8.5	9.5
Sandy	9.0	9.5	12.5	12.0
Moyenne	**9.67**	**9.83**	**10.22**	**10.06**

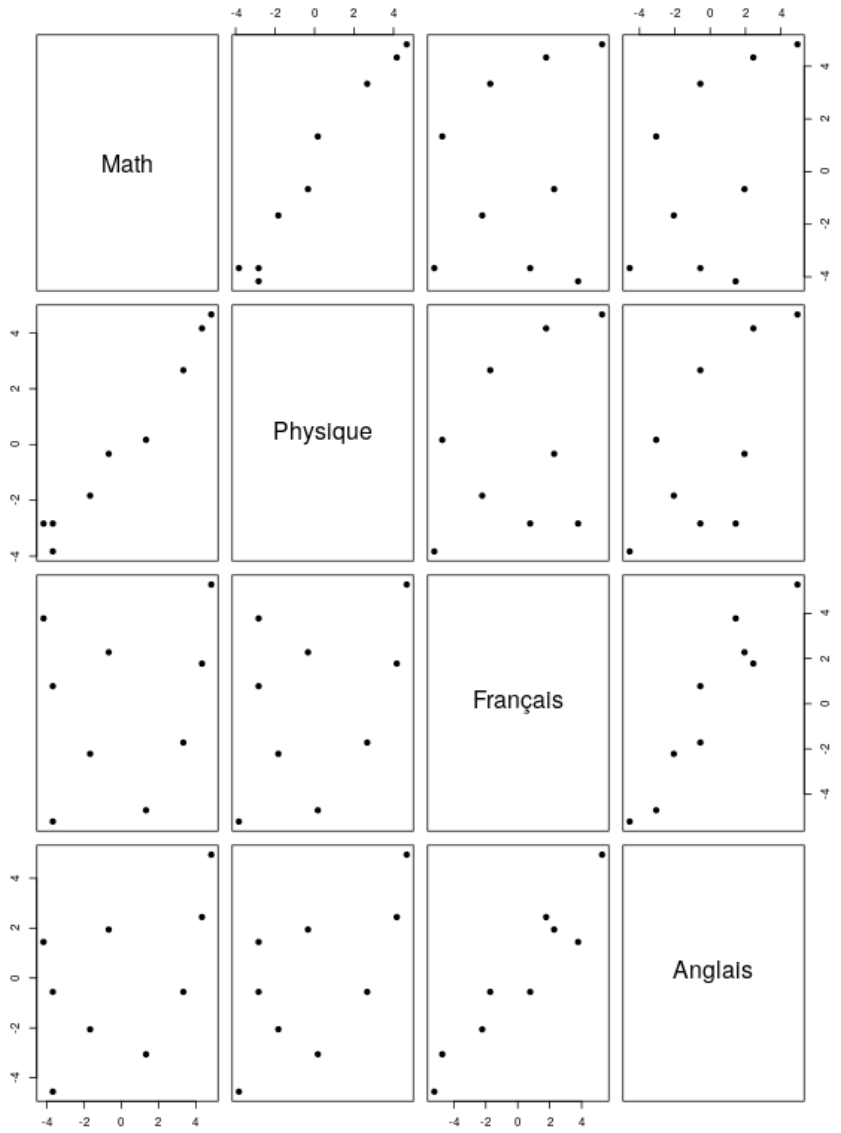
Exemple-jouet : données et visualisation élémentaire

$$X = \begin{pmatrix} -3.67 & -3.83 & -5.22 & -4.56 \\ -1.67 & -1.83 & -2.22 & -2.06 \\ -3.67 & -2.83 & 0.78 & -0.56 \\ 4.83 & 4.67 & 5.28 & 4.94 \\ 4.33 & 4.17 & 1.78 & 2.44 \\ 1.33 & 0.17 & -4.72 & -3.06 \\ -4.17 & -2.83 & 3.78 & 1.44 \\ 3.33 & 2.67 & -1.72 & -0.56 \\ -0.67 & -0.33 & 2.28 & 1.94 \end{pmatrix}$$

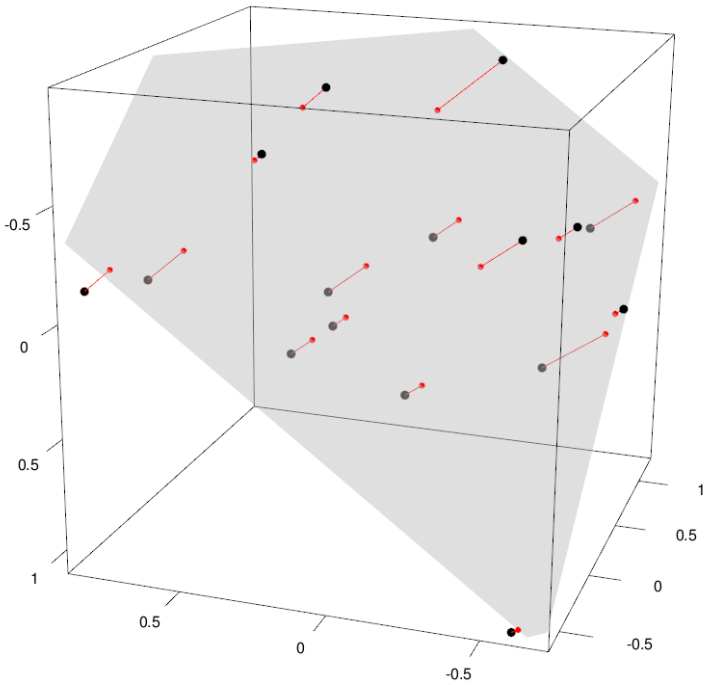


Exemple-jouet : données et visualisation élémentaire

$$X = \begin{pmatrix} -3.67 & -3.83 & -5.22 & -4.56 \\ -1.67 & -1.83 & -2.22 & -2.06 \\ -3.67 & -2.83 & 0.78 & -0.56 \\ 4.83 & 4.67 & 5.28 & 4.94 \\ 4.33 & 4.17 & 1.78 & 2.44 \\ 1.33 & 0.17 & -4.72 & -3.06 \\ -4.17 & -2.83 & 3.78 & 1.44 \\ 3.33 & 2.67 & -1.72 & -0.56 \\ -0.67 & -0.33 & 2.28 & 1.94 \end{pmatrix}$$



Principes de l'ACP



Déterminer les espaces de dimension inférieure à l'espace initial sur lesquels la projection du nuage de points initial est **la moins déformée possible**, *i.e.* celle qui conserve **le plus d'information au sens de la variabilité**.

Du point de vue des variables, cela correspond à chercher les **combinaisons linéaires** qui préservent au mieux la **structure de corrélation** entre les valeurs des données initiales.

Mesure de la variabilité

Échantillon : $x_1, \dots, x_n \in \mathbb{R}^p$

Une façon simple de généraliser la variance dans un cadre multidimensionnel consiste à définir l'**inertie (standard)** comme la somme des variances,

$$I(x) = \sum_{\ell=1}^{p} \sigma^2(x^\ell) = \frac{1}{n} \sum_{k=1}^{n} \sum_{\ell=1}^{p} (x_k^\ell - \overline{x}^\ell)^2.$$

Cette quantité s'écrit également $I(x) = \dfrac{1}{n} \displaystyle\sum_{k=1}^{n} \|x_k - \overline{x}\|^2$ où

$$\overline{x} = \begin{pmatrix} \overline{x}^1 \\ \vdots \\ \overline{x}^p \end{pmatrix} \in \mathbb{R}^p \quad \text{et} \quad \forall v = \begin{pmatrix} v_1 \\ \vdots \\ v_p \end{pmatrix} \in \mathbb{R}^p, \ \|v\|^2 = v^\top v = \sum_{\ell=1}^{p} v_\ell^2.$$

Projection des données

Soit $v \in \mathbb{R}^p$ tel que $\|v\|^2 = 1$.

Pour tout $k \in \{1, \dots, n\}$, la **projection orthogonale** du vecteur observé $x_k - \overline{x}$ sur la **droite engendrée par** v est donnée par

$$\langle x_k - \overline{x}, v \rangle v$$

avec le **produit scalaire** $\langle u, v \rangle = u^\top v = \sum_{\ell=1}^{p} u_\ell v_\ell$ pour tout $u, v \in \mathbb{R}^p$.

Autrement dit, le **résumé des observations le long de** v est donné par les coordonnées des n vecteurs initiaux

$$\langle x_1 - \overline{x}, v \rangle, \dots, \langle x_n - \overline{x}, v \rangle \in \mathbb{R}.$$

Projection des données

Soit $v \in \mathbb{R}^p$ tel que $\|v\|^2 = 1$.
Observations projetées sur $\mathbb{R}v$: $\langle x_1 - \overline{x}, v\rangle, \ldots, \langle x_n - \overline{x}, v\rangle \in \mathbb{R}$.

Les données projetées sont **centrées** par construction,

$$\frac{1}{n}\sum_{k=1}^{n}\langle x_k - \overline{x}, v\rangle = \langle \frac{1}{n}\sum_{k=1}^{n} x_k - \overline{x}, v\rangle = \langle \overline{x} - \overline{x}, v\rangle = 0.$$

L'inertie des données projetées (ici, il s'agit simplement de la variance) vaut

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^{n}\langle x_k - \overline{x}, v\rangle^2 &= \frac{1}{n}\sum_{k=1}^{n} v^\top (x_k - \overline{x})(x_k - \overline{x})^\top v \\ &= v^\top \left(\frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x})(x_k - \overline{x})^\top\right) v = v^\top \Sigma v.\end{aligned}$$

Matrice de covariance de notre nuage de points (symétrique)

Matrice de covariance

La matrice $p \times p$ qui apparaît dans le calcul de l'inertie précédent,

$$\Sigma = \frac{1}{n} \sum_{k=1}^{n} (x_k - \overline{x})(x_k - \overline{x})^\top,$$

s'appelle la **matrice de covariance** car, pour tout $\ell_1, \ell_2 \in \{1, \dots, p\}$,

$$\Sigma_{\ell_1 \ell_2} = \frac{1}{n} \sum_{k=1}^{n} \left(x_k^{\ell_1} - \overline{x}^{\ell_1} \right) \left(x_k^{\ell_2} - \overline{x}^{\ell_2} \right) = \sigma(x^{\ell_1}, x^{\ell_2}).$$

Cette matrice s'écrit également $\Sigma = X^\top W X$ avec $W = n^{-1} Id_n$ et vérifie

- Σ est **symétrique** : $\Sigma_{\ell_1 \ell_2} = \sigma(x^{\ell_1}, x^{\ell_2}) = \sigma(x^{\ell_2}, x^{\ell_1}) = \Sigma_{\ell_2 \ell_1}$,
- Σ est **positive** : $\forall u \in \mathbb{R}^p$, $u^\top \Sigma u = \|Xu\|^2 / n \geqslant 0$.

Première direction principale

Déterminer la droite sur laquelle la projection du nuage de points initial conserve **le plus d'information au sens de la variabilité** revient donc à chercher le vecteur unitaire $v \in \mathbb{R}^p$ tel que $v^\top \Sigma v \in \mathbb{R}$ soit maximal.

Puisque Σ est symétrique et positive, le vecteur v solution de ce problème est donné par **? ? ?**

Première direction principale

Déterminer la droite sur laquelle la projection du nuage de points initial conserve **le plus d'information au sens de la variabilité** revient donc à chercher le vecteur unitaire $v \in \mathbb{R}^p$ tel que $v^\top \Sigma v \in \mathbb{R}$ soit maximal.

Puisque Σ est symétrique et positive, le vecteur v solution de ce problème est donné par le **vecteur propre associé à la plus grande valeur propre de** Σ. Vecteurs propres de norme 1

En effet, le théorème spectral assure qu'il existe une **matrice orthogonale** V (*i.e.* $V^{-1} = V^\top$) de taille $p \times p$ dont les colonnes $v^1, \dots, v^p \in \mathbb{R}^p$ forment une **base orthonormale** de $\mathbb{R}^p$ et correspondent aux **vecteurs propres** de Σ associés aux **valeurs propres** respectives

$$\lambda_1 \geqslant \cdots \geqslant \lambda_p \geqslant 0.$$

Le vecteur $v^1 \in \mathbb{R}^p$ est la **première direction principale** et l'inertie maximale le long d'une droite est donnée par ${v^1}^\top \Sigma v^1 = \lambda_1$.

Part d'inertie expliquée

Le problème unidimensionnel précédent se généralise immédiatement à la recherche d'un espace de dimension $d \leqslant p$ sur lequel la projection du nuage de points initial conserve **le plus d'information au sens de la variabilité**.

La solution est fournie par l'**espace E_d engendré par les d vecteurs propres orthonormaux** $v^1, \ldots, v^d \in \mathbb{R}^p$ qui sont donnés par les d premières colonnes de la matrice V (*i.e.* les d premières directions principales).

L'inertie des données projetées dans E_d vaut

$$\sum_{\ell=1}^{d} {v^\ell}^\top \Sigma v^\ell = \sum_{\ell=1}^{d} \lambda_\ell \left(\leqslant \sum_{\ell=1}^{p} \lambda_\ell = I(x) \right).$$

La **part d'inertie expliquée** est la quantité

$$\frac{1}{I(x)} \sum_{\ell=1}^{d} \lambda_\ell.$$

Composantes principales

Les directions principales $v^1, \ldots, v^p \in \mathbb{R}^p$ forment une **base orthonormale** de $\mathbb{R}^p$ dans laquelle les données initiales peuvent être représentées. Cette représentation correspond à des **combinaisons linéaires** des variables initiales qui préservent au mieux la **structure de corrélation**.

Les coordonnées des données initiales dans la base des directions principales sont fournies par la matrice de taille $n \times p$ définie par

$$C = XV.$$

La matrice C est la **matrice des composantes principales**.

Composantes principales

Les colonnes $c^1, \dots, c^p \in \mathbb{R}^n$ de C s'appellent les **composantes principales** et doivent être considérées comme p variables « virtuelles » obtenues par combinaisons linéaires des variables initiales,

$$\forall \ell \in \{1, \dots, p\}, \ c^\ell = \begin{pmatrix} c_1^\ell \\ \vdots \\ c_n^\ell \end{pmatrix} = Xv^\ell \in \mathbb{R}^n.$$

Les composantes principales sont **centrées** par construction et leur matrice de covariance est donnée par

$$C^\top W C = V^\top X^\top W X V = V^\top \Sigma V = \Lambda = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \lambda_p \end{pmatrix}.$$

Composantes principales

= linéairement indépendantes

Les composantes principales sont des variables **décorrélées** et **ordonnées par la quantité d'information**,

$$\forall \ell, \ell' \in \{1, \dots, p\},\ \sigma(c^\ell, c^{\ell'}) = \begin{cases} \lambda_\ell & \text{si } \ell = \ell', \\ 0 & \text{sinon.} \end{cases}$$

Pour représenter les données initiales dans un espace de dimension $d \leqslant p$, nous utilisons donc les coordonnées fournies par les d premières composantes principales $c^1, \dots, c^d$.

Le cas particulier $d = 2$ correspond au **plan principal** et permet de donner une représentation graphique des données de dimension p (correspondant à la part d'inertie expliquée par les deux premières directions principales).

Exemple-jouet : plan principal

Matrice de covariance $\Sigma = X^\top W X$:

$$\begin{array}{c} \\ \text{Math} \\ \text{Physique} \\ \text{Français} \\ \text{Anglais} \end{array} \begin{array}{c} \begin{array}{cccc} \text{Math} & \text{Physique} & \text{Français} & \text{Anglais} \end{array} \\ \begin{pmatrix} 11.39 & 9.92 & 2.66 & 4.82 \\ 9.92 & 8.94 & 4.12 & 5.48 \\ 2.66 & 4.12 & 12.06 & 9.29 \\ 4.82 & 5.48 & 9.29 & 7.91 \end{pmatrix} \end{array}$$

Inertie du nuage de points initial : $I(x) = \mathrm{Tr}(\Sigma) = 40.31$

Matrice des directions principales :

$$V = \begin{array}{c} \begin{array}{cccc} v^1 & v^2 & v^3 & v^4 \end{array} \\ \begin{pmatrix} -0.515 & 0.569 & 0.185 & 0.614 \\ -0.508 & 0.371 & -0.450 & -0.634 \\ -0.492 & -0.658 & -0.460 & 0.335 \\ -0.484 & -0.325 & 0.742 & -0.329 \end{pmatrix} \end{array}$$

Exemple-jouet : plan principal

Matrice des composantes principales $C = XV$:

X : Matrice des observations centrées
V : Matrice des directions principales

	c^1	c^2	c^3	c^4
Benny	8.61	1.41	0.07	−0.07
Bobby	3.88	0.50	0.01	0.07
Brandy	3.21	−3.47	−0.17	−0.01
Coby	−9.85	−0.60	0.04	0.15
Daisy	−6.41	2.05	−0.08	−0.19
Emily	3.03	4.92	0.08	0.14
Judy	1.03	−6.38	−0.16	0.03
Marty	−1.95	4.20	−0.20	−0.04
Sandy	−1.55	−2.63	0.42	−0.07

Valeurs propres de Σ : $\lambda_1 = 28.23$, $\lambda_2 = 12.03$, $\lambda_3 = 0.03$ et $\lambda_4 = 0.01$

Exemple-jouet : plan principal

Matrice des composantes principales $C = XV$:

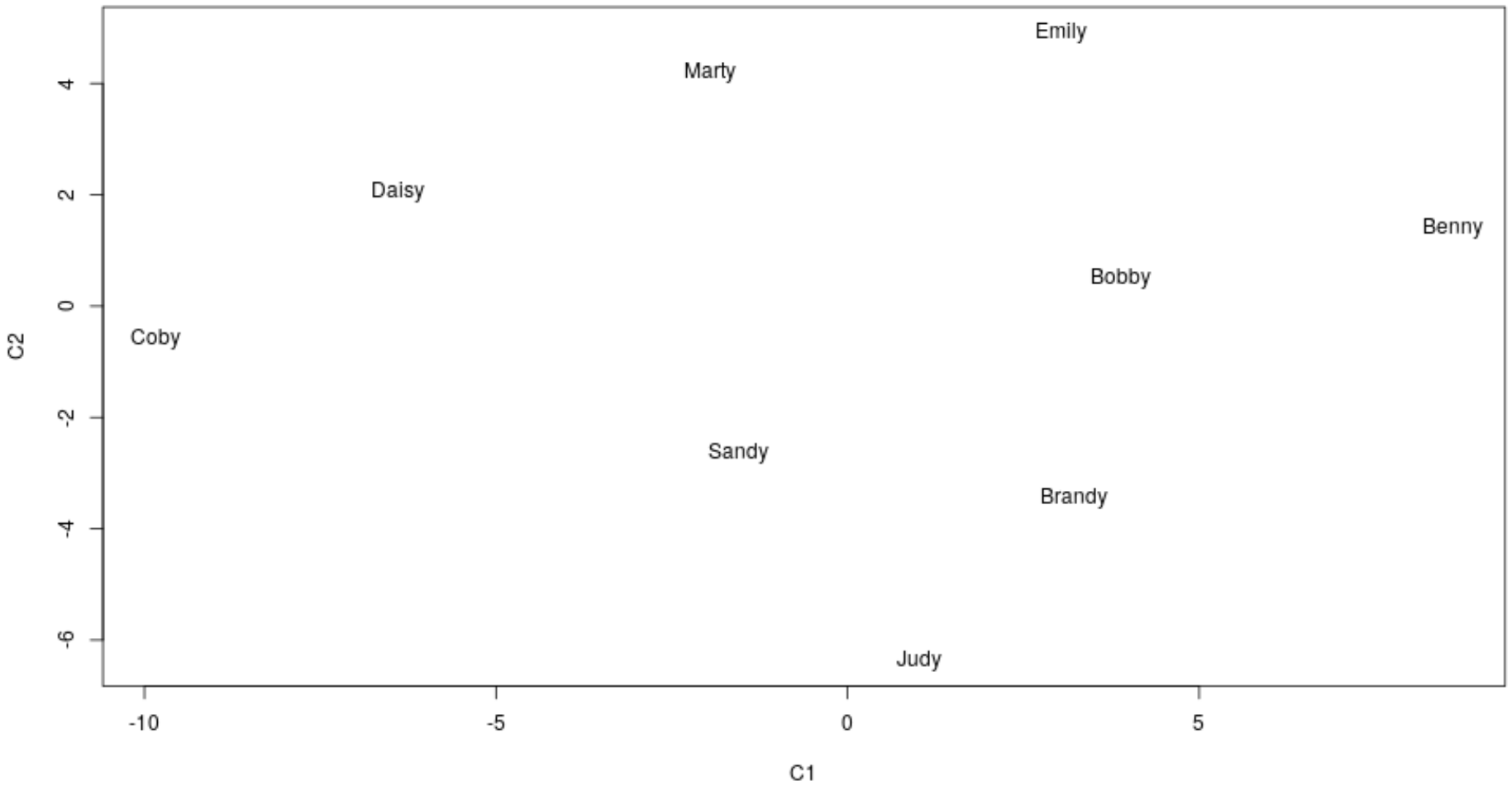
	c^1	c^2	c^3	c^4
Benny	8.61	1.41	0.07	−0.07
Bobby	3.88	0.50	0.01	0.07
Brandy	3.21	−3.47	−0.17	−0.01
Coby	−9.85	−0.60	0.04	0.15
Daisy	−6.41	2.05	−0.08	−0.19
Emily	3.03	4.92	0.08	0.14
Judy	1.03	−6.38	−0.16	0.03
Marty	−1.95	4.20	−0.20	−0.04
Sandy	−1.55	−2.63	0.42	−0.07

Valeurs propres de Σ : $\lambda_1 = 28.23$, $\lambda_2 = 12.03$, $\lambda_3 = 0.03$ et $\lambda_4 = 0.01$

Choix du nombre de valeurs propres : représenter I(x) en fonction de lambda1, lambda1+2, lambda1+2+3 ... et choisir à quel % on s'arrête

Part d'inertie expliquée par le **plan principal** : $(\lambda_1 + \lambda_2)/I(x) = 99.89\%$

Exemple-jouet : plan principal



Représentation des variables

Pour discuter du lien entre les variables initiales et les composantes principales, nous considérons les **coefficients de corrélation linéaire** :

$$\forall \ell, \ell' \in \{1, \dots, p\}, \ \rho(x^\ell, c^{\ell'}) = \frac{\sigma(x^\ell, c^{\ell'})}{\sigma(x^\ell)\sqrt{\lambda_{\ell'}}} = \frac{\sqrt{\lambda_{\ell'}}}{\sigma(x^\ell)} v_\ell^{\ell'}$$

car $X = CV^\top$ par orthogonalité, *i.e.* $x^\ell = \sum_{j=1}^{p} v_\ell^j c^j$.

Pour $\ell \in \{1, \dots, p\}$, le point

$$\Big(\rho(x^\ell, c^1), \rho(x^\ell, c^2)\Big) = \left(\frac{\sqrt{\lambda_1}}{\sigma(x^\ell)} v_\ell^1, \frac{\sqrt{\lambda_2}}{\sigma(x^\ell)} v_\ell^2\right)$$

permet de **représenter graphiquement le lien entre x^ℓ et les deux premières composantes principales**.

Cercle des corrélations

Par construction, pour tout $\ell \in \{1, \dots, p\}$, nous avons

$$\sum_{\ell'=1}^{p} \rho(x^\ell, c^{\ell'})^2 = \sum_{\ell'=1}^{p} \frac{\lambda_{\ell'}}{\sigma^2(x^\ell)} \left(v_\ell^{\ell'}\right)^2 = \frac{1}{\sigma^2(x^\ell)} \left(V \Lambda V^\top\right)_{\ell\ell} = \frac{\Sigma_{\ell\ell}}{\sigma^2(x^\ell)} = 1.$$

Ainsi,

$$\rho(x^\ell, c^1)^2 + \rho(x^\ell, c^2)^2 \leqslant 1$$

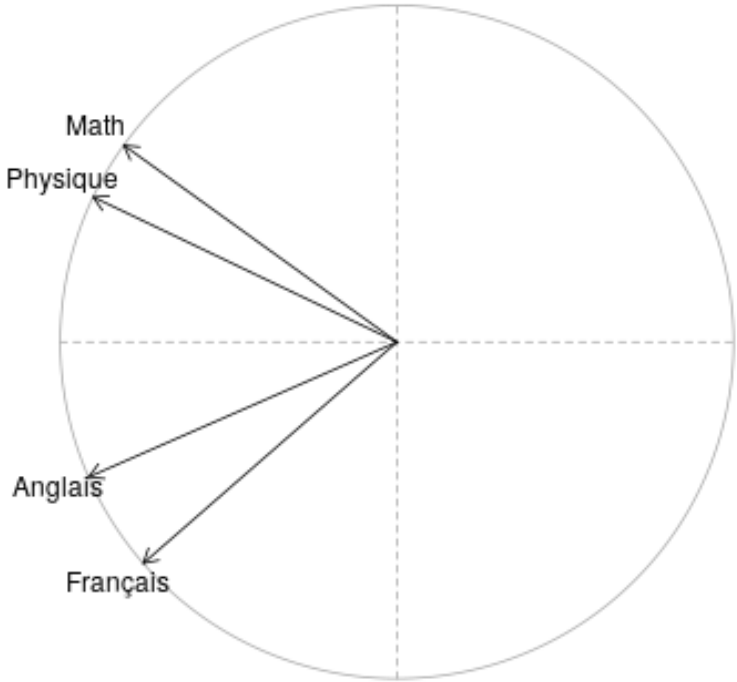
et le point $\Big(\rho(x^\ell, c^1), \rho(x^\ell, c^2)\Big)$ est **dans le disque de rayon unité**. La proximité au cercle traduit la **qualité de la représentation** et la direction indique les **liens avec les deux premières composantes**.

La représentation de toutes les variables initiales sur un même graphique s'appelle le **cercle des corrélations**.

Exemple-jouet : cercle des corrélations

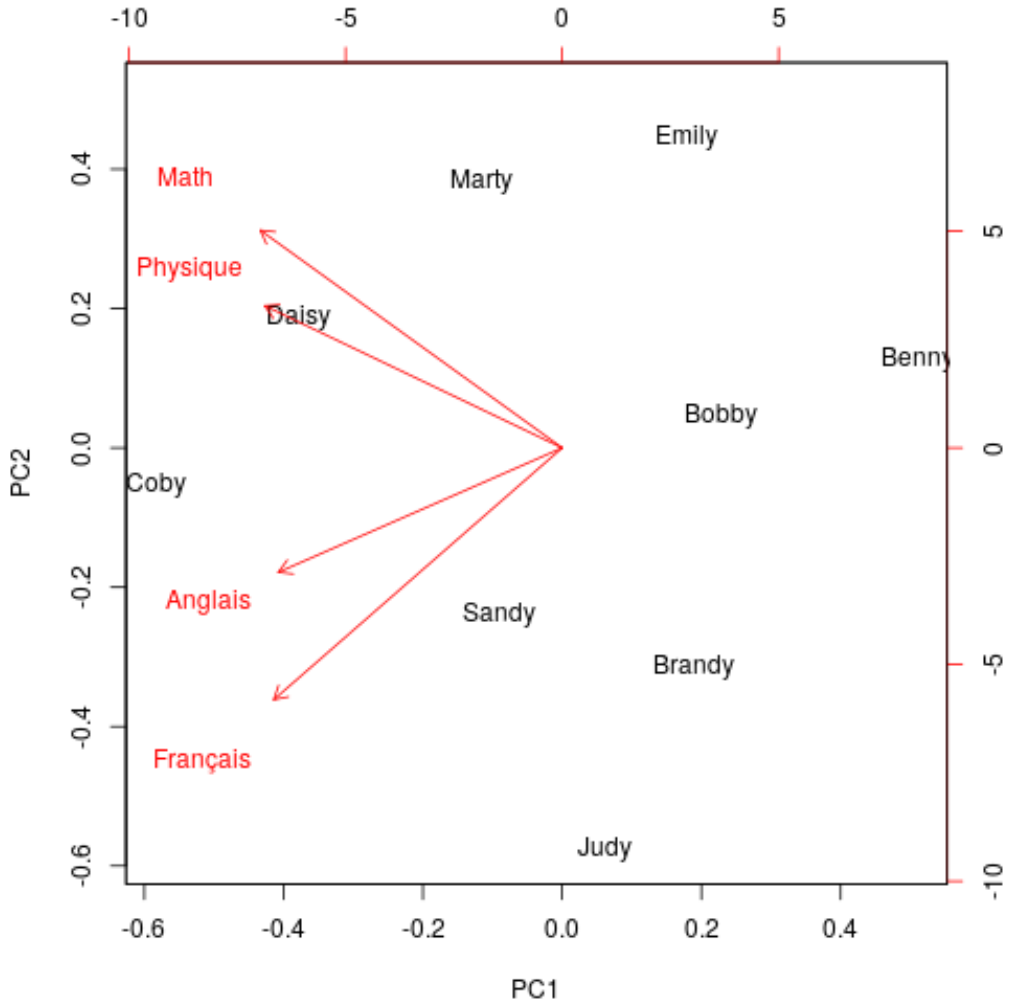
Coefficients de corrélation linéaire :

	c^1	c^2
Math	-0.81	0.58
Physique	-0.90	0.43
Français	-0.75	-0.66
Anglais	-0.91	-0.40



Ecart entre maths et physique faible dans le plan principal

Exemple-jouet : représentation biplot



Représentation simultanée des individus et des variables

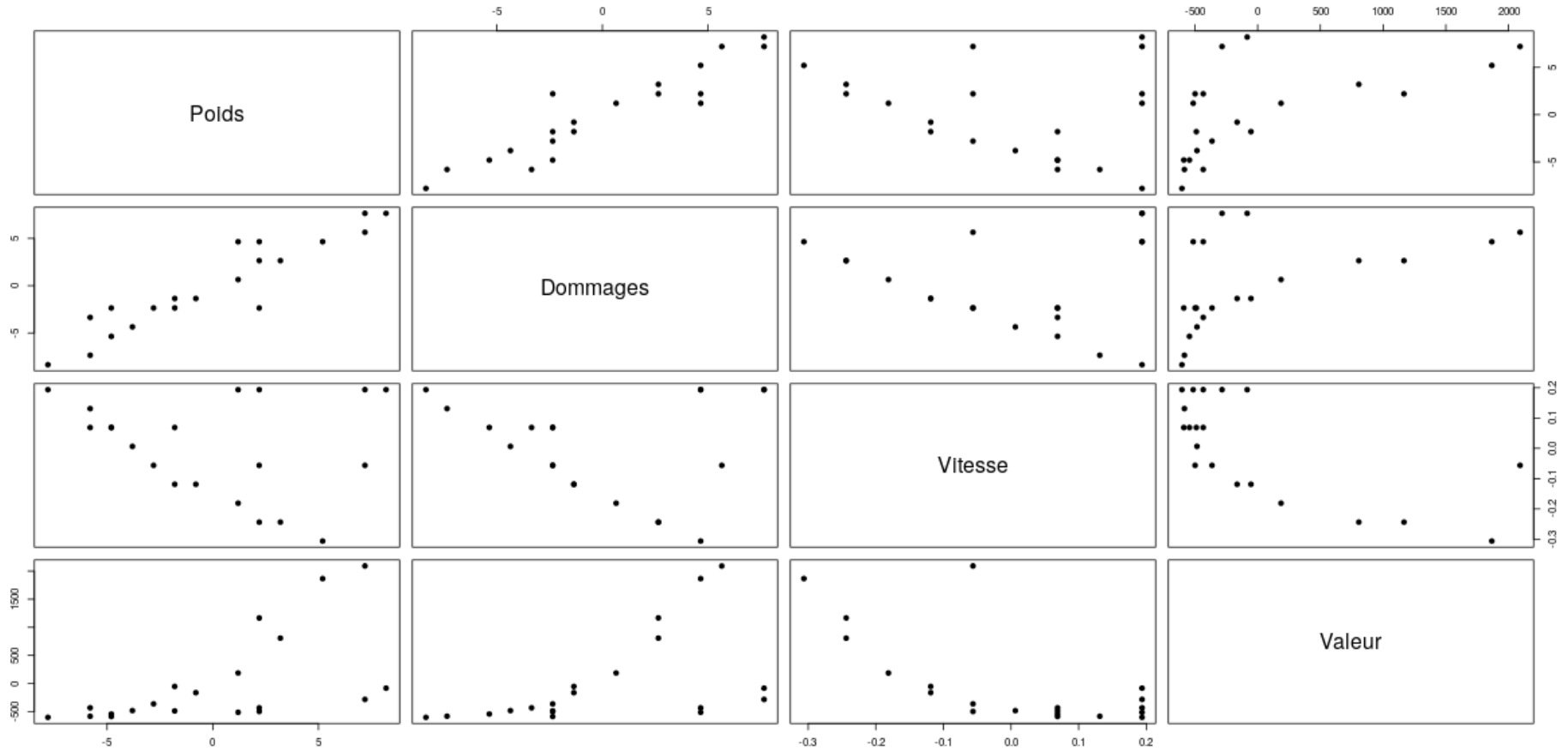
Exemple (un peu plus sérieux mais pas trop)

Considérons les $p = 4$ caractéristiques des $n = 20$ arcs et arbalètes sans enchantement du jeu « *The Elder Scrolls V : Skyrim* » (avec extensions) :

	Poids	Dommages	Vitesse	Valeur
Arc long	5	6	1.0000	30
Arc de chasse	7	7	0.9375	50
Arc nordique du Héros	7	11	0.8750	200
Arc nordique antique	8	12	0.8750	45
Arc impérial	8	9	0.8750	90
Arc de Parjure	11	12	0.8750	145
Arc orque	9	10	0.8126	150
Arc falmer	15	12	0.7500	135
Arc dwemer	10	12	0.7500	270
Arc elfique	12	13	0.6875	470
Arc nordique	11	13	0.6875	580
Arc de verre	14	15	0.6250	820
Arc d'ébonite	16	17	0.5625	1440
Arc de stalhrim	15	17	0.5625	1800
Arc daédra	18	19	0.5000	2500
Arc d'os de dragon	20	20	0.7500	2725
Arbalète	14	19	1.0000	120
Arbalète améliorée	15	19	1.0000	200
Arbalète dwemer	20	22	1.0000	350
Arbalète dwemer améliorée	21	22	1.0000	550

Exemple (un peu plus sérieux mais pas trop)

Comme dans l'exemple-jouet, nous définissons la matrice X des données centrées et une étude exploratoire grossière suggère une structure de corrélation entre les variables.



Exemple (un peu plus sérieux mais pas trop)

En diagonalisant la matrice de covariance $\Sigma = X^\top W X$ où $W = n^{-1} Id_n$, nous obtenons la matrice V des directions principales, la matrice des composantes principales $C = XV$ et les valeurs propres

$$\lambda_1 = 643016.54,\ \lambda_2 = 28.14,\ \lambda_3 = 1.37 \text{ et } \lambda_4 = 0.01.$$

L'inertie du nuage de points initial vaut $I(x) = \mathrm{Tr}(\Sigma) = 643046.1$.

La part d'inertie expliquée par le plan principal est égale à

$$\frac{\lambda_1 + \lambda_2}{I(x)} = 99.99\%.$$

Exemple (un peu plus sérieux mais pas trop)

En diagonalisant la matrice de covariance $\Sigma = X^\top W X$ où $W = n^{-1} Id_n$, nous obtenons la matrice V des directions principales, la matrice des composantes principales $C = XV$ et les valeurs propres

$$\lambda_1 = 643016.54,\ \lambda_2 = 28.14,\ \lambda_3 = 1.37 \text{ et } \lambda_4 = 0.01.$$

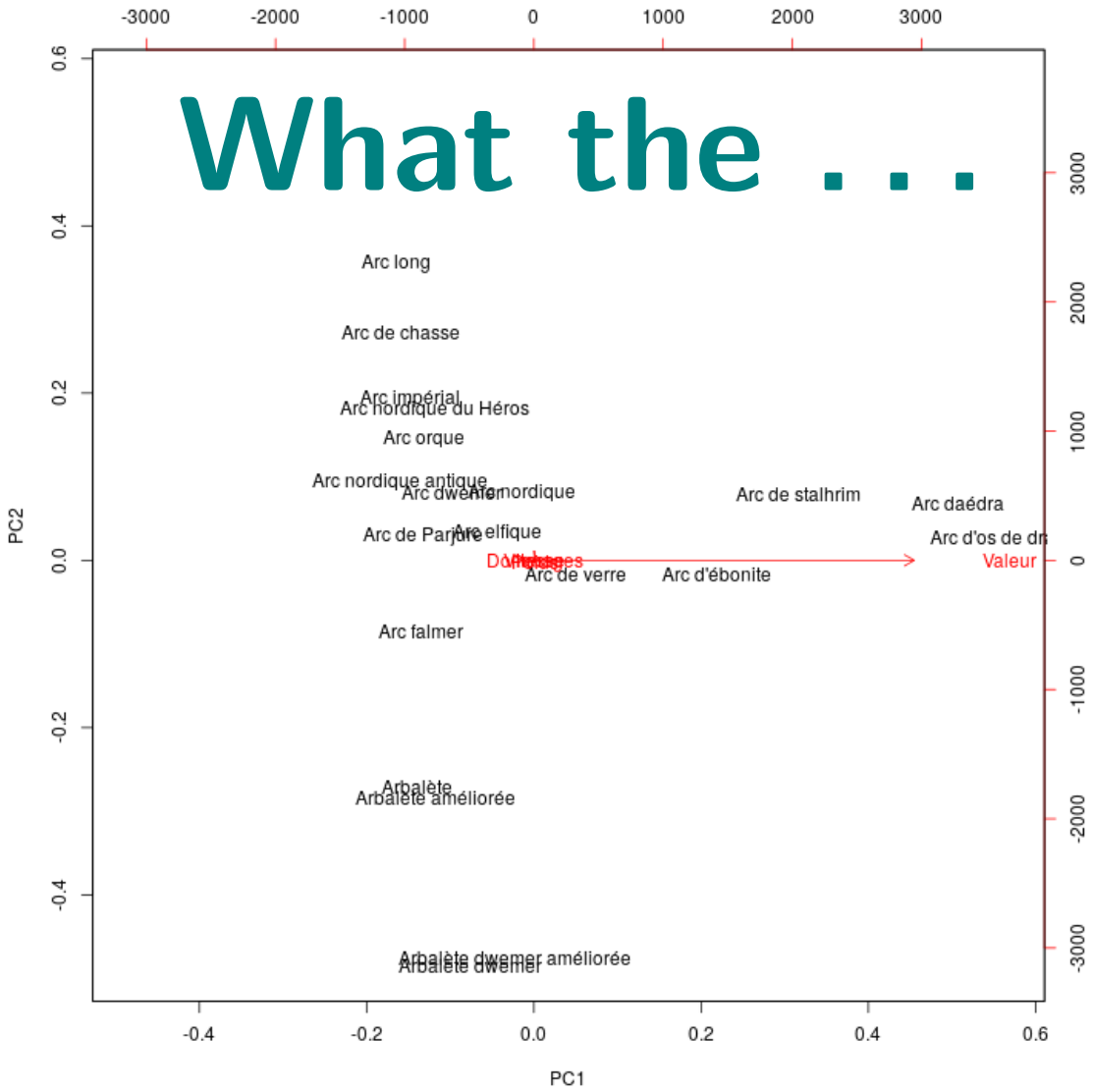
L'inertie du nuage de points initial vaut $I(x) = \mathrm{Tr}(\Sigma) = 643046.1$.

La part d'inertie expliquée par le plan principal est égale à

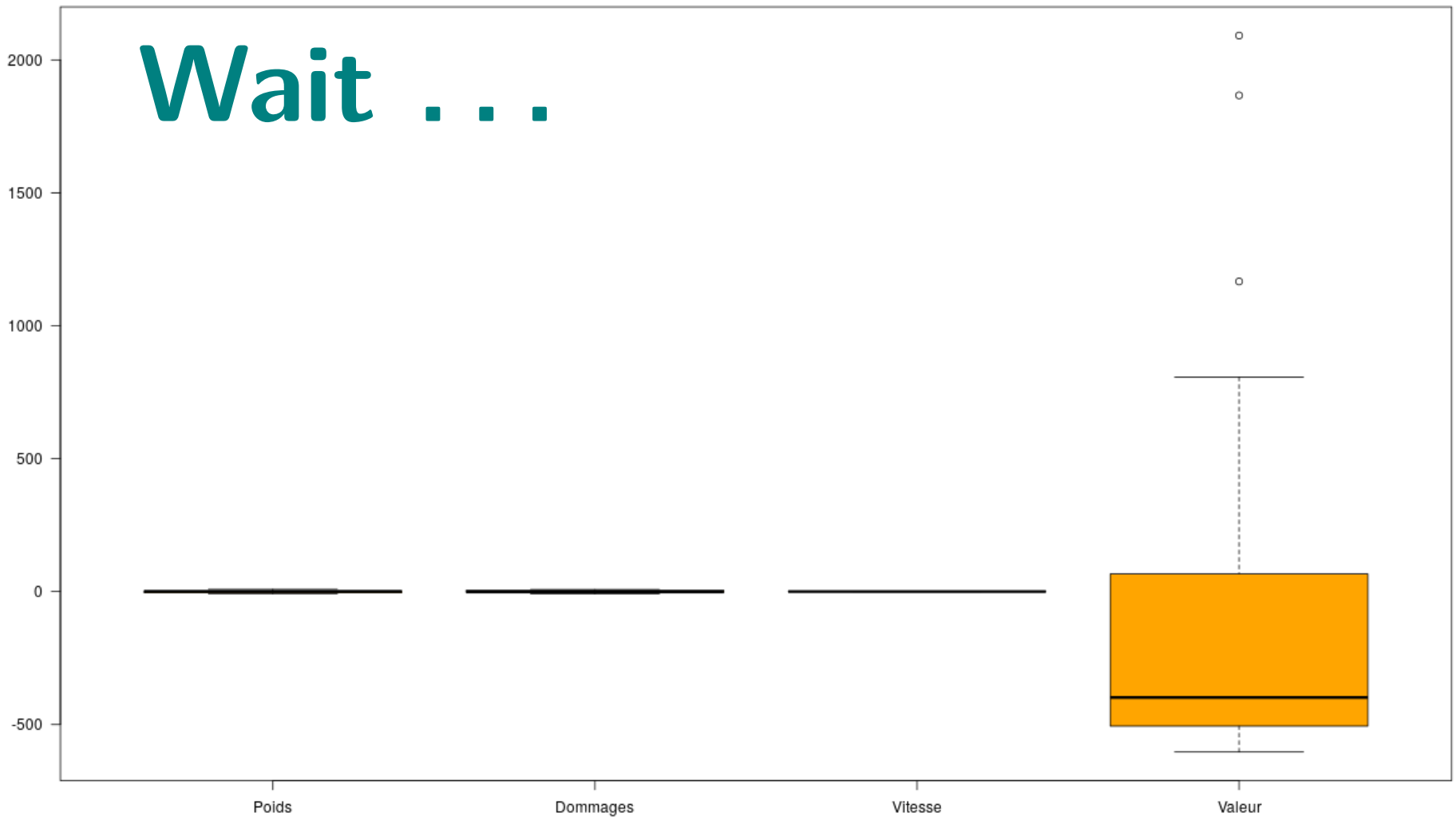
$$\frac{\lambda_1 + \lambda_2}{I(x)} = 99.99\%.$$

Oh yeah ! Go, biplot !

Exemple (un peu plus sérieux mais pas trop)



Exemple (un peu plus sérieux mais pas trop)

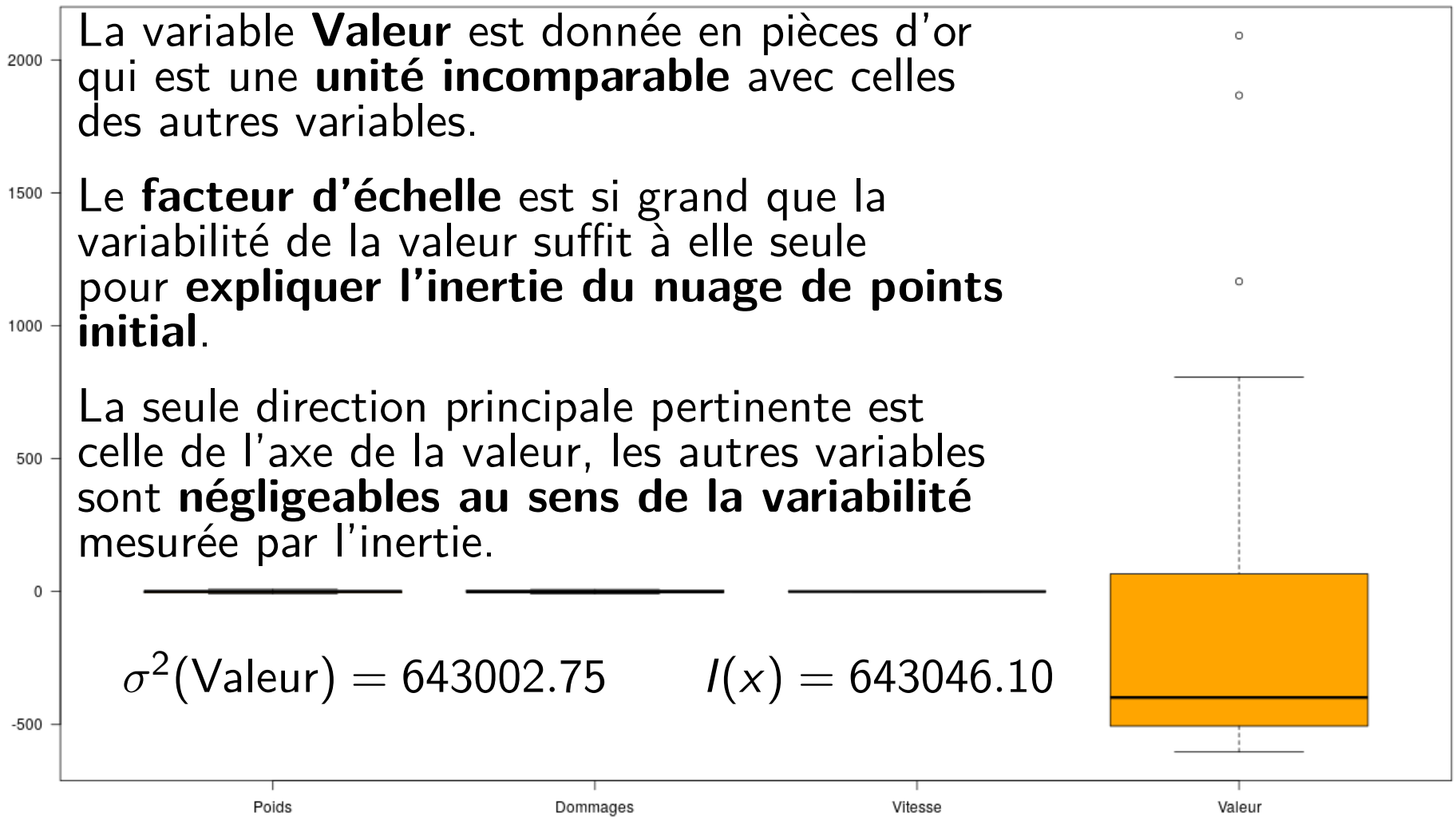


Exemple (un peu plus sérieux mais pas trop)

La variable **Valeur** est donnée en pièces d'or qui est une **unité incomparable** avec celles des autres variables.

Le **facteur d'échelle** est si grand que la variabilité de la valeur suffit à elle seule pour **expliquer l'inertie du nuage de points initial**.

La seule direction principale pertinente est celle de l'axe de la valeur, les autres variables sont **négligeables au sens de la variabilité** mesurée par l'inertie.



$\sigma^2(\text{Valeur}) = 643002.75 \qquad I(x) = 643046.10$

ACP et données réduites

La variance quantifie la variabilité des observations mais **sa valeur dépend de l'unité** utilisée. Un **changement d'échelle** modifie cette mesure,

$$\forall a \in \mathbb{R},\ \sigma^2(ax) = a^2\sigma^2(x).$$

L'inertie comme **somme des variances** attribue donc une importance à chaque variable qui **dépend de l'unité physique**.

Avantages de la réduction :

- exprime les données dans une **échelle neutre**.
- évite qu'une variable concentre toute la variabilité.

Inconvénients de la réduction :

- l'information de l'unité de mesure est perdue.
- un bruit se retrouve avec une variance apparente égale à celle d'une variable informative.

ACP et données réduites

La variance quantifie la variabilité des observations mais **sa valeur dépend de l'unité** utilisée. Un **changement d'échelle** modifie cette mesure,

$$\forall a \in \mathbb{R},\ \sigma^2(ax) = a^2\sigma^2(x).$$

L'inertie comme **somme des variances** attribue donc une importance à chaque variable qui **dépend de l'unité physique**.

Faire l'ACP des données réduites revient à **diagonaliser la matrice de corrélation**,

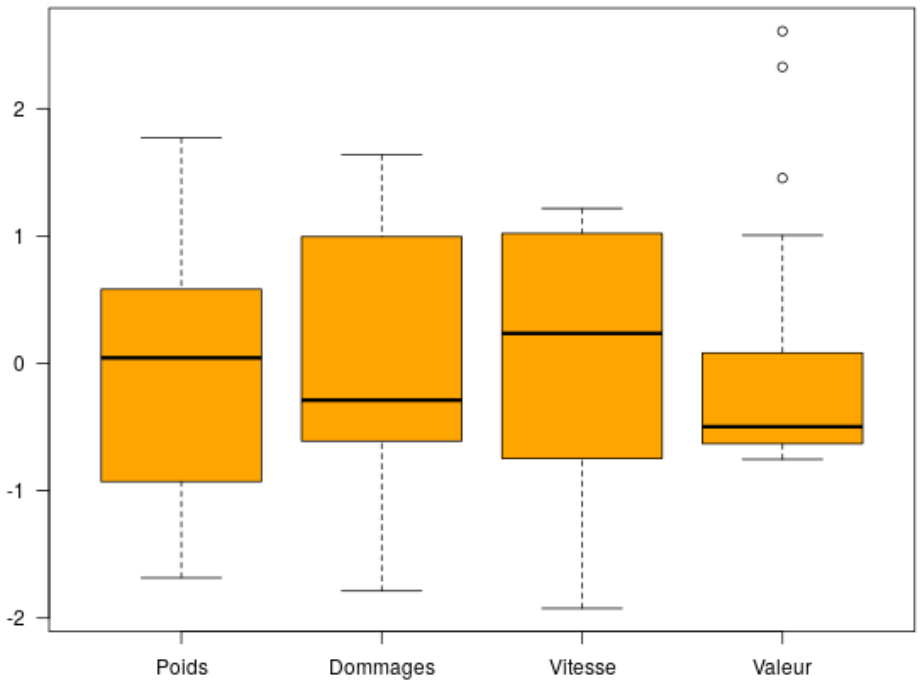
$$\begin{pmatrix} 1.0 & \rho(x^1,x^2) & \dots & \rho(x^1,x^p) \\ \rho(x^1,x^2) & 1.0 & \dots & \vdots \\ \vdots & \vdots & \dots & \rho(x^{p-1},x^p) \\ \rho(x^1,x^p) & \rho(x^2,x^p) & \dots & 1.0 \end{pmatrix}$$

Exemple (un peu plus sérieux mais pas trop) avec réduction

La version centrée-réduite des $p = 4$ variables observées sur les $n = 20$ arcs et arbalètes conduit à considérer la matrice $\tilde{X}$ de taille $n \times p$ définie par

$$\forall k \in \{1, \dots, n\},\ \forall \ell \in \{1, \dots, p\},\ \tilde{X}_{k\ell} = \tilde{x}_k^\ell = \frac{x_k^\ell - \overline{x}^\ell}{\sigma(x^\ell)}.$$

Poids	Dommages	Vitesse	Valeur
−1.68	−1.79	1.22	−0.75
−1.25	−1.57	0.82	−0.73
−1.25	−0.72	0.43	−0.54
−1.04	−0.50	0.43	−0.73
−1.04	−1.15	0.43	−0.68
−0.39	−0.50	0.43	−0.61
−0.82	−0.93	0.04	−0.60
0.47	−0.50	−0.35	−0.62
−0.60	−0.50	−0.35	−0.45
−0.17	−0.29	−0.75	−0.20
−0.39	−0.29	−0.75	−0.07
0.26	0.14	−1.14	0.23
0.69	0.57	−1.53	1.01
0.47	0.57	−1.53	1.45
1.12	1.00	−1.92	2.33
1.55	1.21	−0.35	2.61
0.26	1.00	1.22	−0.64
0.47	1.00	1.22	−0.54
1.55	1.64	1.22	−0.35
1.77	1.64	1.22	−0.10



Exemple (un peu plus sérieux mais pas trop) avec réduction

Matrice de corrélation $\tilde{\Sigma} = \tilde{X}^\top W \tilde{X}$ avec $W = n^{-1} Id_n$:

$$\begin{array}{c} \\ \text{Poids} \\ \text{Dommages} \\ \text{Vitesse} \\ \text{Valeur} \end{array} \begin{array}{c} \begin{array}{cccc} \text{Poids} & \text{Dommages} & \text{Vitesse} & \text{Valeur} \end{array} \\ \begin{pmatrix} 1.00 & 0.93 & -0.21 & 0.60 \\ 0.93 & 1.00 & -0.09 & 0.52 \\ -0.21 & -0.09 & 1.00 & -0.67 \\ 0.60 & 0.52 & -0.67 & 1.00 \end{pmatrix} \end{array}$$

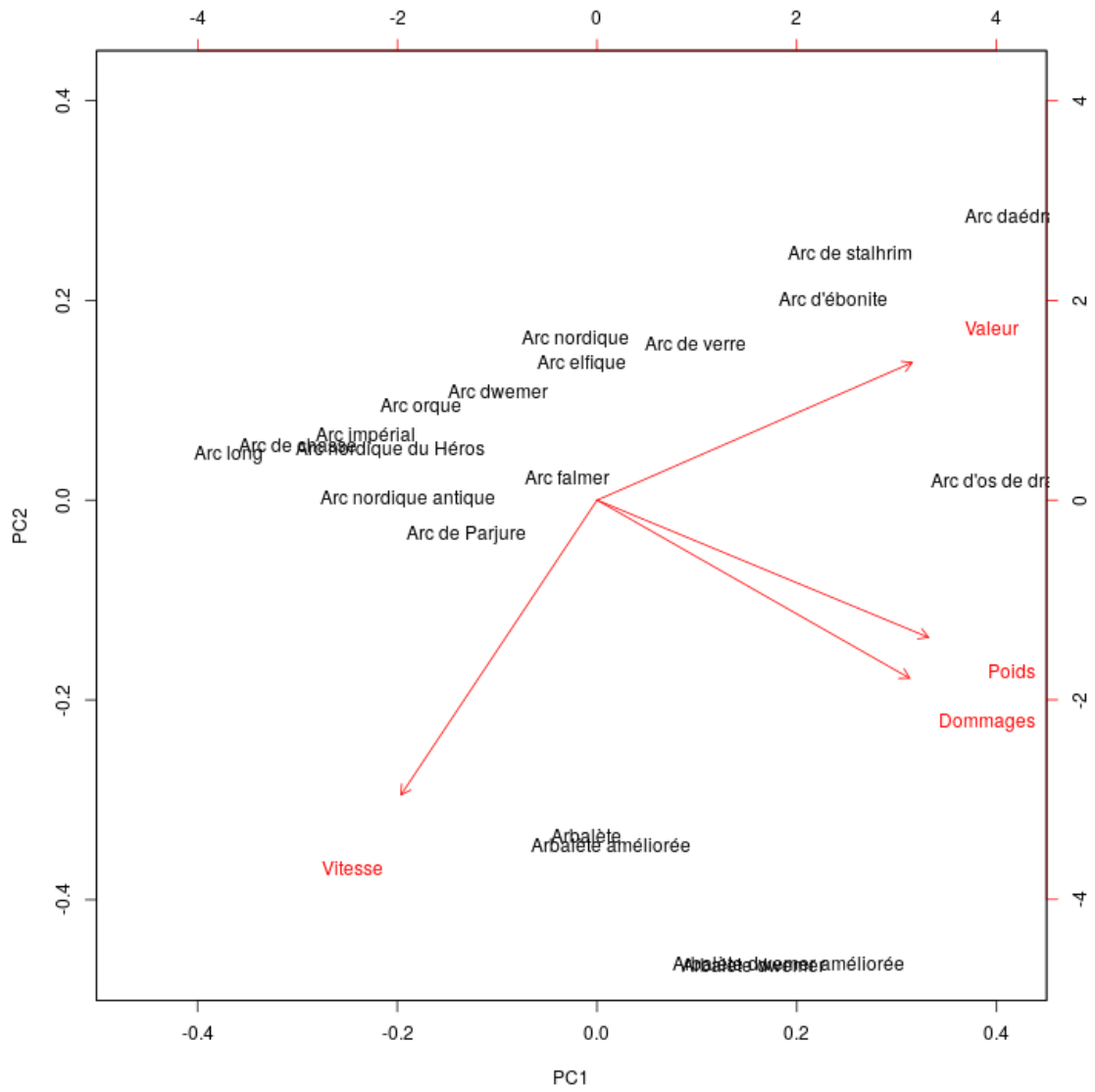
La diagonalisation de $\tilde{\Sigma}$ mène à la matrice $\tilde{V}$ des directions principales et aux valeurs propres

$$\tilde{\lambda}_1 = 2.57,\ \tilde{\lambda}_2 = 1.17,\ \tilde{\lambda}_3 = 0.20 \text{ et } \tilde{\lambda}_4 = 0.06.$$

Inertie du nuage de points initial : $I(\tilde{x}) = \text{Tr}(\tilde{\Sigma}) = 4$

Part d'inertie expliquée par le plan principal : $\dfrac{\tilde{\lambda}_1 + \tilde{\lambda}_2}{I(\tilde{x})} = 93.45\%$

Exemple (un peu plus sérieux mais pas trop) avec réduction



ACP et données réduites (point de vue dual)

La matrice des données réduites s'écrit également $\tilde{X} = XM^{1/2}$ où $M^{1/2}$ est la matrice **symétrique** et **positive** définie par

$$M^{1/2} = \begin{pmatrix} \frac{1}{\sigma(x^1)} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma(x^2)} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \frac{1}{\sigma(x^p)} \end{pmatrix}.$$

Dans le cas unidimensionnel de la recherche d'une direction $\tilde{v} \in \mathbb{R}^p$ de plus grande variabilité, nous avons vu qu'il fallait rendre maximale l'inertie des données projetées sur la droite $\mathbb{R}\tilde{v}$,

$$\tilde{v}^\top \tilde{\Sigma} \tilde{v} = \langle \tilde{\Sigma} \tilde{v}, \tilde{v} \rangle.$$

ACP et données réduites (point de vue dual)

Inertie des données projetées : $\tilde{v}^\top \tilde{\Sigma} \tilde{v} = \langle \tilde{\Sigma} \tilde{v}, \tilde{v} \rangle$.
En posant $\tilde{v} = M^{1/2} v$, cette inertie s'écrit

$$\begin{aligned}
\tilde{v}^\top \tilde{\Sigma} \tilde{v} &= \tilde{v}^\top \tilde{X}^\top W \tilde{X} \tilde{v} \\
&= \left(M^{1/2} v\right)^\top \left(X M^{1/2}\right)^\top W \left(X M^{1/2}\right) \left(M^{1/2} v\right) \\
&= v^\top M (X^\top W X) M v = v^\top M \Sigma M v = \langle \Sigma M v, v \rangle_M
\end{aligned}$$

Manque les 1/2

avec, pour tout $u, v \in \mathbb{R}^p$, $\langle u, v \rangle_M = u^\top M v$.

Le **produit scalaire** $\langle \cdot, \cdot \rangle_M$ modifie la géométrie de $\mathbb{R}^p$ en donnant une **importance différente à chaque coordonnée** (*i.e.* à chaque variable) en fonction des variances observées. Ainsi, l'ACP calculée sur les données réduites n'est rien d'autre que l'**ACP des données initiales calculée avec une structure euclidienne** induite par la matrice M.

Exemple (plus sérieux)

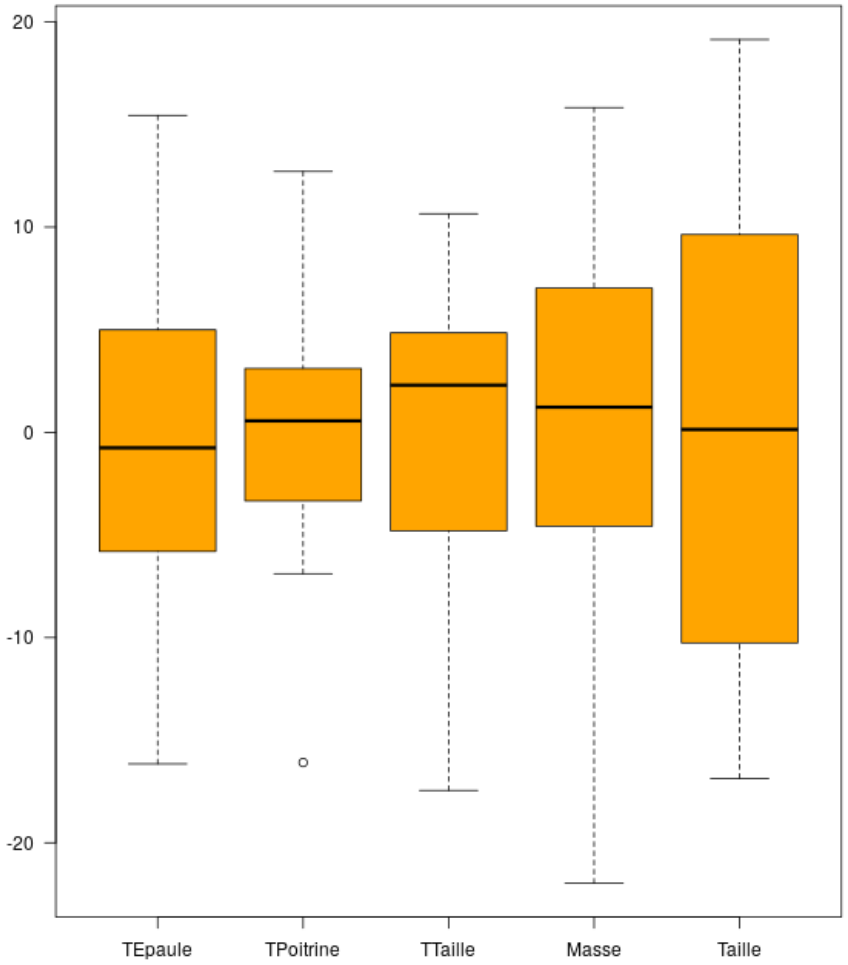
Nous disposons de $p = 5$ variables morphologiques mesurées sur $n = 20$ individus contenant 10 hommes et 10 femmes.

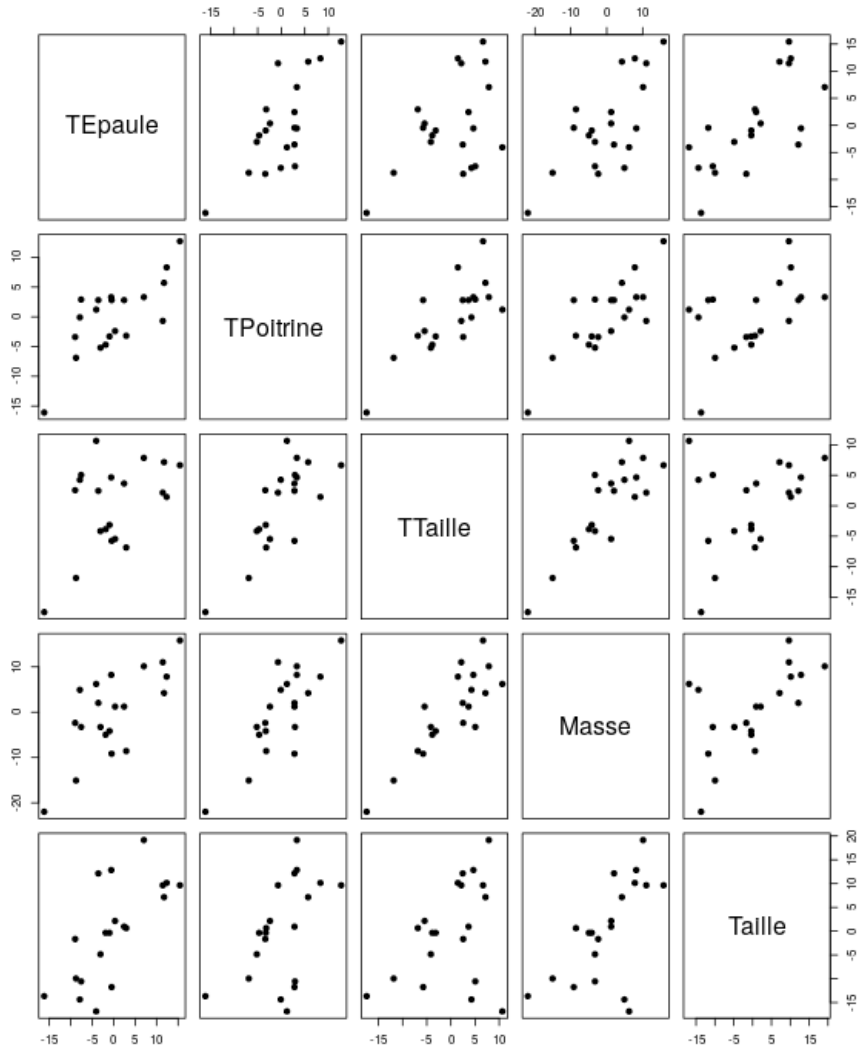
`TEpaule` : tour d'épaules (cm)
`TPoitrine` : tour de poitrine (cm)
`TTaille` : tour de taille (cm)
`Masse` : masse (kg)
`Taille` : taille (cm)

	TEpaule	TPoitrine	TTaille	Masse	Taille
H1	106.2	89.5	71.5	65.6	174.0
H2	110.5	97.0	79.0	71.8	175.3
H3	115.1	97.5	83.2	80.7	193.5
H4	104.5	97.0	77.8	72.6	186.5
H5	107.5	97.5	80.0	78.8	187.2
H6	119.8	99.9	82.5	74.8	181.5
H7	123.5	106.9	82.0	86.4	184.0
H8	120.4	102.5	76.8	78.4	184.5
H9	111.0	91.0	68.5	62.0	175.0
H10	119.5	93.5	77.5	81.6	184.0
F1	105.0	89.0	71.2	67.3	169.5
F2	100.2	94.1	79.6	75.5	160.0
F3	99.1	90.8	77.9	68.2	172.7
F4	107.6	97.0	69.6	61.4	162.6
F5	104.0	95.4	86.0	76.8	157.5
F6	108.4	91.8	69.9	71.8	176.5
F7	99.3	87.3	63.5	55.5	164.4
F8	91.9	78.1	57.9	48.6	160.7
F9	107.1	90.9	72.2	66.4	174.0
F10	100.5	97.1	80.4	67.3	163.8

Exemple (plus sérieux)

X : matrice des données **centrées**





Exemple (plus sérieux)

Matrice de covariance $\Sigma = X^\top W X$ avec $W = n^{-1} Id_n$:

	TEpaule	TPoitrine	TTaille	Masse	Taille
TEpaule	65.21	35.85	26.68	52.55	58.13
TPoitrine	35.85	35.64	32.20	43.42	30.78
TTaille	26.68	32.20	48.24	53.76	26.31
Masse	52.55	43.42	53.76	81.42	56.55
Taille	58.13	30.78	26.31	56.55	103.84

La diagonalisation de Σ mène à la matrice V des directions principales et aux valeurs propres

$$\lambda_1 = 242.87,\ \lambda_2 = 57.17,\ \lambda_3 = 22.31, \lambda_4 = 8.18 \text{ et } \lambda_5 = 3.81.$$

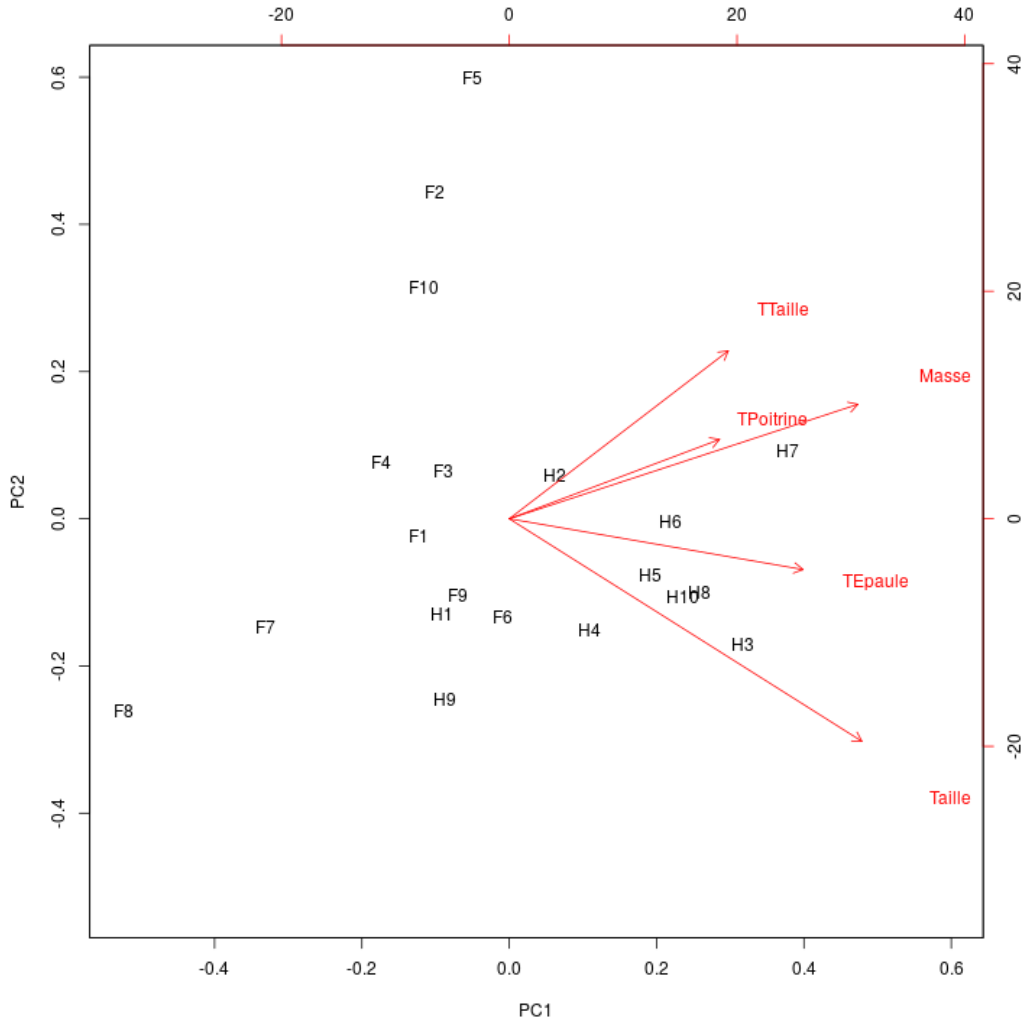
Inertie du nuage de points initial : $I(x) = 334.3$

Part d'inertie expliquée par le plan principal : $\dfrac{\lambda_1 + \lambda_2}{I(x)} = 89.74\%$

Exemple (plus sérieux)

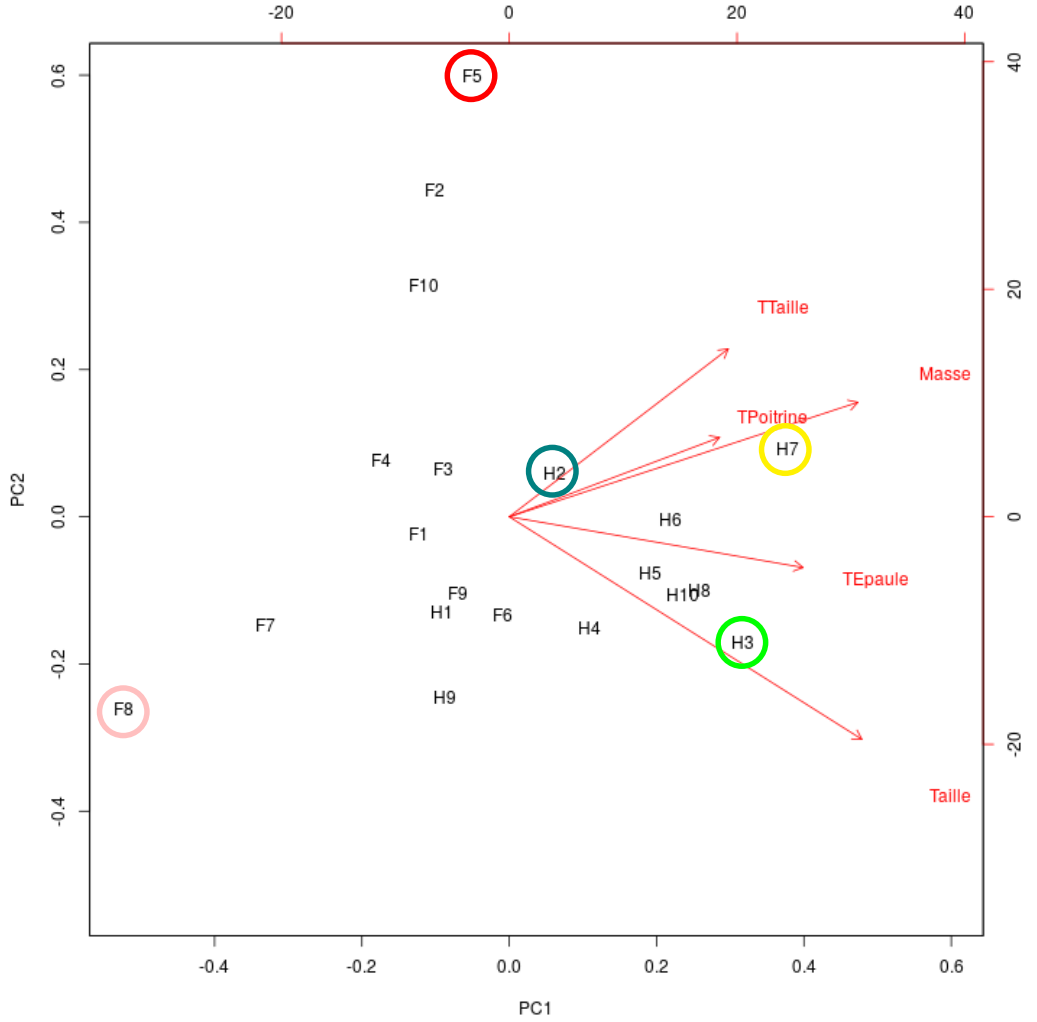
PC1 : « Gabarit »
Sépare les grands (valeurs élevées pour les 5 variables) à droite et les petits à gauche.

PC2 : « Embonpoint »
Sépare les variables liées à la taille et à la carrure (en bas) et celles liées à la masse et aux tours de taille et de poitrine (en haut).



Exemple (plus sérieux)

	TEpaule	TPoitrine	TTaille	Masse	Taille
H1	106.2	89.5	71.5	65.6	174.0
H2	110.5	97.0	79.0	71.8	175.3
H3	115.1	97.5	83.2	80.7	193.5
H4	104.5	97.0	77.8	72.6	186.5
H5	107.5	97.5	80.0	78.8	187.2
H6	119.8	99.9	82.5	74.8	181.5
H7	123.5	106.9	82.0	86.4	184.0
H8	120.4	102.5	76.8	78.4	184.5
H9	111.0	91.0	68.5	62.0	175.0
H10	119.5	93.5	77.5	81.6	184.0
F1	105.0	89.0	71.2	67.3	169.5
F2	100.2	94.1	79.6	75.5	160.0
F3	99.1	90.8	77.9	68.2	172.7
F4	107.6	97.0	69.6	61.4	162.6
F5	104.0	95.4	86.0	76.8	157.5
F6	108.4	91.8	69.9	71.8	176.5
F7	99.3	87.3	63.5	55.5	164.4
F8	91.9	78.1	57.9	48.6	160.7
F9	107.1	90.9	72.2	66.4	174.0
F10	100.5	97.1	80.4	67.3	163.8

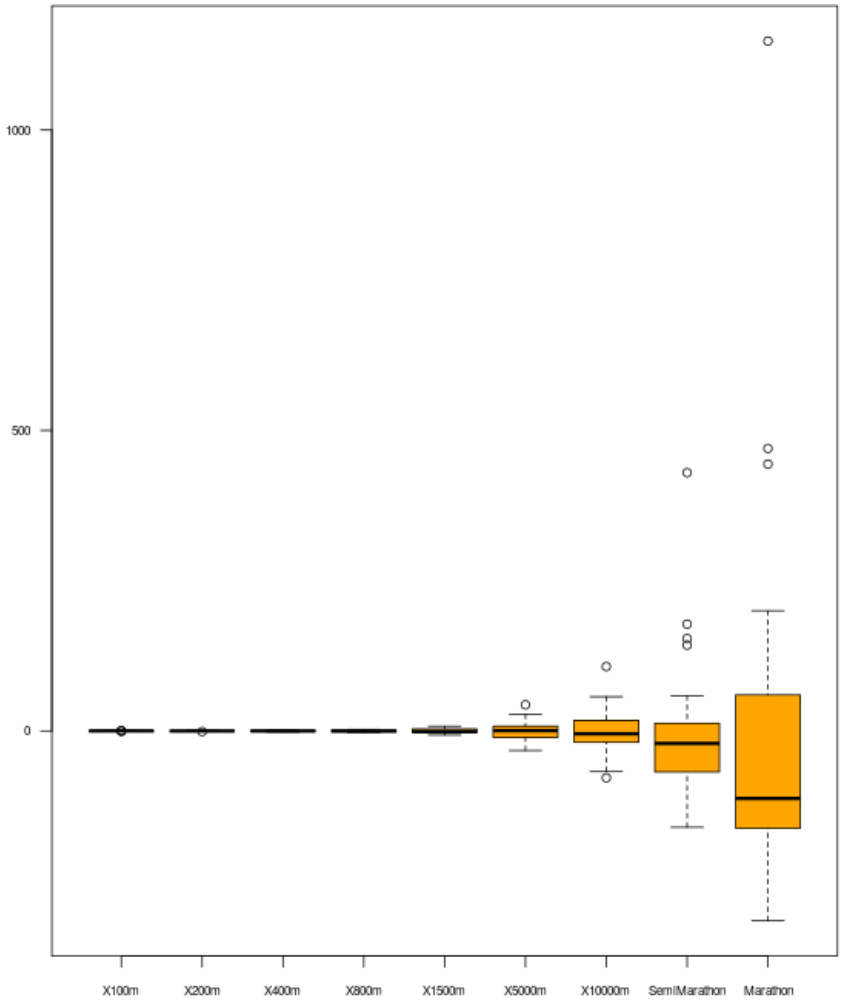


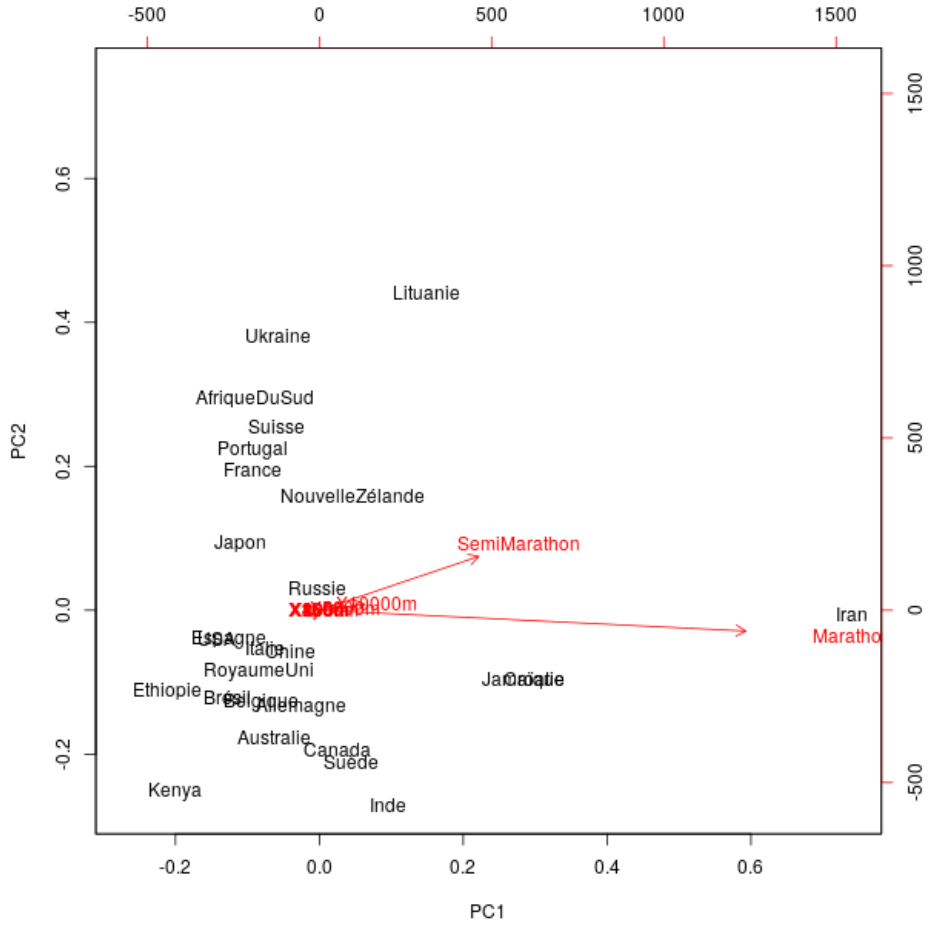
Exemple (un dernier)

Voici $n = 26$ records nationaux (en sec.) de $p = 9$ épreuves d'athlétisme :

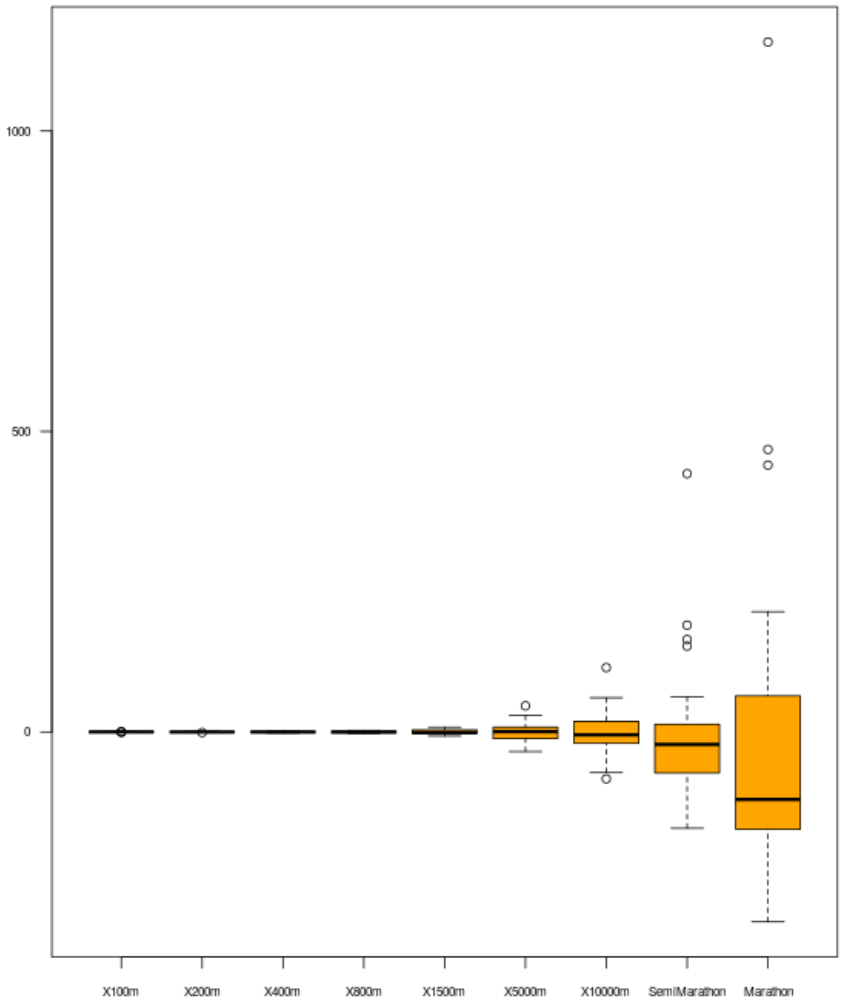
	X100m	X200m	X400m	X800m	X1500m	X5000m	X10000m	SemiMarathon	Marathon
Australie	9.93	20.06	44.38	104.40	211.96	775.76	1649.73	3602	7671
Belgique	10.02	20.19	44.78	103.86	214.13	769.71	1612.30	3605	7640
Brésil	10.00	19.89	44.29	101.77	213.25	799.43	1648.12	3573	7565
RoyaumeUni	9.87	19.87	44.36	101.73	209.67	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

Exemple (un dernier)





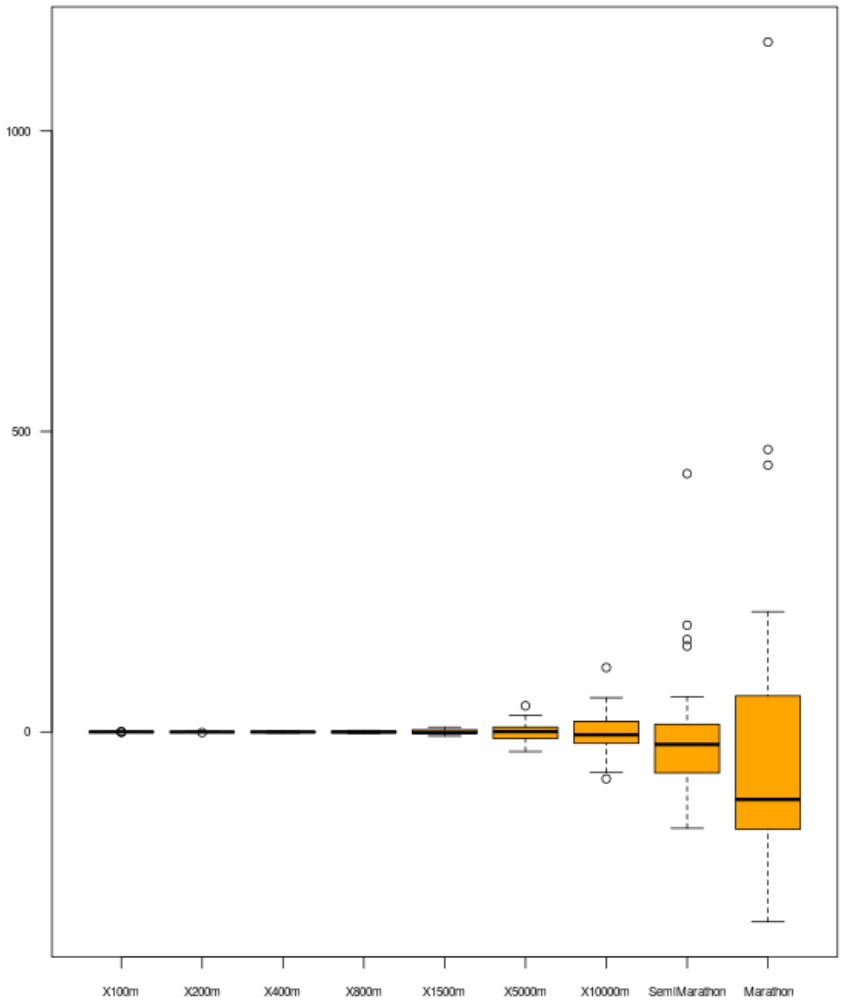
Exemple (un dernier)



Les épreuves de longue durée capturent presque toute la variabilité du jeu de données.

Qu'allons-nous perdre en réduisant les variances ?

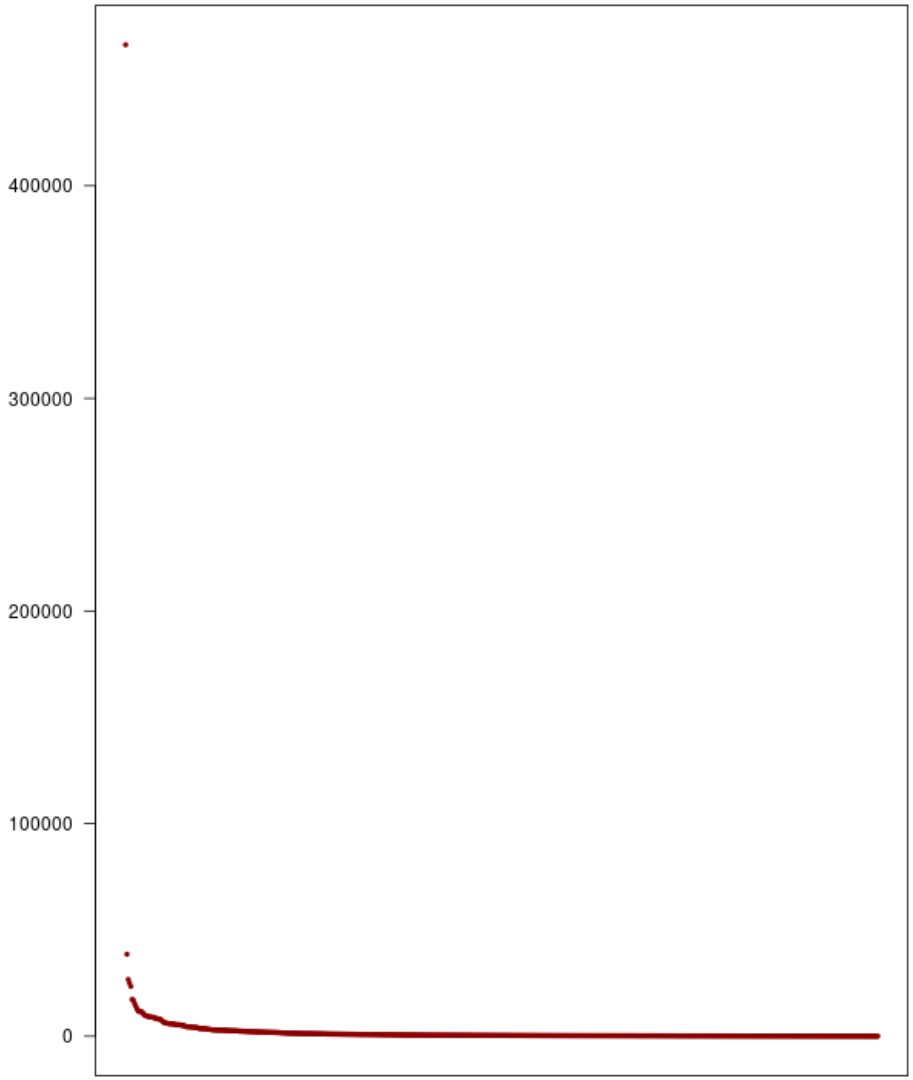
Exemple (un dernier)

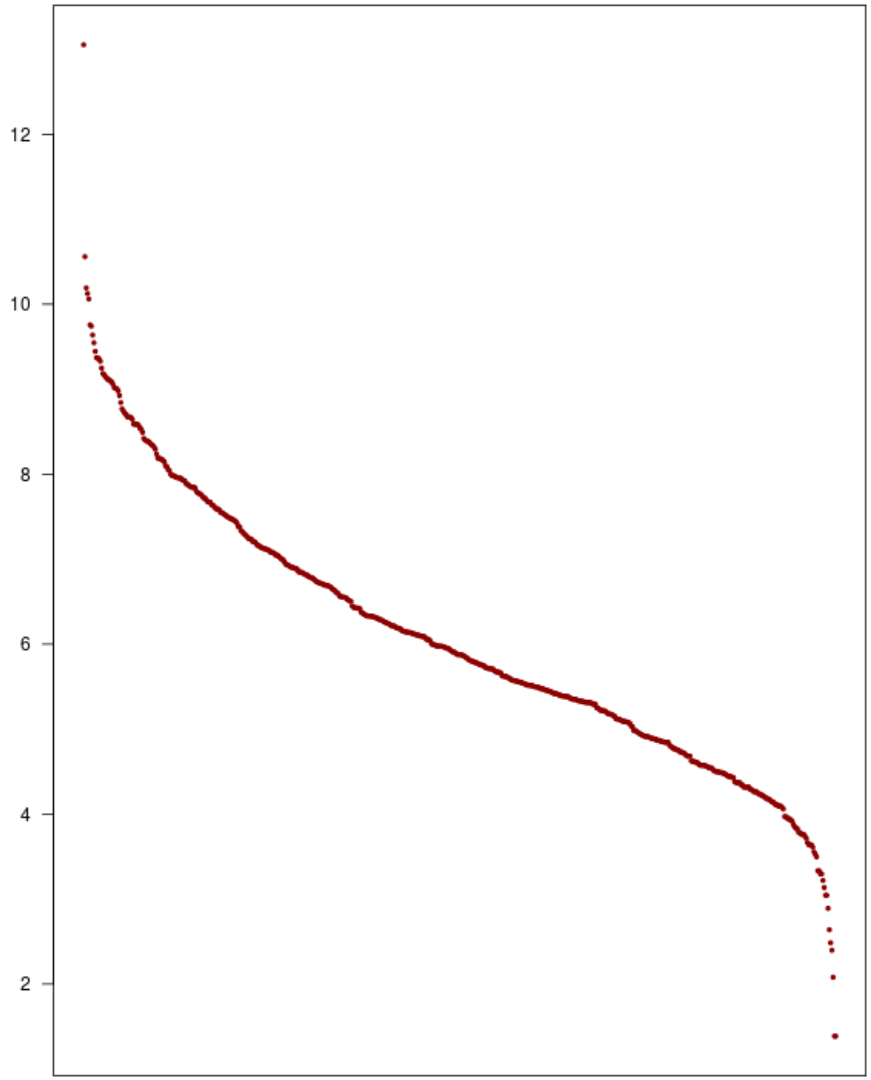


Les épreuves de longue durée capturent presque toute la variabilité du jeu de données.
Qu'allons-nous perdre en réduisant les variances ?

Les observations suggèrent une **structure d'échelle exponentielle** entre les différentes variables. Une réduction exprimerait les observations dans une même échelle neutre mais ferait également **disparaître cette structure entre les variables**. Il faut considérer une **transformation globale** plutôt que des transformations par variable.

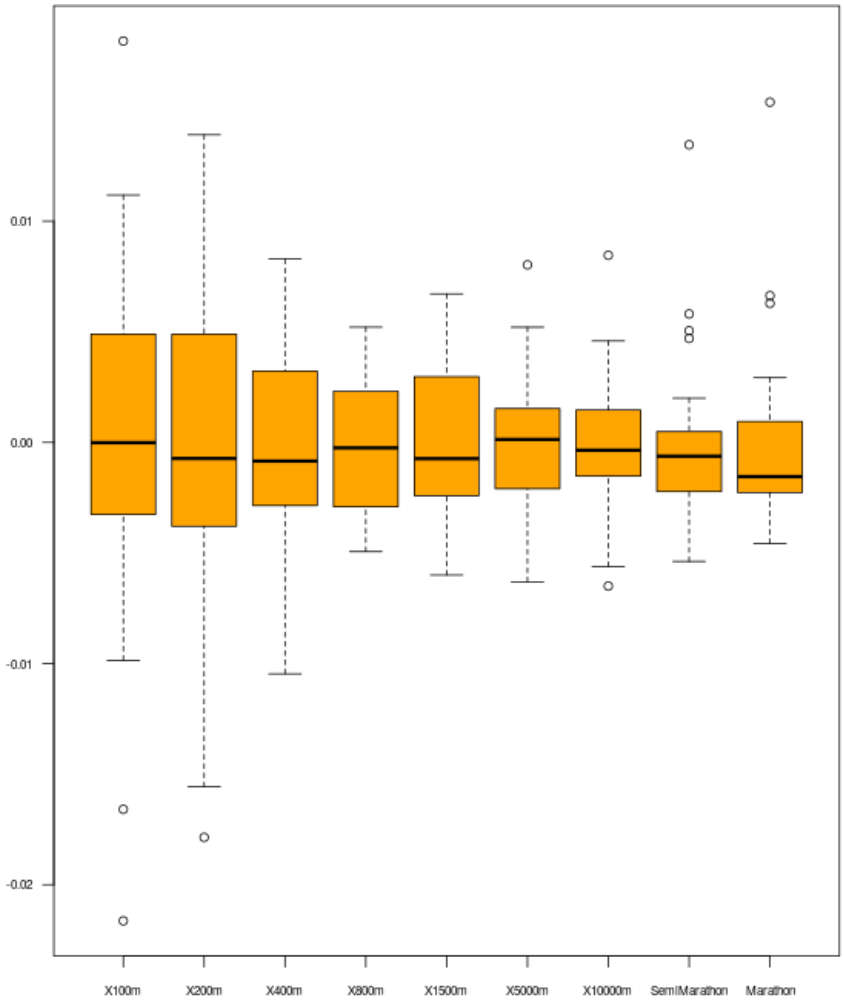
Transformation logarithmique

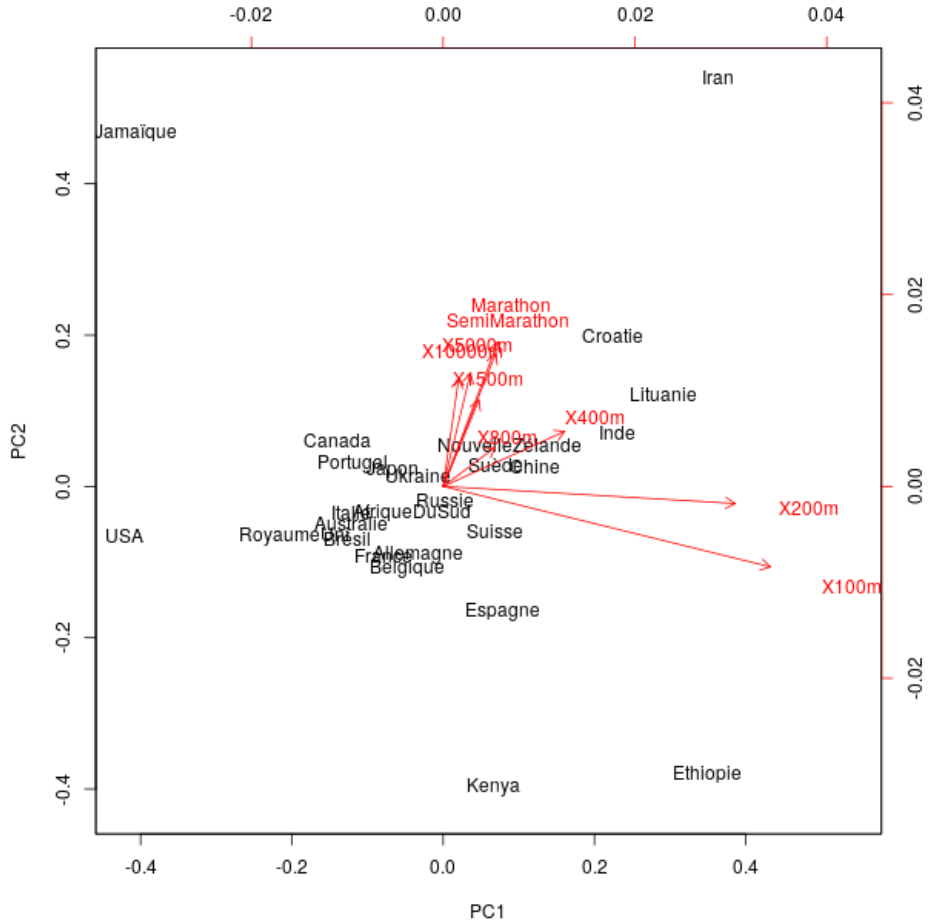




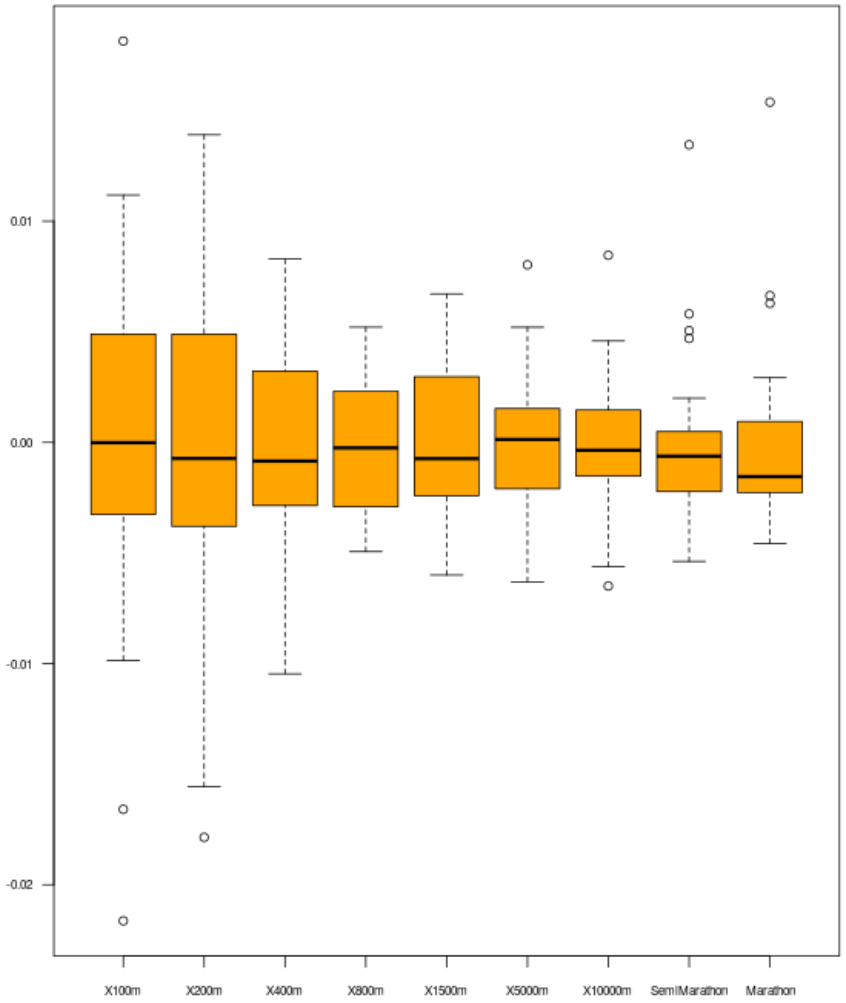
Données : population des 589 communes de Haute-Garonne en 2014, *INSEE*.

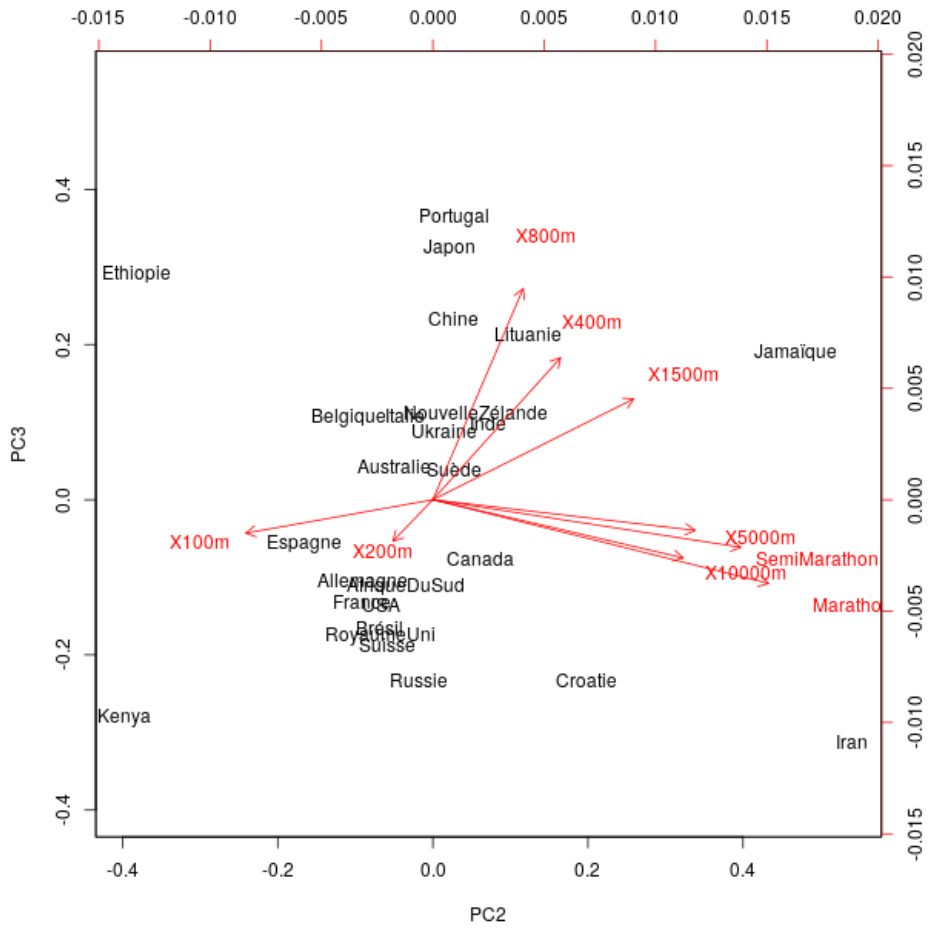
Exemple (un dernier) avec transformation logarithmique





Exemple (un dernier) avec transformation logarithmique





2.0 Application à la réduction de dimension

Regression en composante principale (PCR)

Contexte : **La régression**.
On observe $(X_i, Y_i)_{i=1}^n$ i.i.d avec les $(Y_i)_{i=1}^n$ des sorties réelles que l'on cherche à expliquer en fonctions de $(X_i)_{i=1}^n$, des covariables dans $\mathbb{R}^d$.

Fléau de la dimension

Ajuster un modèle de regression peut être difficile notament lorsque la dimension d devient grande.
C'est ce qu'on appel **le fléau (ou la malédiction) de la dimension**.

Solution $\rightsquigarrow$ Appliquée une ACP sur les covariable pour réduire la dimension.

Regression (linéaire) en composante principale

Soit $p < d$ le nombre de composantes principales de X sur lequel ont veut régresser Y.
Notons v_k la k-ième composante principale. On cherche alors à ajuster le modèle linéaire suivant :

$$Y = \sum_{k=1}^{p} \alpha_k v_k$$

Attention

- Les coefficient α_k sont calculés avec les X_i projetés sur les v_k, $k = 1, \ldots, p$.
- Il n'y a pas d'intercept car on centre toujours avant de faire une ACP.

PCR : retour aux variables de départ

Par construction les composantes principales v_k, $k = 1, \ldots, p$ sont construites comme combinaisons linéaires des x^l, $l = 1, \ldots, d$, notons

$$v_k = \sum_{l=1}^{d} \beta_{l,k} x^l, \quad k = 1, .., d.$$

On peut alors ecrire la regression avec les variable de base

$$y = \sum_{l=1}^{d} \left(\sum_{k=1}^{p} \alpha_k \beta_{l,k} \right) x^l.$$

A noter que l'intercept apparait après la "dénormalisation" des données.

PCR : formulation générale

Soit X (centré-réduit). ACP : $X = TP^\top$, avec $T^\top T$ diagonal.

$$Y = T_k \beta + \varepsilon.$$

Estimation des paramètres (moindres carrés sur les scores) :

$$\hat{\beta}_T = (T_k^\top T_k)^{-1} T_k^\top Y,$$

et retour sur l'échelle des variables originales :

$$\hat{\beta}_{PCR} = P_k \hat{\beta}_T.$$

PCR : formulation générale

Soit X (centré-réduit). ACP : $X = TP^\top$, avec $T^\top T$ diagonal.

$$Y = T_k\beta + \varepsilon.$$

Estimation des paramètres (moindres carrés sur les scores) :

$$\hat{\beta}_T = (T_k^\top T_k)^{-1} T_k^\top Y,$$

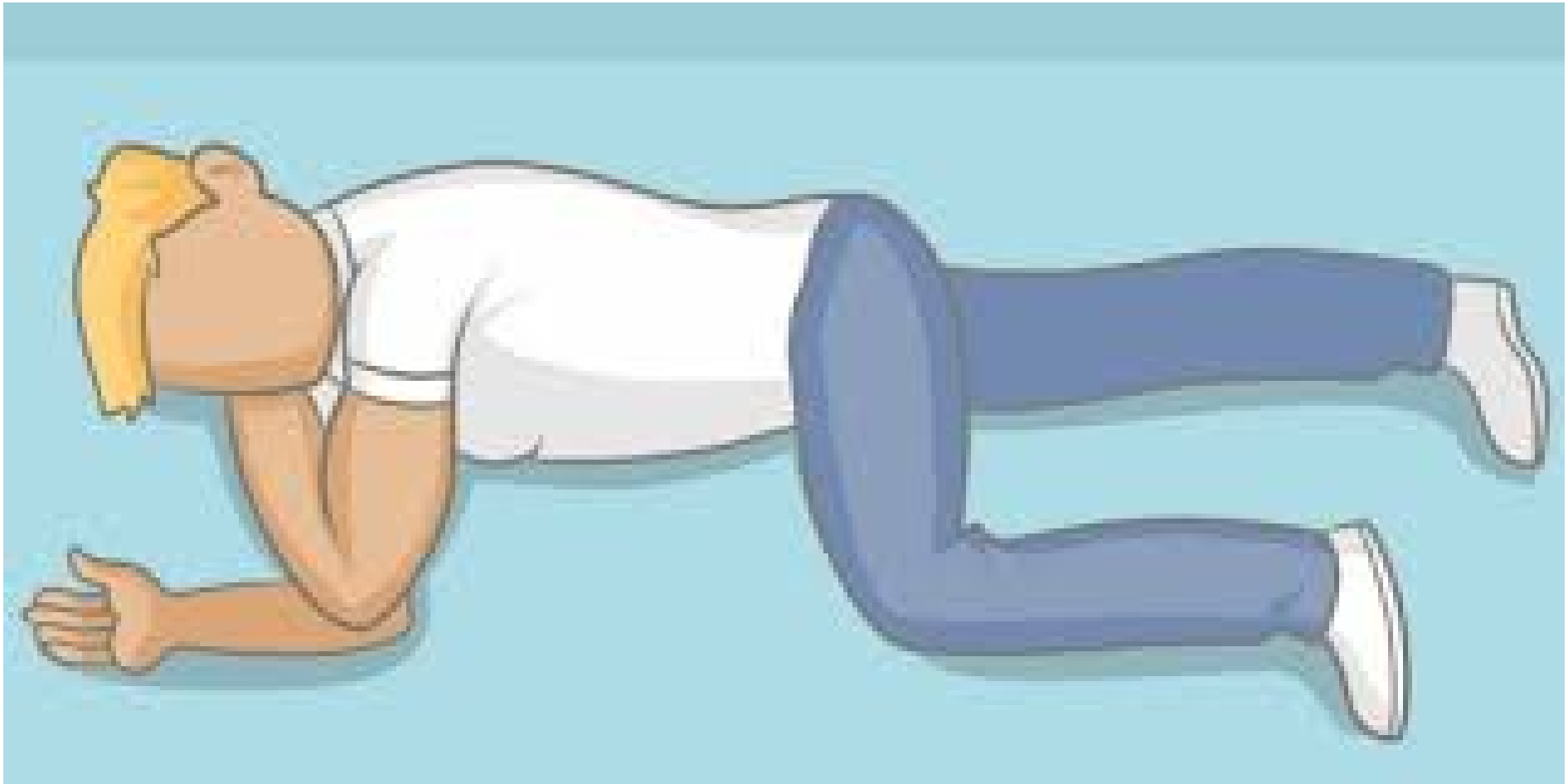
et retour sur l'échelle des variables originales :

$$\hat{\beta}_{PCR} = P_k \hat{\beta}_T.$$

Problème : les composantes T_k sont construites sans tenir compte de Y.

Dans un cadre linéaire on peut faire mieux !

La regression en moindres carrés partiel (PLS).



La regression PLS

On cherche à ajuster un modèle $Y = X\beta + \varepsilon$, avec ε un vecteur de bruit. Ici Y est les sorties et X la matrice des covariable.

Idée

C'est une méthode **itérative**, on cherche deux matrices de projection P et Q telles que

P1,Q1 de norme 1

- On construit les premières lignes, $(P_1, Q_1) = \underset{(u,v)}{argmax}\ cov(Xu, Yv)^2$.
- Pour $k > 1$ on construit
  - $X_k = X_{k-1} - P_{k-1}P_{k-1}^t X_{k-1}$
  - $Y_k = Y_{k-1} - Q_{k-1}Q_{k-1}^t Y_{k-1}$
  - $(P_k, Q_k) = \underset{(u,v)}{argmax}\ cov(X_{k-1}u, Y_{k-1}v)^2$.
- On choisit le nombre de composante r sur lequel on veut régresser, et on fair une régression linéaire "classique" sur les r première composante de PX.

PLS : Sortie en dimension 1

On veut régresser $Y \in \mathbb{R}$ sur $X \in \mathbb{R}^d$.
On cherche alors une matrice U étant la solution du problème suivant :
Pour tout $h = 1, \dots, r < d$,

$$u_h = \underset{u}{argmax}\ cov(Y, Xu) = \underset{u}{argmax}\ u^t X^t Y Y^t X u$$

sous la contrainte $\|u_h\| = 1$.

- Les vecteurs u_h sont appelés **vecteurs de chargement ou pondérations**.
- Les variable $e_h := Xu_h$ sont appelées **variables latentes**.

La regression PLS à r composante est alors la régression linéaire de Y sur les variables latentes e_h.

PLS : Algorithme

Algo : PLS 1 sur r composantes

Centrer X

Pour $h = 1, \ldots, r$:

- $u_h = \frac{X^t Y}{\|X^t Y\|}$
- $e_h = X u_h$
- $X = X - e_h e_h^t X$

Ajuster $T = E\beta$, $\beta \in \mathbb{R}^d$.

Pour le cas général l'algorithme le plus courment utilisé est **NIPALS**.

3.0 Analyse factorielle discriminante (AFD)

Motivation : visualiser des groupes

En 1962, le biologiste russe Lubischew a publié une étude de $n = 74$ coléoptères issus de 3 espèces notées A, B et C. Pour chaque insecte, $p = 6$ variables morphologiques ont été mesurées.

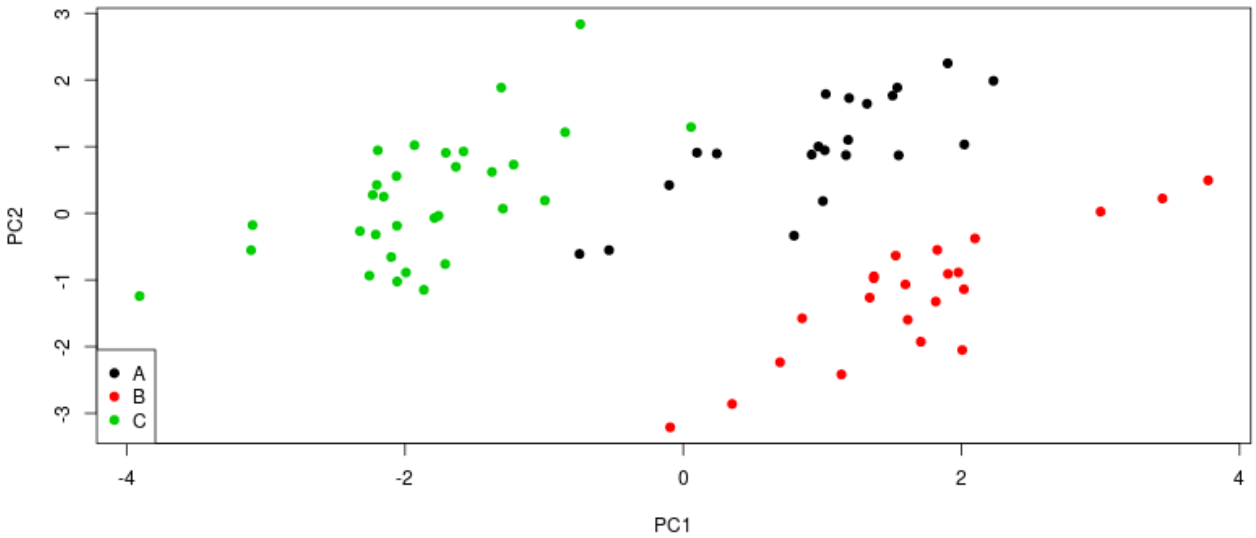
```
      V1  V2 V3  V4 V5  V6 Species
[ 1] 191 131 53 150 15 104       A
[ 2] 185 134 50 147 13 105       A
[ 3] 200 137 52 144 14 102       A
[ 4] 173 127 50 144 16  97       A
...
[22] 158 141 58 145  8 107       B
[23] 146 119 51 140 11 111       B
[24] 151 130 51 140 11 113       B
[25] 122 113 45 131 10 102       B
...
[44] 186 107 49 120 14  84       C
[45] 211 122 49 123 16  95       C
[46] 201 114 47 130 14  74       C
[47] 242 131 54 131 16  90       C
...
```

Objectif : déterminer des combinaisons de variables qui permettent de **discriminer** au mieux les 3 espèces et donner une **représentation graphique** de cette discrimination.

Motivation : visualiser des groupes

Première idée naïve

Faire une ACP sur les observations (réduites) des 6 variables quantitatives et faire apparaître la variable catégorielle `Species` sur le résultat.



Est-ce la « meilleure » façon de faire ? La variable catégorielle est utilisée uniquement **a posteriori**, pouvons-nous utiliser sa connaissance de manière plus pertinente ?

Apprentissage supervisé

Le cadre statistique que nous considérons ici est celui de l'**apprentissage supervisé** dans lequel nous disposons de n observations de :

- p variables réelles $x^1, \dots, x^p$,
- 1 variable catégorielle t à valeurs dans $\{\tau_1, \dots, \tau_q\}$.

Autrement dit, les données sont de la forme suivante :

$$(x_1, t_1), \dots, (x_n, t_n) \in \mathbb{R}^p \times \{\tau_1, \dots, \tau_q\}.$$

Objectif

Étudier ou prédire la **modalité** de t en fonction des variables $x^1, \dots, x^p$.

Décomposition de la matrice de covariance

Matrice de covariance : $\Sigma = \frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x})(x_k - \overline{x})^\top$

Les observations de la variable t induisent une **partition** naturelle des observations en q groupes $G_1, \ldots, G_q$ (supposés non vides),

$$\forall m \in \{1, \ldots, q\},\ G_m = \{k \in \{1, \ldots, n\} \text{ tels que } t_k = \tau_m\}.$$

Pour chaque $m \in \{1, \ldots, q\}$, nous pouvons définir le vecteur $\overline{x}_m \in \mathbb{R}^p$ des moyennnes des observations du groupe G_m,

$$\overline{x}_m = \begin{pmatrix} \overline{x}_m^1 \\ \vdots \\ \overline{x}_m^p \end{pmatrix} \text{ avec } \forall \ell \in \{1, \ldots, p\},\ \overline{x}_m^\ell = \frac{1}{n_m}\sum_{k \in G_m} x_k^\ell$$

où n_m est la taille du groupe G_m.

Décomposition de la matrice de covariance

Matrice de covariance : $\Sigma = \frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x})(x_k - \overline{x})^\top$

Faisons apparaître ces moyennes par groupe dans l'expression de la matrice de covariance,

$$\begin{aligned}
\Sigma &= \frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x})(x_k - \overline{x})^\top \\
&= \frac{1}{n}\sum_{m=1}^{q}\sum_{k\in G_m}\left((x_k - \overline{x}_m) + (\overline{x}_m - \overline{x})\right)\left((x_k - \overline{x}_m) + (\overline{x}_m - \overline{x})\right)^\top \\
&= \frac{1}{n}\sum_{m=1}^{q}\sum_{k\in G_m}(x_k - \overline{x}_m)(x_k - \overline{x}_m)^\top + \frac{1}{n}\sum_{m=1}^{q}\sum_{k\in G_m}(\overline{x}_m - \overline{x})(\overline{x}_m - \overline{x})^\top
\end{aligned}$$

car $\sum_{k\in G_m}(x_k - \overline{x}_m) = 0$ par définition.

Décomposition de la matrice de covariance

Matrice de covariance : $\Sigma = \frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x})(x_k - \overline{x})^\top$

Ainsi,

$$\Sigma = \Sigma_w + \Sigma_b$$

où nous avons défini :

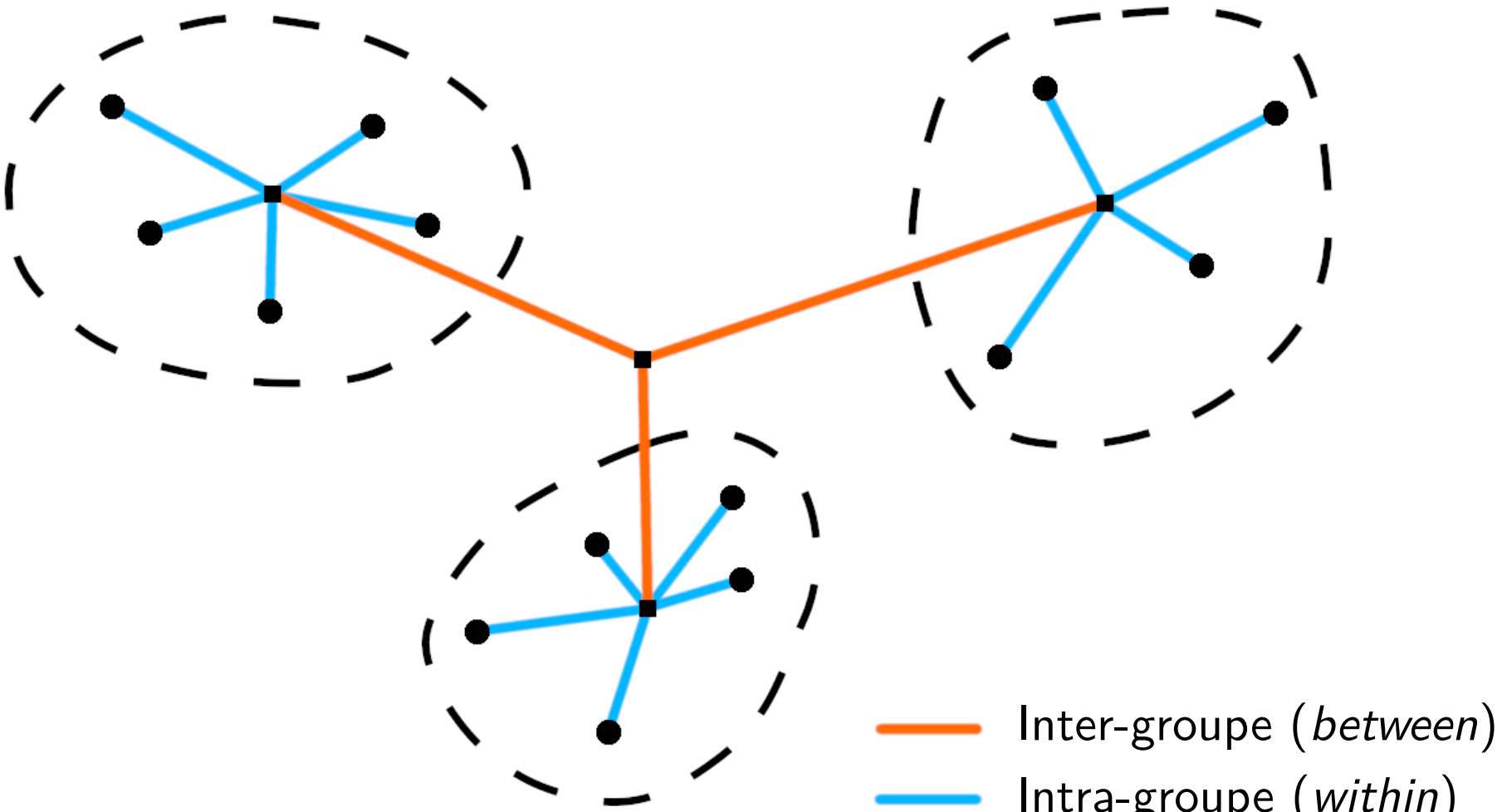
- la **matrice de covariance intra-groupe** (*within*)

$$\Sigma_w = \frac{1}{n}\sum_{m=1}^{q} n_m \times \frac{1}{n_m}\sum_{k\in G_m}(x_k - \overline{x}_m)(x_k - \overline{x}_m)^\top,$$

- la **matrice de covariance inter-groupe** (*between*)

$$\Sigma_b = \frac{1}{n}\sum_{m=1}^{q} n_m(\overline{x}_m - \overline{x})(\overline{x}_m - \overline{x})^\top.$$

Décomposition de la matrice de covariance (visuel)



Principe de l'AFD (cas unidimensionnel)

Nous avons vu dans la section consacrée à l'ACP que l'inertie des données projetées sur la droite engendrée par un vecteur $v \in \mathbb{R}^p$ unitaire est égale à

$$v^\top \Sigma v = \underbrace{v^\top \Sigma_w v}_{\text{Inertie intra-groupe}} + \underbrace{v^\top \Sigma_b v}_{\text{Inertie inter-groupe}}$$

où la décomposition de l'inertie découle de celle de la matrice de covariance.

L'inertie intra-groupe quantifie la **variabilité à l'intérieur des groupes** et l'inertie inter-groupe celle **entre les groupes**.

Principe de l'AFD (cas unidimensionnel)

Nous avons vu dans la section consacrée à l'ACP que l'inertie des données projetées sur la droite engendrée par un vecteur $v \in \mathbb{R}^p$ unitaire est égale à

$$v^\top \Sigma v = \underbrace{v^\top \Sigma_w v}_{\text{Inertie intra-groupe}} + \underbrace{v^\top \Sigma_b v}_{\text{Inertie inter-groupe}}$$

où la décomposition de l'inertie découle de celle de la matrice de covariance.

L'inertie intra-groupe quantifie la **variabilité à l'intérieur des groupes** et l'inertie inter-groupe celle **entre les groupes**.

Des **groupes homogènes et bien séparés** correspondent donc à une **petite inertie intra-groupe** et une **grande inertie inter-groupe**.

Principe de l'AFD (cas unidimensionnel)

Pour trouver une direction qui permette de **discriminer** au mieux les groupes d'observations $G_1, \ldots, G_q$, il faut chercher à rendre simultanément **l'inertie inter-groupe maximale** et **l'inertie intra-groupe minimale**.

Pour cela, l'analyse factorielle discriminante consiste à déterminer un vecteur unitaire $v \in \mathbb{R}^p$ qui maximise le rapport de l'inertie inter-groupe sur l'inertie totale,

$$\frac{v^\top \Sigma_b v}{v^\top \Sigma v}.$$

Remarque : nous supposerons dans la suite que la matrice de covariance Σ est inversible.

Encore un peu d'algèbre linéaire (ne dites pas non...)

Si $v \in \mathbb{R}^p$ est un vecteur propre de $\Sigma^{-1}\Sigma_b$ associé à la valeur propre λ, alors nous avons

$$\frac{v^\top \Sigma_b v}{v^\top \Sigma v} = \lambda.$$

Les valeurs propres de $\Sigma^{-1}\Sigma_b = \Sigma^{-1/2}(\Sigma^{-1/2}\Sigma_b)$ coïncident avec celles de $\Sigma^{-1/2}\Sigma_b\Sigma^{-1/2}$ qui est symétrique et positive puisque Σ_b est une matrice de covariance. Par conséquent, la matrice $\Sigma^{-1}\Sigma_b$ est diagonalisable et ses valeurs propres sont positives.

Par construction, Σ_b est engendrée par q vecteurs **centrés**, son rang vaut donc au plus $q-1$. Il en va de même pour $\Sigma^{-1}\Sigma_b$ et les $q-1$ valeurs propres (potentiellement) non triviales sont notées

$$\lambda_1 \geqslant \cdots \geqslant \lambda_{q-1} \geqslant 0.$$

Encore un peu d'algèbre linéaire (ne dites pas non...)

Si $v \in \mathbb{R}^p$ est un vecteur propre de $\Sigma^{-1}\Sigma_b$ associé à la valeur propre λ, alors nous avons

$$\frac{v^\top \Sigma_b v}{v^\top \Sigma v} = \lambda.$$

La solution du problème de l'AFD unidimensionnel est donc donnée par le vecteur propre unitaire v^1 de la matrice $\Sigma^{-1}\Sigma_b$ associé à la plus grande valeur propre λ_1.

Principe de l'AFD (cas général)

Nous nous intéressons maintenant à la recherche d'un espace E_d engendré par d vecteurs $v^1, \dots, v^d \in \mathbb{R}^p$ libres qui permette de discriminer au mieux les groupes d'observations $G_1, \dots, G_q$.

L'inertie des données projetées dans E_d s'exprime à l'aide d'un **déterminant** (admis) et le principe de l'AFD consiste à maximiser le rapport

$$\frac{\det(V^\top \Sigma_b V)}{\det(V^\top \Sigma V)}$$

où V est la matrice de taille $p \times d$ dont les colonnes sont données par les vecteurs $v^1, \dots, v^d$.

Principe de l'AFD (cas général)

Nous nous intéressons maintenant à la recherche d'un espace E_d engendré par d vecteurs $v^1, \dots, v^d \in \mathbb{R}^p$ libres qui permette de discriminer au mieux les groupes d'observations $G_1, \dots, G_q$.

Il est encore possible de montrer que la solution est donnée par les vecteurs propres $v^1, \dots, v^d$ de la matrice $\Sigma^{-1}\Sigma_b$ associés au d plus grandes valeurs propres $\lambda_1 \geqslant \cdots \geqslant \lambda_d$.

Attention

La **dimension maximale est limitée** par le nombre de modalités de la variable catégorielle t. En effet, le rang de la matrice $\Sigma^{-1}\Sigma_b$ est au plus $q-1$ et il ne peut pas y avoir plus de directions pour discriminer les q groupes.

En particulier, pour une variable binaire ($q = 2$), il n'y a qu'une seule direction discriminante **quel que soit le nombre p de variables explicatives**.

L'AFD est une variante d'ACP

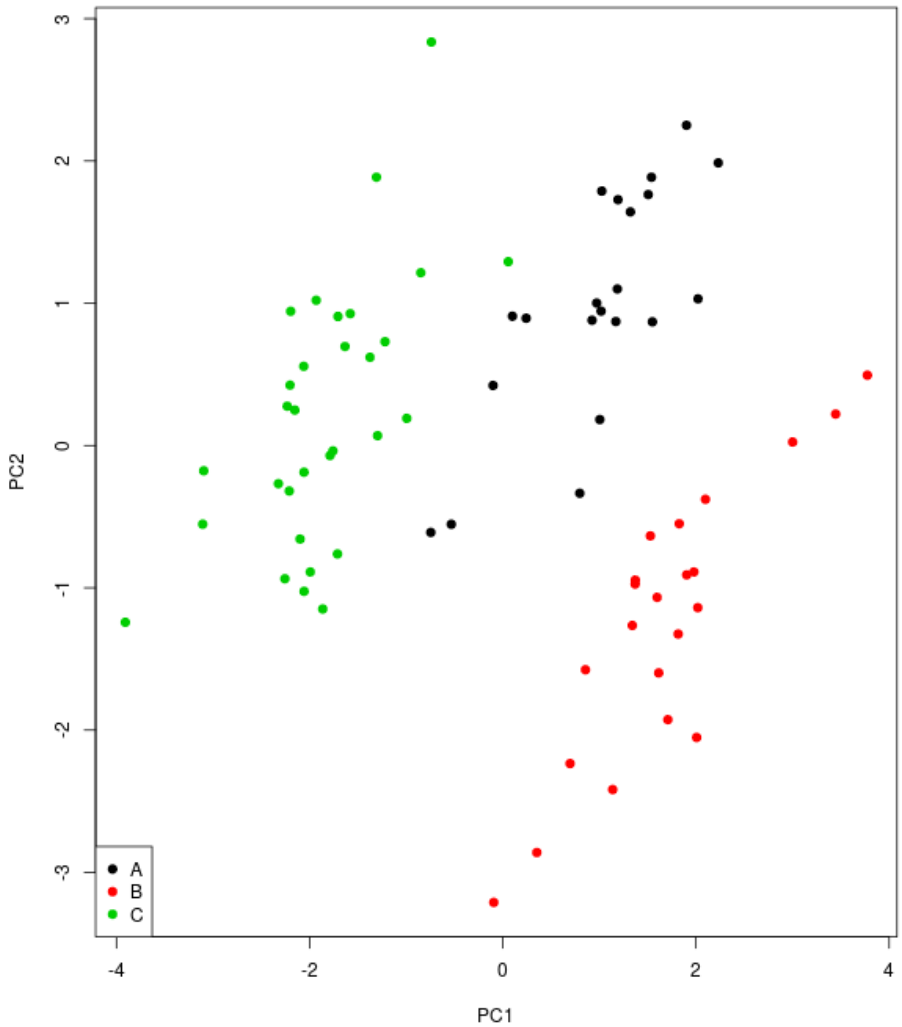
À un **reparamétrage** près, l'AFD peut se comprendre comme une variante de l'**ACP des vecteurs moyens** $\overline{x}_1, \dots, \overline{x}_m \in \mathbb{R}^p$.

Bien que ce point de vue soit souvent utilisé dans la littérature pour son apparente simplicité, la formulation du problème en ces termes demande **plus de technicité**. Il faut en particulier prendre le point de vue dual évoqué dans le cas de l'ACP sur les données réduites (avec $M = \Sigma^{-1}$ pour l'AFD).

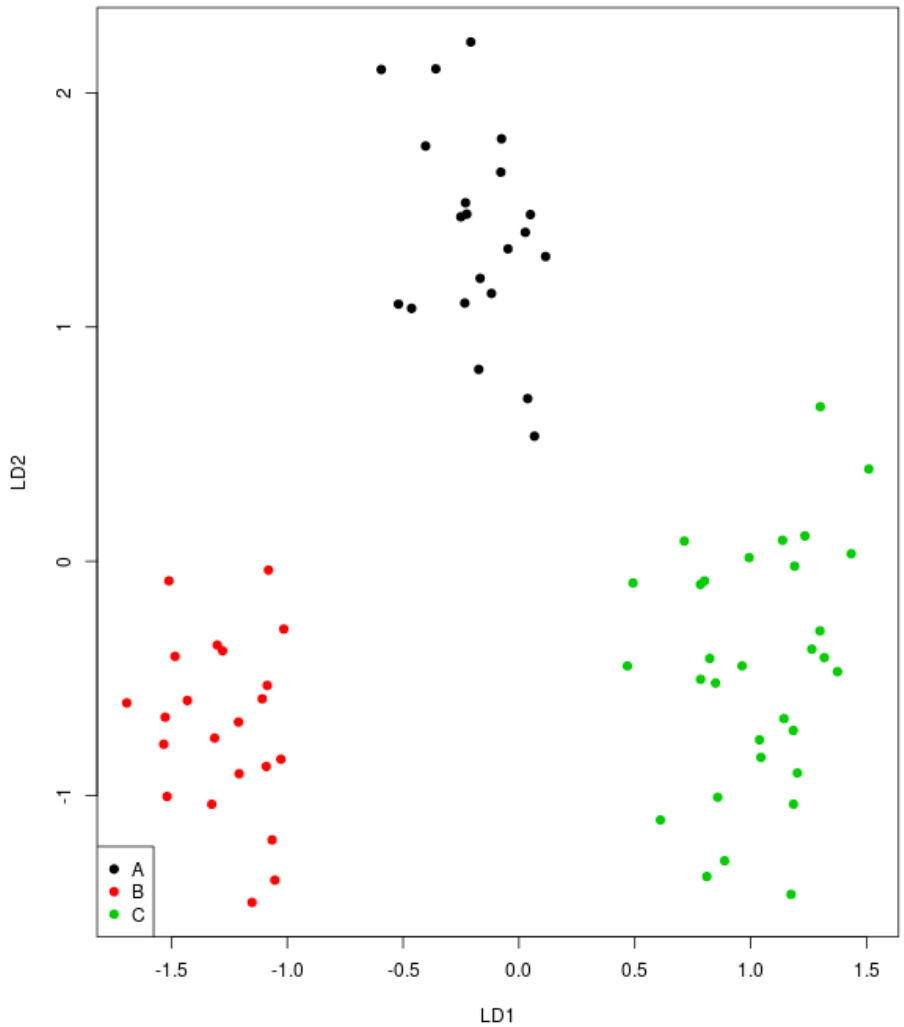
L'avantage principal de ce parallèle est de proposer une **représentation graphique** des données dans le plan engendré par les deux premières directions discriminantes, *i.e.* le plan qui permet de discriminer au mieux les groupes d'observations $G_1, \dots, G_q$.

Exemple des bébêtes

ACP



AFD



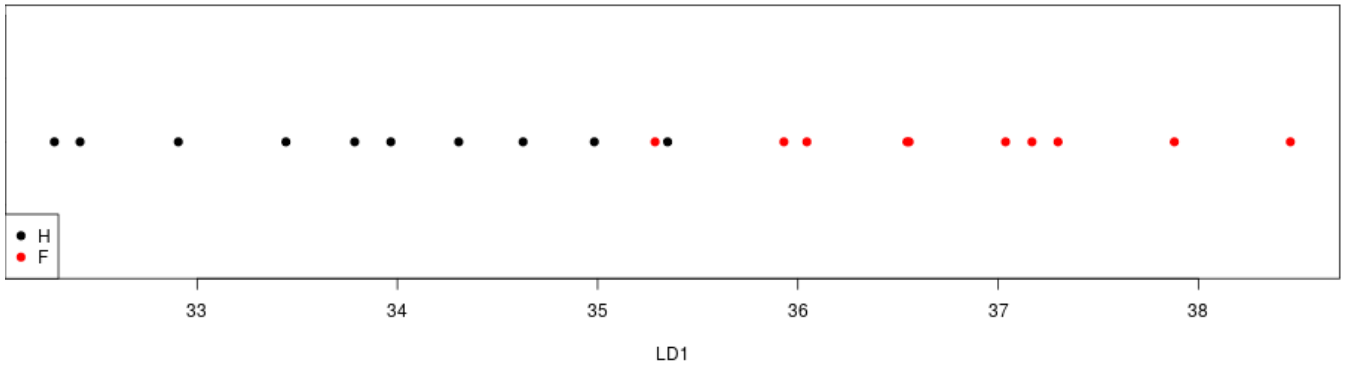
Exemple binaire

Reprenons les données morphologiques constituées de 10 hommes et 10 femmes. Les moyennes par groupe sont données dans le tableau suivant,

	TEpaule	TPoitrine	TTaille	Masse	Taille
Homme	113.80	97.23	77.88	75.27	182.55
Femme	102.31	91.15	72.82	65.88	166.17

La variable catégorielle est binaire, il n'y a donc au plus que 1 **seule direction discriminante**.

	LD1
TEpaule	0.119
TPoitrine	-0.022
TTaille	0.107
Masse	-0.112
Taille	0.139



Selon ces coefficients, la variable qui joue le rôle le plus faible dans la discrimination entre les hommes et les femmes est le tour de poitrine . . .

AFD « décisionnelle »

L'AFD peut être utilisée dans un cadre d'apprentissage supervisé pour **prédire la modalité d'un nouvel individu** à partir des observations des variables quantitatives.

Une méthode simple pour affecter un nouvel individu à une classe donnée consiste à **utiliser le vecteur moyen dont il est le plus proche**. Cette approche souffre cependant de conservatisme et il existe des **règles de classification** plus complexes (plus proches voisins adaptatifs, ...).

Une faiblesse de l'AFD pour répondre à la prédiction d'une modalité apparaît lorsque le nombre de modalités est **faible** et celui des variables quantitatives est **élevé**. Par construction, il y aura peu de directions discriminantes et la méthode se retrouvera limitée. D'autres méthodes existent pour répondre à cette question (**arbres de décision**, **régression logistique**, **réseaux de neurones**, ...) et feront l'objet de cours à venir.